Inferring Adolescent Social Networks Using Partial Ego-Network Substance Use Data

Ju-Sung Lee

Department of Social and Decision Sciences
College of Humanities and Social Sciences
and
The Center for the Computational Analysis of
Social and Organizational Systems

Carnegie Mellon University
Pittsburgh, Pennsylvania

May 15, 2008

Dissertation Committee:

Kathleen M. Carley (Chair)

Jonathan P. Caulkins

Shelby Stewman

Submitted in partial fulfillment of the requirements for the Degree of
Doctor of Philosophy in Sociology

Copyright ©2008 Ju-Sung Lee, All Rights Reserved

Abstract

This dissertation explores the social network processes involved in adolescent substance use. Over the past three decades, researchers have focused on, with increasing clarity, the specific dynamics of peer selection and peer influence in their attempts to understand how adolescents first use a substance, what compels them to continue use, and why some of them quit. However, the exact nature of interplay between those two key social processes continues to be elusive, due to the lack of both robust longitudinal network data and sophisticated network methodologies capable of addressing such data; it is only in recent years that advances in the field have improved these deficiencies. The research presented here adopts an alternative approach using a large cross-sectional data set that is not without its limitations, but still manages to produce specific parameters for selection and influence some of which are surprisingly similar to those reported in some recent work on this topic.

Inferences to describe adolescent networks are drawn from partially-formed ego-network data contained in the 1998 and 1999 survey years of the National Survey on Drug Use and Health; a modest level of precision in these analyses is achievable thanks to the large sample size. A custom Poisson/binomial/multinomial mixture is employed to extract precise peer network properties from ordinal response data having categories of proportions which implicitly cover the [0,1] interval. In order to understand the transitions in these network properties in tandem with the changing levels of substance use, from one age to the next, I exploit 1) the monotonic relationships between age and both peer network size and age-specific substance use, and 2) the distinct nature of an adolescent's peer group at the time of his or her first use of a substance, or initiation. The ego-network parameters also allow for the construction of network distributions: whole networks representing small hypothetical populations of adolescents, say a grade or an entire school. These distributions can serve as more than mere curiosities when used as a basis for a dynamic model of network change and substance use. However, this work stops short of constructing such a model. Instead, as an epilogue, and also as a prelude to future research, I address potential causation by calculating inter-latencies between substance use and certain factors, such as attachment to parents, which are relevant to prevention and intervention strategies.

Acknowledgments

Research Support

This work was supported in part by the Center for the Computational Analysis of Social and Organizational Systems (CASOS), the National Defense Science and Engineering Fellowship, the Department of Social and Decision Sciences, the Three Rivers Network, and the following grants: NSF GERT 9354995, NSF SES-0452487 and 0452598, NSF IGERT (CASOS) DGE-9972762, and ONR ONDCP N00014-06-1-0104. Any opinions, findings, and conclusions expressed in this material are those of the author and do not necessarily reflect the views or policies of the National Science Foundation or the United States government.

Special Thanks

Having been immersed in the graduate experience for longer than I am willing to admit, I have come to realize that an advisor, who not only appreciates her student's need to infuse his work with a modicum of cleverness and precision, but who also has such patience and generosity as to allow him the freedom to take some necessary missteps, is most certainly a rarity. For that, and more, I wholeheartedly thank my advisor Kathleen Carley. Balancing this source of support is a pair of giants whose critical minds were instrumental in honing my own and responsible for much improvement in this work since its birth. Though I regret not having availed myself more of their talents, I am enormously grateful to Jonathan Caulkins and Shelby Stewman for their time and perspective.

There is very little in this life that can compare to the company of those who can easily share in one's pain and pleasure. My dear friends Tiffany DeFoe, Keith Hunter, Andrew Petersen, Elisabeth Ploran, Danae Hoose, Eric Stone, Diane Lavsa, Ethan Shayne, Melissa Chan, and Joanna Tamburino have, in their own special ways, supplied me with just the intellectual and emotional support I needed to keep believing in myself. Special thanks also go to Stacey Ackerman-Alexeeff and Darren Homrighausen for being my source for all things LaTeX and assuring me that my methods might not be entirely crazy. I am also grateful to the staff of the Crazy Mocha Shadyside for being a source of comfort, when I often needed to drown my

dissertation frustrations in a half-caf americano.

Finally, no statement of gratitude is ever complete without an acknowledgment of those people who had first rights in shaping one's thoughts and character. Without my family's unwavering confidence in my choices, this journey would have been far more difficult; and I share the joy and satisfaction of completing this work with them.

Dedication

To my father, an almost lifelong smoker who found the will to quit.

To my nieces, Dylann, Sean, and Corey, with the hope that my work, and others like it, will benefit their lives.

And to my Lucy September, who is finding her way back into the race.

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

Why does a teenager start to smoke? And why do smoking teens seem to "hang out" with only other smoking peers? For most Americans familiar with our culture of substance use, these questions seem rhetorical and elicit high school memories of either clandestine meetings with friends at safe smoking locations or eyeing such groups from a distance with a modicum of envy or disdain. Still, a general acceptance of the culture of substance use has given rise to social norms, even protective ones; for instance, in my high school, students would willingly identify themselves to smokers hiding in bathrooms by uttering the key phrase "it's cool" upon entry.

The question of what exactly explains this agglomeration of adolescent substance users, of tobacco and other products such as alcohol and marijuana, continues to challenge researchers who have an interest in seeing a reduction in the levels of substance use among youth. It is often the case that we observe groups of people conspicuously sharing one or few common traits, be they behavioral, attitudinal, or sociodemographic ones. In sociological studies, this phenomenon is called 'homophily', a broad term that describes people's tendencies to seek out others who are similar to themselves. While it is tempting to simply admit clusters of people to be a foregone conclusion, especially since we see and participate in them so often in social life, the mechanisms involved in their assembly can be far subtle and not easily identifiable, despite having been under scientific scrutiny from the early 20th century (Almack, 1922; Bott, 1928; Wellman, 1929; Hubbard, 1929; Hagman, 1933)[1] until present day, when recent advances in network methodologies have provided investigators a precise means for measuring the extent of homophily in a spectrum of social contexts, including adolescent substance use. Some notable sociological research into interpersonal homophily has surveyed organizations (McPherson et al., 1992, 2001), communities (Verbrugge, 1977), and even the entire population of the U.S. (Burt, 1984, 1985; Marsden, 1987).

Over the past three decades, researchers have focused on two mechanisms believed to be mostly responsible for the prevalence of the substance use among adolescents.

[1] Most of these early studies looked at friendship choices among young school children.

For one, homophily, generally called ‘selection’ in the substance use literature, implies substance users seek out fellow substance users. Alternatively, this homogeneity in use behavior in peer groups can be the result of something else: non-users over time trying a substance for the first time as a result of some kind of social influence. Social or peer influence, commonly known as ‘peer pressure’ in lay writings, entail the change of one’s opinions, attitudes, or behaviors such that they ultimately conform to those of others, or ones others are perceived to maintain. Research into peer influence not surprisingly has its early roots in social psychology (Festinger, 1950, 1953, 1954) with some formalization recently developed in the neighboring field of sociology (Friedkin, 1990; Friedkin and Johnsen, 1990; Friedkin, 1998). Influence is considered to play a prominent role in all aspects of substance use, especially when a teen tries a drug for the first time; this event is known as ‘initiation’. Even when influence is identified as the culprit mechanism, we still might want to specify the brand of influence responsible, which can range from being a subtle inducement or opportunity to mimic friends’ behavior to being an imposition when such behavior attains a normative status and non-conformity becomes tantamount to expulsion. Alternatively, an adolescent might selectively shift the composition of his or her friendship circle in order to reduce the dissonance that arises when an inclination to try a substance or continue its use is frowned upon. Failure to achieve this change, for instance with parents, can render the adolescent susceptible to the anti-substance influence.

Some early studies, conducted in the late 1970s when network methodologies were in their infancy, offer findings that point to peer influence as the overriding determinant of substance use homogeneity within adolescent social groups (Kandel, 1978a; Jessor and Jessor, 1978; Akers et al., 1979; Brook et al., 1983), while other work at the time considers selection to have much larger role. Kandel (1978b) suggests that this homogeneity, in specifically marijuana use, is due to both influence and selection almost equally, while Cohen (1977) claims homophilic selection accounts for group membership much more so than the pressure to conform. More recent work, by Bauman and Ennett (1996), alerts us to a bias for over-estimating the influence effect, and their findings echo those of Cohen (1977): that selection plays a more prominent role. Some research observe how commonly adolescents first try a substance in the company of using friends, highlighting the preponderance of influence-induced initiation (Hahn et al., 1990; Kirke, 2004a,b). Alternatively, selection of friends can occur through other channels (e.g., similarities in music tastes (Steglich et al., 2006a) or involvement in sports activities (Pearson et al., 2006)), and once a relationship is established, substance use influence (to initiate or to continue using) takes hold.

Contained in all these studies is a framework for quantifying and analyzing peer relations and behaviors that falls under the class of study broadly known as social network analysis (Wasserman and Faust, 1994). The categories of network data cover the gamut from the simplest, sets of dyadic relations, to ego-centric or ego-networks, to snowball, and other link-trace, samples to complete networks, all of which qualities are enhanced by multiple collections over time, bestowing them the status of longitudinal data. While sample sizes can strongly determine the precision of estimates, which is generally the case for quantitative methodologies, the wide range of scope in the ways we can measure social relations obliges us to consider how well the data addresses inter-dependencies inherent in almost all social structures. Not surprisingly, earlier network studies, including the aforementioned substance use articles, report findings on just dyadic relations, in which the measures of interest are restricted to activities between pairs of individuals; for example, Kandel (1978b) looks at pairs of best friends. Since then, the methodologies and the quality of network data have co-evolved.

Ego-centric networks, or simply ego-networks, comprise relational data from the perspective of the respondent, called the 'ego', and include tie data from the respondent to his or her immediate 'alters', those individuals in the respondents' lives who qualify into one or more social categories requested by the researcher, such as friend, co-worker, etc. A well-known, nationally sampled ego-centric study is the Social Network Module of the 1985 General Social Survey. This portion of the survey requested data on people 'with whom, in the last six months, you discussed an important personal matter' in an attempt to obtain intimate, confidante relationships (Burt, 1984, 1985; Marsden, 1987). In studies that employ ego-centric networks, researchers generally collect, in addition to the ego-alter relations, personal information, such as sociodemographic and behavioral data which are key in understanding the nature of homophilic relationships; normally, the respondent provides his or own own data and also that of their alters. Sometimes, perceived alter-to-alter relations are requested from the ego. These triangular or triadic sets of relations are of particular interest to network analysts. Structures that reflect or induce balance, transitivity, equivalence, and small-world clustering highlight the substantial complexity inherent in triadic relationships. Since it requires only a marginal amount of additional effort to collect ego-centric data, over dyadic data, they are more commonly collected and studied; some recent substance use studies that favor the use of ego-networks include work by Gainey et al. (1995) and Ellickson and Bell (1990).

In the various brands of link-trace, or chain referral, sampling designs, such as snowball sampling (Frank, 1977, 1979; Frank and Snijders, 1994), respondents are selected by their affiliation to other respondents; hence these sampling methods result in chain-like structures and are often employed to access hidden or hard-to-access populations, such as needle-injecting hard substance users or sexual networks. Recent modifications have been made to address the bias introduced with this kind of sampling in a technique called respondent-driven sampling (Heckathorn, 1997; Heckathorn et al., 1999; Heckathorn, 2002, 2007; Salganik and Heckathorn, 2004). However, size for size, complete network data is most robust yet the most difficult to collect as it entails relational data between all actors from a pre-defined population. The closed system nature of complete networks restricts the size of the sampled population generally to a few hundred at most. However, it remains the most useful kind of network data for studying diffusive type events and behavior (Morris, 1993).

Like the majority of social science data, network data, especially the larger sets, tend to be cross-sectional. Two major data sources for adolescent substance use in the United States do contain some network data: the Monitoring the Future (MTF) and the National Survey on Drug Use and Health (NSDUH), formerly known as the National Household Survey on Drug Abuse (NHSDA). However, as expected, the quality of their data is restricted given the enormous sample sizes (10,000+) per survey collection year; while the array of survey items is comprehensive, the network data is ego-centric and only partial in that: they lack alter-alter ties and responses are not given in exact quantities, but rather ordinal categories of substance use among the respondents' friends.

The recognition of limitations in cross-sectional data has fueled efforts into collecting longitudinal data and developing appropriate dynamic network analysis, with which researchers have been making statements on the selection/influence dynamic. Notable work of this type include studies by Bearman et al. (2004); Pearson et al. (2006); Hall and Valente (2007). Still, other studies continue to exploit the statistical advantages afforded by the large, cross-sectional, substance use data sets (Everingham and Rydell, 1994; Caulkins et al., 1999; Caulkins, 2000a; Caulkins et al., 2004). However, the dynamic models in these works do not differentiate structure or specify interactions between individuals.

With this dissertation, I aspire to contribute to the ongoing efforts into understanding adolescent substance use, by applying a novel inference technique on partial ego-network data contained in one of the large, nationally sampled data sets and then using derived estimates to triangulate selection/influence coefficients. The first section of this dissertation introduces the probability model, a Poisson/binomial/multinomial mixture, used to extract precise estimates about the size of adolescent peer networks and the number of those peers who use substances from ordinal response data having categories of proportions which implicitly cover the [0,1] interval; for those first analyses, I focus on a single substance, tobacco (i.e. cigarette smoking). The second section exploits these ego-network parameters in generating distributions of complete networks and highlights some network measures of interest to substance use and network researchers alike. The third section expands the probability model to incorporate joint analysis on ego-network data for two and then three substances. However, due to technological limitations, a full suite of analyses cannot be performed; the time required for each estimation is extremely long, and the sample size is not large enough to support some of the joint analyses. Next, peer network estimates from adolescents who have recently initiated are inferred and, when combined with estimates on the non-using population, they give us measures for the risk of initiating. These initiation parameters are then used to predict transitions in ego-network properties for both users and non-users alike, from which I draw precise estimates about influence and selection and compare them with estimates from two other recent adolescent substance use studies. Finally, I perform a relatively naïve analysis on age-based curves which I treat as temporal, so that I might infer ordering and latencies between substance use and some covariates relevant to issues of prevention and intervention.

Chapter 2

Inferring Parameters of Adolescent Ego-Networks

2.1 Data Source

Large sets of substance use network data are generally ego-centric and cross-sectional. While efforts in the past decade have focused on collecting complete, longitudinal networks, their samples sizes tend to be smaller. An earlier study by Kirke (1996) and a recent study conducted by Hall and Valente (2007) are relatively small, where sample size n is 267 and 880, respectively. The sample size of the social networks in National Longitudinal Study of Adolescent Health (Add Health) (Bearman et al., 2004) is of moderate size, $n = \sim$15,000, with complete networks comprising entire schools. Current non-complete/ego-centric, cross-sectional substance use networks are on the order of 25,000 for the National Survey on Drug Use and Health (NSDUH), formerly known as the National Household Survey on Drug Abuse (NHSDA) and ∼250,000 for the Monitoring the Future study (MTF). The data from each of these studies have different limitations and strengths. Kirke's data is relatively small, comes from a discrete geographic area in Dublin County Ireland having a population of 2,500, and though complete, the data is static (i.e. collected at only one time point), so no dynamic inferences can be drawn. The data from the Add Health and Hall/Valente studies are longitudinal, permitting some dynamic analysis, but the investigators restrict friendship size to maximum of five, with friends' substance use in Add Health being limited to just three best friends, much in the same way earlier studies limited the number of friends (e.g. Ennett and Bauman (1993)). Furthermore, Hall and Valente only report on cigarette smoking behavior and, like Kirke's data, their population is geographically bound, to several areas of Los Angeles county.

As for the larger data sets, both the NSDUH and the MTF offer only ego-centric data in the form of broad ordinal categories of friends' substance use, such as 'None', or 'Few', or 'All'. Furthermore, while the sample size of the MTF study seems sufficiently large, adolescents are sampled at only three grade levels (8th, 10th, 12th) limiting the granularity of dynamic statements we can make. And, because it employs five response categories of friends' use, rather than four of the NSDUH, the network parameter estimation procedure is further complicated. Despite the smaller sample size of the NSDUH, its broader age range and simplifying ego-network response categories makes it the candidate data source, one from which we can draw age transition inference. Furthermore, the NSDUH includes partial ego-network data on adults' substance use, information that is relevant in the prediction of adolescent use but not included in the other data sets.

The NSDUH is funded by the Substance Abuse and Mental Health Services Administration (SAMHSA) and informs the Federal government on the use of alcohol, tobacco, and various illicit substances. While survey administration commenced in 1971, the earliest publicly available data is the 1979 survey year. Since 1990, the survey has been conducted yearly and samples individuals ages 12 and older who live in households and over-samples 12–17 year olds; the survey does not sample the prison or homeless population. Surveys are answered in the privacy of each respondent's home with most of the computerized responses answered completely in private. The survey takes approximately one hour to complete, and respondents are compensated $30 in cash for their participation. Confidentiality is emphasized through the absence of records on participants' names and through protection under the Confidential Information Protection and Statistical Efficiency Act of 2002.

Some concerns about the accuracy of sensitive self-report data have been allayed with validity studies conducted in 2000 and 2001 (Harrison et al., 2007). For tobacco, there was 84.6% agreement between self report in the past 30 days and urine test results. About 5.8% reported no use and tested positive and 9.6% reported use in the past 30 days and did not test positive. For marijuana, there was 89.8% agreement between self report in the past 30 days and urine test results. About 4.4% reported no use and tested positive and 5.8% reported use in the past 30 days and did not test positive.

In the 1998 and 1999 NSDUH survey years, respondents were asked to state how many of their friends smoked, consumed alcohol, and used marijuana. These egocentric response items will be used to construct a distribution of complete networks, necessary for making dynamic statements on influence and selection. While these measures convey respondents' perceptions of friends' use, D'Amico and McCarthy (2006); Iannotti and Bush (1992); Iannotti et al. (1996); Kawaguchi (2004); and Olds et al. (2005) demonstrate that perceived peer substance use behavior is just as, and often more predictive, of respondents' use than the actual peer use behavior.

This work will solely employ youth respondent data (i.e. 12–17 year olds) because a) this population is oversampled, and hence, their data are more robust and b) the youth-related covariates in the NSDUH important to intervention or prevention strategies exist only for that age range in our data.[1] Furthermore, earlier initiation, in

[1] Another reason to focus on this age range is that the majority of these youths live in households, as opposed to other types of residences such as dormitories or barracks. Also, the acquisition of

all of the substances examined, translates to a higher chance of persistent use through adulthood (Everett et al., 1999).[2] So, it behooves us to focus the analyses on those initiation ages.

2.2 Notation

$p(...)$ will refer to density of a distribution and will not necessarily be within the interval [0,1]. Pr{...} will refer to either the density of a discrete distribution or the probability of a discrete event; in either case, its referent values will be in the interval [0,1]. More often than not probabilities or likelihoods of models will tend to be extremely small. Hence, most of probability or likelihood results will be reported as their natural logarithms. The notations L, $\mathcal{L}$, and $\mathcal{L}_r$ denote, respectively, the likelihood, log-likelihood, and log of the likelihood ratio between a given model and the mode. The density function of a distribution, p, will be used directly as the likelihood:

$$\begin{aligned} L(\theta|y) &= p(y|\theta) \\ \mathcal{L}(\theta|y) &= \log[p(y|\theta)] \end{aligned}$$

The density p refers (but is not necessarily equal) to the probability of observing the data, y, given a hypothetical parameter, θ. In reversing the conditional, we obtain the likelihood of the parameter θ given the data y. While it is necessarily the case that density function sums to 1 (i.e. $\int_{-\infty}^{+\infty} p(y|\theta)d\theta = 1$ or $\sum_{\theta=-\infty}^{+\infty} p(y|\theta) = 1$), this does not hold true for the likelihood: $\sum_{y=-\infty}^{+\infty} L(\theta|y) \neq 1$.

The mode of some distribution x is denoted by a hat: $\hat{x}$. In most cases, this will be identical to its mean. In table headings, μ_x, or just some parameter x, refers to the mean of the distribution around x and σ_x will refer to its standard deviation.

The expression $(x_0, x_1, \ldots)$ refers to a tuple or vectors of quantities. The operations performed on it follow standard vector arithmetic. For instance $(x_0, x_1, \ldots)+(1, 2, \ldots)$ $= (x_0 + 1, x_1 + 2, \ldots)$. We name tuples with a parameter label in parentheses, e.g. $(n_{\mathrm{FDCIG}}) = (n_{\mathrm{None}}, n_{\mathrm{Few}}, n_{\mathrm{Most}}, n_{\mathrm{All}})$, or in bold-type, e.g. $\boldsymbol{\theta} = (\theta_{00}, \theta_{10}, \theta_{01}, \theta_{11})$.

2.3 Terminology

For the sake of brevity, covariates of interest, in this dissertation, will often be called by their formal, abbreviated names. The nomenclature of tags identifying a specific

tobacco ceases being illegal for teens over the age of 17.

[2] Probability of last month use is significantly lower for those respondents who initiated outside of the 9-12 year-old range than for those whose initiation age is within that range. In Appendix A, an analysis of peak initiation ages is provided along with evidence that shows early initiation incurs higher rates of use in adult years.

substance is:

When we see ...	It refers to some frequency or recency of ...
CIG	Cigarette Smoking
ALC	Alcohol Consumption
MRJ or MJ	Marijuana/Hashish Use
COC	Cocaine Use

Recency of use for each substance was obtained through similar response items in which respondents were solicited for the time frame in which they last used the above substances. For the naming of the recency of use indicator variables, the above tags are joined with the following suffices:

Have you ever used substance x in y?

where $x \in$ {CIG, ALC, MRJ, COC} and $y \in$

Informal	Formal
"your lifetime" or "ever"	xFLAG
"the past three years"	xRC3
"the past year"	xYR
"the past month"	xMON

The response to these variables is either a No or a Yes, indicating use within the stated time frame; these variables are converted to a binary indicator, 0 or 1.

Adolescent respondents were also asked to provide the degree of substance use among their friends and adults whom they know; the responses were not confirmed by the implicit alters and constitute the perceived ego-centric measures:

How many of x use substance y?

where $y \in$ {CIG, ALC, MJ} and $x \in$

x	Formal	Informal
"your friends"	FDy	"Friends' Use"
"adults that you know"	ADOy	"Adults' Use"

The response to these variables is one of {None, Few, Most, All}.

Respondents were also asked to provide their age at which they tried a substance, assuming they had initiated on that substance:

How old were you when you tried x for the first time?

where $x \in$ {CIG, ALC, MJ, COC}

x	Formal	Informal
CIG or ALC	xTRY	"initiation age" or
MJ or COC	xAGE	"age of first use"

	Have You Ever Used Cigarettes in ...				
		Past	Past	Past	
	Lifetime	3 Yrs	Year	Month	Recency
Intercept	-8.350***	-7.735***	-8.068***	-9.523***	-8.772***
	(0.158)	(0.163)	(0.183)	(0.227)	(0.149)
Is Male	0.161***	0.123ˆ	0.113	0.185	0.157**
	(0.031)	(0.032)	(0.035)	(0.041)	(0.028)
Age	0.323***	0.260***	0.237***	0.269***	0.297***
	(0.010)	(0.010)	(0.011)	(0.014)	(0.009)
Friends' Use	1.122***	1.224***	1.275***	1.416***	1.242***
	(0.023)	(0.024)	(0.025)	(0.029)	(0.021)
Adults' Use	0.375***	0.288***	0.315***	0.327***	0.352***
	(0.024)	(0.025)	(0.027)	(0.031)	(0.022)
n	24916	24916	24916	24916	24916
Pseudo-R^2	0.335	0.325	0.322	0.345	0.319
p value	0.000	0.000	0.000	0.000	0.000
AIC	25677	24247	20817	15855	47925
BIC	25718	24287	20857	15895	47990

Table 2.1: Predicting Recency of Smoking with Common Covariates with Binary (and Ordinal) Logistic Regression Models. *All models above employ the precision weighted data provided in the data. Significance and standard errors are drawn from unweighted fit due to an artificial shrinkage on these introduced by the precision weights. Significance is denoted by: ˆ= $p <$ 0.10, ** = $p <$ 0.01, *** = $p <$ 0.001. Lower values of AIC and BIC denote better fitting models.*[2]

2.4 Relevant Covariates

Before introducing the decomposition of the ego-centric measures, we supplement findings from prior research on substance use and peer networks by quickly confirming the relationship between substance use and network covariates among the NSDUH respondents; the network variables in these regressions will serve as the core of the network inference. For the length of this chapter, we will focus on just tobacco use (i.e. cigarette smoking) in order to easily present the methodology. Later, similar analyses will be performed on alcohol consumption and marijuana use.

According the regression results displayed in Table 2.1, friends' use dominates in the prediction of respondents' smoking across the indicators of recency as well as a composite ordinal recency variable.[4] The effect of a single jump in the category of

[2] The Akaike Information Criterion (AIC) is a goodness of fit measure, penalizing the likelihood by the number of parameters: AIC = $2k - 2\log(L)$ where k is the number of parameters (or predictors including intercept) and L is the likelihood. The lower AIC represents a better fitting model. The Bayesian Information Criterion (BIC) is similar to the AIC except it penalizes free parameters more strongly: BIC = $k\log(n) - 2\log(L)$, where n is the number of observations.

[4] Recency is an ordinal variable, which exclusively separates the previous four indicators: 0=Never

friends' use is four to seven times greater in magnitude, depending on the recency variable, than the increase in the probability of use incurred by aging a single year. Males are more prone to delinquent or non-sanctioned behavior due to a combination of biology and peer-influence, so gender is expected to be a modestly significant predictor; yet, its effect is dwarfed by the other predictors. With age, it is clear that being alive longer affords more opportunities to encounter smokers, and hence the opportunities to smoke simply increase as time passes; in other studies, a more appropriate hazard model is employed to predict the initiation age (Edelen et al., 2006). Hence, a cohort will increasingly include more smokers over time.

2.5 Probability Model

In this section, our probability model is built incrementally, and we incorporate additional parameters and other covariates when appropriate. To begin, we focus on the NSDUH ego-network response item for peer-group smoking, FDCIG:

How many of your friends smoke (FDCIG)?

	None	Few	Most	All
raw	9,865	10,626	3,804	756
weight adjusted	9,621	10,652	3,901	878

The *raw* row in the table shows how the 25,052 youth respondents between the ages of 12 and 17 answered the item,[5] while the *weight adjusted* row reports the same total respondents broken down into same categories, but proportionally to the precision weights.[6] From now on, this work will refer to the *weight adjusted* data, unless otherwise noted. A facile calculation reveals that a majority (58%) of the adolescent respondents had at least one friend who smoked.

2.5.1 Simple Binomial Model

If it was the case that we knew exactly the proportion of an adolescent's friends who smoke, we could start inferring the prevalence of smoking among their peers, with a

Use, 1=Within Lifetime, but not within last 3 years, 2=Within last 3 years, but not within last year, 3=Within last year, but not in last month, 4=In last month. Appropriately, we employ an ordinal logistic regression. To save space, the mean of the four Intercepts is reported.

[5] While the 1998 and 1999 surveys contain a total of 25,463 youth respondents, 412 have missing FDCIG data.

[6] We sum the NSDUH's respondent weight variable (ANALWT) for each response category, determine their percentages, and multiply by the sample size.

ANALWT	None	Few	Most	All
sum	17,377,662	19,239,907	7,045,315	1,585,141
prop.	0.38405	0.42521	0.15570	0.03503
×25052 =	9,621	10,652	3,901	878

simple binomial model:

$$n_{smoke} \sim \text{Binomial}(n_{friends}, \theta)$$

In this simple model, we assume, unrealistically, that all respondents have exactly $n_{friends}$ and the probability that a single one of their friends smokes is θ. The number of friends who smoke, n_{smoke}, is then distributed binomially. More specifically, the probability that a youth has x friends who smokes is

$$\Pr\{n_{smoke} = x\} = \binom{n_{friends}}{x} \theta^x (1-\theta)^{(n-x)}$$

Say a youth has five friends ($n_{friends} = 5$) and is embedded in a social world in which a quarter of all peers smoke ($\theta = .25$), then the probability that two of those friends smoke ($n_{smoke} = 2$) is $\binom{5}{2} \cdot 0.25^2 \cdot 0.25^{(5-2)} \approx 0.264$. Expanding this further, we compute the probabilities that 0, 1, 2, 3, 4, or 5 of his or her friends smoke, rounded to four decimal places; the results are 0.2373, 0.3955, 0.2637, 0.0879, 0.0146, and 0.0010.

Unfortunately, the data reveals neither exactly how many friends smoke nor how many total friends a respondent has; these are crucial pieces of information necessary for analyzing friendship networks and later inferring how non-users and users might be connected. Instead, we can map the four response categories to the space of possibly using friends, and the accompanying probabilities, offered by the binomial model. For example, if a respondent has five friends and answers 'None', then we know that none of those five smoke. It is also reasonable to assume that if the respondent answered 'Few', we can say that one or two friends smoke. 'Most' implies three or four smoking friends, and 'All' means all five. At this stage, we ignore reporting error; for example, a respondent may actually have only four out of five friends who smoke, but in misremembering, answers the query with 'All' or 'Few'. Also, for now, we ignore other interpretations of 'Few' and 'Most'; for example, it is possible that some respondents regard 3 out of 5 as 'Few'.

The probability associated with each of the responses to the NSDUH survey question "How many of your friends smoke?" (FDCIG) can be now calculated. Given $\theta = 0.25$ and $n_{friends} = 5$, the probabilities that a youth would respond to the FDCIG item with 'None', 'Few', 'Most', or 'All' are:

$$\begin{aligned}
\Pr\{y = \text{`None'}\} &= \Pr\{n_{smoke} = 0\} \\
&= 0.2373\ldots \\[1em]
\Pr\{y = \text{`Few'}\} &= \Pr\{n_{smoke} = 1 \text{ or } n_{smoke} = 2\} \\
&= 0.3955\ldots + 0.2636\ldots \\
&= 0.6591\ldots \\[1em]
\Pr\{y = \text{`Most'}\} &= \Pr\{n_{smoke} = 3 \text{ or } n_{smoke} = 4\} \\
&= 0.0878\ldots + 0.0146\ldots \\
&= 0.1025\ldots \\[1em]
\Pr\{y = \text{`All'}\} &= \Pr\{n_{smoke} = 5\} \\
&= 0.0009\ldots
\end{aligned}$$

Next, we seek the likelihood of the tabulated FDCIG data fitting our parameters, $\theta = 0.25$ and $n_{friends} = 5$; that is, how likely is it that these parameters are correct given our data? The friends' use data is expressed notationally as:

$$n_{\text{None}} = 9621,\ n_{\text{Few}} = 10652,\ n_{\text{Most}} = 3901, \text{ and } n_{\text{All}} = 878$$

or equivalently:

$$(n_{\text{FDCIG}}) = (9621, 10652, 3901, 878)$$

Since these probabilities determine how a set of items fall into four categories, the appropriate likelihood distribution is the multinomial:[7]

$$\begin{aligned}
&\Pr\{(n_{\text{FDCIG}}) = (9621, 10652, 3901, 878)|\theta = 0.25, n_{friends} = 5)\} \\
&\quad = \binom{n_{\text{None}} + n_{\text{Few}} + n_{\text{Most}} + n_{\text{All}}}{n_{\text{None}}\ n_{\text{Few}}\ n_{\text{Most}}\ n_{\text{All}}} \cdot \Pr\{y = \text{`None'}\}^{n_{\text{None}}} \\
&\qquad \cdot \Pr\{y = \text{`Few'}\}^{n_{\text{Few}}} \cdot \Pr\{y = \text{`Most'}\}^{n_{\text{Most}}} \cdot \Pr\{y = \text{`All'}\}^{n_{\text{All}}} \\
&\quad \approx \binom{25052}{9621\ 10652\ 3901\ 878} \cdot 0.2373^{9621} \cdot 0.6592^{10652} \cdot 0.1025^{3901} \cdot 0.0010^{878} \\
&\quad \approx e^{-4749.1} \text{ or } \exp(-4749.1)
\end{aligned}$$

[7] Multinomial distribution: The probability that n items will fill m possible categories, such that there are θ_1 items in the first category, θ_2 items in the second, and so forth given that the probability for an item to appear in the i-th category is p_i, is:

$$p(\theta) = \binom{n}{\theta_1 \cdots \theta_m} p_1^{\theta_1} \cdots p_m^{\theta_m}, \text{ where } \sum_{i=1}^m \theta_i = n,\ \sum_{i=1}^m p_i = 1, \text{ and } \binom{n}{\theta_1\ \theta_2 \cdots \theta_m} = \frac{n!}{\theta_1! \theta_2! \cdots \theta_m!}$$

Since the actual probability/likelihood here is extremely small, its log is reported, $\mathcal{L} = -4749.10$.[8] We can compare this result with the best log-likelihood obtainable, which occurs at the mode of the distribution: the set of probabilities $(n_{\text{FDCIG}})/\Sigma(n_{\text{FDCIG}}) = (0.384, 0.425, 0.156, 0.035)$.[9] This best log-likelihood is $\hat{\mathcal{L}} = e^{-14.44}$, leaving our model $e^{(-14.44--4749.10)}$ or $e^{4734.67}$ times worse than some idealized model that produces the modal probabilities:

$$\begin{aligned}
\mathrm{L}(\theta = 0.25|n_{friends} = 5, y_{\text{FDCIG}}) &= e^{-4749.10} \\
\mathcal{L}(\theta|y) = \log(\mathrm{L}(\theta|y)) &= -4749.10 \\
\mathcal{L}_r(\theta|y) = \log\left(\frac{\mathrm{L}(\theta|y)}{\hat{\mathrm{L}}}\right) = \mathcal{L}\theta|y) - \hat{\mathcal{L}} &= -4749.10 - -14.44 \\
&= -4734.67
\end{aligned}$$

Indeed, the fit is not so stellar. If we want to know the best fitting θ for $n_{friends} = 5$, we would derive the maximum likelihood estimator of the mode using the first derivative of the log likelihood. First, the likelihood as a function of θ:

$$\begin{aligned}
&p(y = (n_{\text{FDCIG}})|\theta) = \mathrm{L}(\theta|y = (n_{\text{FDCIG}})) = \\
&\quad \binom{n_{\text{None}} + n_{\text{Few}} + n_{\text{Most}} + n_{\text{All}}}{n_{\text{None}}\ n_{\text{Few}}\ n_{\text{Most}}\ n_{\text{All}}} \cdot \\
&\quad \left[\binom{5}{0}\theta^0(1-\theta)^5\right]^{n_{\text{None}}} \cdot \left[\binom{5}{1}\theta^1(1-\theta)^4 + \binom{5}{2}\theta^2(1-\theta)^3\right]^{n_{\text{Few}}} \cdot \\
&\quad \left[\binom{5}{3}\theta^3(1-\theta)^2 + \binom{5}{4}\theta^4(1-\theta)^1\right]^{n_{\text{Most}}} \cdot \left[\binom{5}{5}\theta^5(1-\theta)^0\right]^{n_{\text{All}}}
\end{aligned}$$

We calculate the log of the likelihood, ignoring the constant which will become zero in the derivative:

$$\begin{aligned}
&\mathcal{L}(\theta|y = (n_{\text{FDCIG}})) = \log(\mathrm{L}(\theta|y = (n_{\text{FDCIG}})) = \\
&\quad \text{constant} + n_{\text{None}} \cdot 5\log(1-\theta) + n_{\text{Few}} \cdot \log(5\theta(1-\theta)^4 + 10\theta^2(1-\theta)^3) \\
&\quad + n_{\text{Most}} \cdot \log(10\theta^3(1-\theta)^2 + 5\theta^4(1-\theta)) + n_{\text{All}} \cdot \log(\theta^5)
\end{aligned}$$

[8] Incidentally, this likelihood value was calculated using the actual, non-rounded $\Pr\{y = \text{'None'}\}$, $\Pr\{y = \text{'Few'}\}$, etc., and not the rounded probabilities in the displayed equation.

[9] These probabilities are obtained from the data itself:

$$\begin{aligned}
p_{None} &= 9621/(9621 + 10652 + 3901 + 878) &\simeq 0.384, \\
p_{Few} &= 10652/(9621 + 10652 + 3901 + 878) &\simeq 0.425, \\
p_{Most} &= 3901/(9621 + 10652 + 3901 + 878) &\simeq 0.156, \\
p_{All} &= 878/(9621 + 10652 + 3901 + 878) &\simeq 0.035.
\end{aligned}$$

To obtain the mode, we first compute the first derivative of the log-likelihood:

$$\begin{aligned}\mathcal{L}'(\theta) = \frac{d\mathcal{L}}{d\theta} &= n_{\text{None}} \cdot \frac{5}{1-\theta} + n_{\text{Few}} \cdot \frac{5(1-\theta)^4 - 30(1-\theta)^2\theta^2}{5(1-\theta)^4 - 10(1-\theta)^2\theta^2} \\ &\quad + n_{\text{Most}} \cdot \frac{30(1-\theta)^2\theta^2 - 5\theta^2}{10(1-\theta)^2\theta^3 + 5(1-\theta)\theta^4} + n_{\text{All}} \cdot 5\theta\end{aligned}$$

We set $(n_{\text{None}}, n_{\text{Few}}, n_{\text{Most}}, n_{\text{All}})$ to $(9621, 10652, 3902, 878)$, and then set the first derivative $\frac{dL}{d\theta}$ to 0. When we solve for θ, we obtain $\theta = 0.252274$.[10] The log-likelihood $\mathcal{L}(\theta = 0.252274|(n_{\text{FDCIG}}))$ is -4747.60, which is $\exp(-4747.60 - -4749.10)$ or 4.48 times a better fit than our earlier estimate of $\theta = 0.25$.

We also need to know the uncertainty surrounding our estimate; the standard deviation (s.d.) of the mode can be obtained by calculating the second derivative at the mode of θ (i.e. $\hat{\theta}$):

$$\begin{aligned}\mathcal{L}''(\theta) &= n_{\text{None}} \cdot \frac{5}{(1-\theta)^2} + n_{\text{Few}} \cdot -\left(\frac{5(1-\theta)^4 - 30(1-\theta)^2\theta^2)^2}{5(1-\theta)^4\theta + 10(1-\theta)^3\theta^2)^2}\right) \\ &\quad + n_{\text{Most}} \cdot \frac{-20(1-\theta)^3 - 60\theta(1-\theta)^2 + 60\theta^2(1-\theta)}{5(1-\theta)^4\theta + 10(1-\theta)^3\theta^2} + n_{\text{All}} \cdot 5 \\ \mathcal{L}''(\hat{\theta}) &= -578485.1, \text{ when we substitute } \theta \text{ with } \hat{\theta}\end{aligned}$$

The second derivative gives us an estimate of the covariance matrix,[11] $[-\mathcal{L}''(\hat{\theta})]^{-1} = V_{\hat{\theta}}$, in this case just the variance; then, the s.d. of θ is 0.001314782, which is quite small. If it was true that all youths had exactly five friends, then it is most likely that just over quarter of their friends smoke or the probability that a single friend smoke is roughly 0.2522 ± 0.0026 (or two s.d.'s). If we were restricted to selecting a single constant for $n_{friends}$ for all respondents, we might select a different value. But, first, we need to generalize our interpretation of the response categories.

For some chosen $n_{friends}$,

	Possible values for n_{smoke}	
Response to FDCIG	Minimum	Maximum
None	0	0
Few	1	$\lfloor n_{friends}/2 \rfloor$
Most	$\lfloor n_{friends}/2 \rfloor + 1$	$n_{friends} - 1$
All	$n_{friends}$	$n_{friends}$

[10] The numerator of the final polynomial is $1337250\theta^4 - 9780650\theta^5 + 215149000\theta^6 - 118666000\theta^7 - 188733350\theta^8 + 312981500\theta^9 - 167612000\theta^{10} + 31315000\theta^{11}$. The other roots are -0.8624722, 0, $0.9998713 \pm 0.0001287i$, $1.0001287 \pm 0.0001287i$, and 1.9626492, none of which are both non-imaginary and lie in between 0 and 1.

[11] The inverse of the negative second derivative at the mode yields the covariance matrix of a multivariate normal approximation, or just the univariate variance in the single parameter model.

			Posterior quantiles for θ				
$n_{friends}$	$\mathcal{L}(\hat{\theta})$	$\hat{\sigma}$	2.5%	25%	mode	75%	97.5%
3	-118.12	0.00164	0.277	0.279	0.281	0.282	0.284
4	-3696.36	0.00154	0.297	0.299	0.300	0.301	0.303
5	-4747.60	0.00131	0.250	0.251	0.252	0.253	0.255
6	-10729.79	0.00129	0.265	0.266	0.267	0.268	0.270
7	-12104.73	0.00114	0.234	0.236	0.237	0.237	0.239
8	-18880.34	0.00114	0.246	0.248	0.248	0.249	0.251
9	-20306.83	0.00104	0.225	0.226	0.227	0.228	0.229
10	-27426.21	0.00103	0.235	0.236	0.237	0.237	0.239

Table 2.2: *Log-likelihood and other statistics for best θ under a fixed $n_{friends}$ binomial model.*

Essentially, a respondent who offered 'None' is claiming that zero friends smoke. If the respondent marked 'Few', then s/he is telling us, under basic assumptions, that the number of smoking friends lies between one and half the total number of friends. 'Most' means that the number of smoking friends is somewhere between just over half and just one less than the total number of friends. Finally, 'All' implies that the number of smoking friends is equal to the number of friends s/he has in mind.

According to Table 2.2, the best parameters are $n_{friends} = 3$ and $\theta = 0.281$, with a log-likelihood of -118.12, significantly better than any likelihood under a different $n_{friends}$. If we had to work with a single value for total friends, we would pick $n_{friends} = 3$ and assume that the chances of a friend smoking is significantly more (in the statistical sense) than 0.252. At this point, we might interpret $\theta = 0.281$ to be the proportion of youths, of ages 12–17, who smoke; this is a population-level statistic. For comparison, we offer, from the recency of use data, the proportions of respondents who claimed to have smoked at some point in their lives, in the past three years, in the past year, and in the past month.[12]

Have You Smoked ... ?	No	Yes	p(Yes)	$p_{unweighted}$(Yes)
Ever	16182	9281	0.364	0.364
In Past 3 Years	17564	7899	0.310	0.309
In Past Year	19456	6007	0.236	0.230
In Past Month	21248	4215	0.166	0.156

Comparing this data to the θ results in Table 2.2, inferred under the naïve binomial model, we conclude for now that the practical definition of perceived "smoking" lies somewhere between past year and past three year use; the majority of the posterior θ distributions lies between 0.236 and 0.310. That is, when a respondent thinks of a

[12] The variable names for these categories are CIGFLAG, CIGYR, and CIGMON. Also, these categories overlap; a youth who claimed to have smoked in the past year would be also counted as having smoked once in their lifetime.

smoking friend, that friend is perceived to have smoked within the past three years or within the past year.

2.5.2 Poisson/Binomial Mixture Model

Since it is unrealistic to assume all teens have exactly the same number of friends, it behooves us to enhance the model and allow $n_{friends}$ to arise from some discrete random distribution:

$$\begin{aligned} n_{friends} &\sim X \text{ (i.e. some discrete distribution)} \\ n_{smoke} &\sim \text{Binomial}(n_{friends}, \theta) \end{aligned}$$

The expression for each of the response categories now expands to:

$$\begin{aligned} \Pr\{y = \text{`None'}\} &= \sum_{i=0}^{\infty}\left(\Pr\{n_{friends} = i\} \cdot \binom{i}{0}\theta^0(1-\theta)^i\right) \\ \Pr\{y = \text{`Few'}\} &= \sum_{i=2}^{\infty}\left(\Pr\{n_{friends} = i\} \cdot \sum_{j=1}^{\lfloor i/2 \rfloor}\binom{i}{j}\theta^j(1-\theta)^{(i-j)}\right) \\ \Pr\{y = \text{`Most'}\} &= \sum_{i=3}^{\infty}\left(\Pr\{n_{friends} = i\} \cdot \sum_{j=\lfloor i/2 \rfloor+1}^{i-1}\binom{i}{j}\theta^j(1-\theta)^{(i-j)}\right) \\ \Pr\{y = \text{`All'}\} &= \sum_{i=1}^{\infty}\left(\Pr\{n_{friends} = i\} \cdot \binom{i}{i}\theta^i(1-\theta)^0\right) \end{aligned}$$

Note that for 'Few', we consider only $n_{friends} \geq 2$; for 'Most', $n_{friends} \geq 3$; and for 'All', $n_{friends} \geq 1$. These categories are undefined for values below their respective minimums; e.g. in order to have been able to respond with 'Few', a respondent needed to have at least 2 friends in mind, or else the only applicable categories would have been 'None' or 'All'. For our discrete distribution X, we choose the Poisson distribution, which is commonly used to model count data.[13]

$$\begin{aligned} n_{friends} &\sim \text{Pois}(\lambda) \\ n_{smoke} &\sim \text{Binomial}(n_{friends}, \theta) \end{aligned}$$

A new parameter λ refers to the mean number of friends among the population of respondents; the variance of a Poisson is identical to its mean; hence, the Poisson is a single parameter distribution. The functional form is as follows; we abbreviate $n_{friends}$ as just n:

$$\Pr\{y = \text{`None'}\} = \sum_{n=0}^{\infty}\left(\frac{\lambda^n e^{-n}}{n!} \cdot \binom{n}{0}\theta^0(1-\theta)^n\right)$$

[13] The Poisson distribution is more appropriate than the binomial in predicting number of friends. In Appendix B, we justify its use by fitting to several empirical distributions of friend counts. Later, in this chapter, we will explore other candidate prior distributions such as the negative-binomial.

$$
\begin{aligned}
\Pr\{y = \text{`Few'}\} &= \sum_{n=2}^{\infty}\left(\frac{\lambda^n e^{-n}}{n!}\cdot\sum_{m=1}^{\lfloor n/2\rfloor}\binom{n}{m}\theta^m(1-\theta)^{(n-m)}\right)\\
\Pr\{y = \text{`Most'}\} &= \sum_{n=3}^{\infty}\left(\frac{\lambda^n e^{-n}}{n!}\cdot\sum_{m=\lfloor n/2\rfloor+1}^{n-1}\binom{n}{m}\theta^m(1-\theta)^{(n-m)}\right)\\
\Pr\{y = \text{`All'}\} &= \sum_{n=1}^{\infty}\left(\frac{\lambda^n e^{-n}}{n!}\cdot\binom{n}{n}\theta^n(1-\theta)^0\right)
\end{aligned}
$$

While the modes $\hat{\lambda}$ and $\hat{\theta}$ can be obtained using *conditional maximization*, that method does not give us the variance around the modes. Instead, we employ the *Newton-Raphson* algorithm which requires calculations of the first and second derivatives for each parameter; given proper starting points, the algorithm will converge on the mode and covariance matrix of a unimodal, multivariate normal distribution.[14] As earlier, we attempt to compute these derivatives analytically, but now find that the process quickly becomes unwieldy, even for low values of λ. For example, say we want to solve θ for $\lambda = 0.08$ (or any other λ for which the maximum $n_{friends}$ is unlikely to be more than 3):

$$
\begin{aligned}
&\mathcal{L}(\lambda,\theta|n_{\text{FDCIG}}) =\\
&\quad \text{constant} +\\
&\quad n_{\text{None}}\cdot\log\left[\frac{\lambda^0 e^{-\lambda}}{0!}(1-\theta)^0 + \frac{\lambda^1 e^{-\lambda}}{1!}(1-\theta)^1 + \frac{\lambda^2 e^{-\lambda}}{2!}(1-\theta)^2 + \right.\\
&\qquad\qquad\left.\frac{\lambda^3 e^{-\lambda}}{3!}(1-\theta)^3\right] +\\
&\quad n_{\text{Few}}\cdot\ \log\left[\frac{\lambda^2 e^{-\lambda}}{2!}\binom{2}{1}\theta^1(1-\theta)^1 + \frac{\lambda^3 e^{-\lambda}}{3!}\binom{3}{1}\theta^1(1-\theta)^2\right] +\\
&\quad n_{\text{Most}}\cdot\log\left[\frac{\lambda^3 e^{-\lambda}}{3!}\binom{3}{2}\theta^2(1-\theta)^1\right] +\\
&\quad n_{\text{All}}\cdot\ \log\left[\frac{\lambda^1 e^{-\lambda}}{1!}\binom{1}{1}\theta^1 + \frac{\lambda^2 e^{-\lambda}}{2!}\binom{2}{2}\theta^2 + \frac{\lambda^3 e^{-\lambda}}{3!}\binom{3}{3}\theta^3\right]
\end{aligned}
$$

[14] The Newton-Raphson algorithm:(Gelman et al., 2003)

1. Choose a starting point, $\overline{\theta}_0$, where $\overline{\theta}$ is a vector of all parameters; in our case: $\overline{\theta} = \{\lambda, \theta\}$.
2. For convergence steps, t = 1,2,3,....
   (a) Compute $\mathcal{L}'(\overline{\theta}_{t-1})$ and $\mathcal{L}''(\overline{\theta}_{t-1})$. The Newton's method step at time t is based on the quadratic approximation to $\mathcal{L}(\overline{\theta})$ centered at $\overline{\theta}_{t-1}$.
   (b) Set the new iterate, $\overline{\theta}_t$, to maximize the quadratic approximation; thus,
   $$\overline{\theta}_t = \overline{\theta}_{t-1} - [\mathcal{L}''(\overline{\theta}_{t-1})]^{-1}\mathcal{L}'(\overline{\theta}_{t-1}).$$

We compute the first derivative:

$$\begin{aligned}
\frac{d\mathcal{L}}{d\theta} &= n_{\text{None}} \cdot \frac{-\frac{\lambda}{e^{\lambda}} - \frac{\lambda}{e^{\lambda}}(1-\theta) - \frac{\lambda}{2e^{\lambda}}(1-\theta)^2}{e^{-\lambda} + \frac{\lambda}{e^{\lambda}}(1-\theta) + \frac{\lambda}{2e^{\lambda}}(1-\theta)^2 + \frac{\lambda}{6e^{\lambda}}(1-\theta)^3} + \\
&\quad n_{\text{Few}} \cdot \frac{\frac{\lambda^2}{e^{\lambda}}(1-\theta) + \frac{\lambda^2}{e^{\lambda}}\theta + \frac{\lambda^3}{2e^{\lambda}}(1-\theta)^3 - \frac{\lambda^3}{e^{\lambda}}\theta}{\frac{\lambda^2}{e^{\lambda}}\theta(1-\theta) + \frac{\lambda^3}{2e^{\lambda}}\theta(1-\theta)^2} + \\
&\quad n_{\text{Most}} \cdot \frac{\frac{\lambda^2}{e^{\lambda}}2(1-\theta) - \frac{\lambda^3}{e^{\lambda}}\theta^2}{\frac{\lambda^3}{e^{\lambda}}\theta^2(1-\theta)} + \\
&\quad n_{\text{All}} \cdot \frac{\frac{\lambda}{e^{\lambda}} + \frac{\lambda^2}{e^{\lambda}}(1-\theta) + \frac{\lambda^3}{e^{\lambda}}\theta^2}{\frac{\lambda}{e^{\lambda}}\theta + \frac{\lambda^2}{2e^{\lambda}}\theta^2 + \frac{\lambda^3}{3e^{\lambda}}\theta^3} \\
&= n_{\text{None}} \cdot \frac{3\lambda(5-4\theta+\theta^2)}{-6+\lambda(-10+15\theta-6\theta^2+\theta^3)} + \\
&\quad n_{\text{Few}} \cdot \frac{2+\lambda+3\lambda\theta^2-4(1+\lambda)\theta}{(\theta-1)\theta(-2+\lambda(\theta-1))} + \\
&\quad n_{\text{Most}} \cdot \frac{2-3\theta}{\theta-\theta^2} + \\
&\quad n_{\text{All}} \cdot \frac{6(1+\lambda\theta+\lambda^2\theta^2)}{\theta(6+3\lambda\theta+2\lambda^2\theta^2)}
\end{aligned}$$

Finding the numerator of the final expression, necessary to derive the modes $\hat{\lambda}$ and $\hat{\theta}$, and then finding the second derivative is a tiresome task, even more so when we will need to solve for larger values of λ. Instead, we can use finite differencing to estimate the derivatives for λ and θ:

$$\begin{aligned}
\mathcal{L}'(\lambda) = \frac{d\mathcal{L}}{d\lambda} &\approx \frac{\mathcal{L}(\lambda+\delta,\theta|y) - \mathcal{L}(\lambda-\delta,\theta|y)}{2\delta} \\
\mathcal{L}'(\theta) = \frac{d\mathcal{L}}{d\theta} &\approx \frac{\mathcal{L}(\theta+\delta,\lambda|y) - \mathcal{L}(\theta-\delta,\lambda|y)}{2\delta} \\
\mathcal{L}''(\lambda,\lambda) &= \frac{d^2\mathcal{L}}{d\lambda d\lambda} = \frac{d}{d\lambda}\left(\frac{d\mathcal{L}}{d\lambda}\right) \\
&\approx \frac{\mathcal{L}'(\lambda+\delta|y,\theta) - \mathcal{L}'(\lambda-\delta|y,\theta)}{2\delta} \\
&\approx [(\mathcal{L}(\lambda+\delta+\delta,\theta|y) - \mathcal{L}(\lambda-\delta+\delta,\theta|y)) - \\
&\quad (\mathcal{L}(\lambda+\delta-\delta,\theta|y) - \mathcal{L}(\lambda-\delta-\delta,\theta|y))]/(4\delta\delta) \\
\mathcal{L}''(\theta,\theta) &= \frac{d^2\mathcal{L}}{d\theta d\theta} = \frac{d}{d\theta}\left(\frac{d\mathcal{L}}{d\theta}\right) \\
&\approx \frac{\mathcal{L}'(\theta+\delta,\lambda|y) - \mathcal{L}'(\theta-\delta,\lambda|y)}{2\delta} \\
&\approx [(\mathcal{L}(\theta+\delta+\delta,\lambda|y) - \mathcal{L}(\theta-\delta+\delta,\lambda|y)) -
\end{aligned}$$

$$
\begin{aligned}
& (\mathcal{L}(\theta+\delta-\delta,\lambda|y)-\mathcal{L}(\theta-\delta-\delta,\lambda|y))]/(4\delta\delta)\\
\mathcal{L}''(\lambda,\theta) &= \frac{d^2\mathcal{L}}{d\lambda d\theta}=\frac{d}{d\theta}\left(\frac{d\mathcal{L}}{d\lambda}\right)\\
&\approx \frac{\mathcal{L}'(\lambda,\theta+\delta|y)-\mathcal{L}'(\lambda,\theta-\delta|y)}{2\delta}\\
&\approx [(\mathcal{L}(\lambda+\delta,\theta+\delta|y)-\mathcal{L}(\lambda-\delta,\theta+\delta|y))-\\
&\quad (\mathcal{L}(\lambda+\delta,\theta-\delta|y)-\mathcal{L}(\lambda-\delta,\theta-\delta|y))]/(4\delta\delta)\\
\mathcal{L}''(\theta,\lambda) &= \frac{d^2\mathcal{L}}{d\theta d\lambda}=\frac{d}{d\lambda}\left(\frac{d\mathcal{L}}{d\theta}\right)\\
&\approx \frac{\mathcal{L}'(\theta,\lambda+\delta|y)-\mathcal{L}'(\theta,\lambda-\delta|y)}{2\delta}\\
&\approx [(\mathcal{L}(\theta+\delta,\lambda+\delta|y)-\mathcal{L}(\theta-\delta,\lambda+\delta|y))-\\
&\quad (\mathcal{L}(\theta+\delta,\lambda-\delta|y)-\mathcal{L}(\theta-\delta,\lambda-\delta|y))]/(4\delta\delta)
\end{aligned}
$$

where we select a δ low enough to approximate the derivative, typically 0.0001. The following results summarize the fit for $\mathcal{L}(\lambda,\theta|y)$:

$$
\begin{aligned}
&\hat{\lambda}=3.324,\hat{\theta}=0.326,\\
&\sigma_\lambda=0.02303,\sigma_\theta=0.00221,\\
&\mu_{smoke}=1.085,\sigma_{smoke}=0.00803,\\
&\Sigma=\begin{bmatrix} 0.0005300 & -0.00002080\\ -0.0000208 & 0.00000486\end{bmatrix},\\
&\mathcal{L}(\hat{\lambda},\hat{\theta}|y)=-783.56
\end{aligned}
$$

Figure 2.1 demonstrates that an assumption of normality in the error surrounding the estimates is completely appropriate. Figure 2.2 demonstrates the unimodality of the joint posterior: there is a unique λ,θ pair that best explains the observed data.[15] The Poisson-binomial mixture model predicts youths to have mean of roughly $3\frac{1}{3}$ friends and that almost $\frac{1}{3}$ of these friends smoke, yielding an average of just over one smoking friend in each ego-network of peers. Furthermore, the negative covariance between λ and θ is explainable: when λ increases holding for θ, the probabilities $\boldsymbol{p}_{\text{FDCIG}}=(p_{\text{None}},p_{\text{Few}},p_{\text{Most}},p_{\text{All}})$ shift towards ‘Few’. In order to balance this effect, a lower θ is required to achieve a similar $\boldsymbol{p}_{\text{FDCIG}}$.

Despite employing a relaxed assumption on the friends parameter, this model underperforms the fixed binomial model for $n_{friends}=3$, which as we recall produced

[15] The effectiveness of the Newton-Raphson algorithm in these analyses obviates the need to resort to more advanced methods of posterior estimation such as Monte Carlo Markov Chains (MCMC), which in comparison tests yields identical results, as expected.

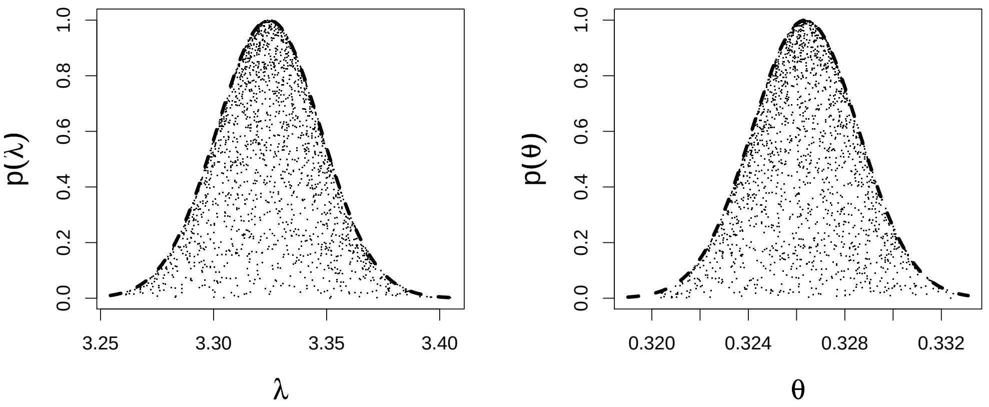


Figure 2.1: Normal Approximations for λ and θ. *After drawing 5,000 pairs of* $\{\lambda,\theta\}$ *from the appropriate multivariate normal as defined by the means for* λ *and* θ *and their accompanying covariance matrix, we plot each pair's* λ *and* θ *separately to its actual marginal likelihood (normalized by the likelihood at the mode):* $p(\lambda|y)/p(\hat{\lambda}|y)$ *and* $p(\theta|y)/p(\hat{\theta}|y)$*; the dashed lines depict the densities for each normal approximation:* $\lambda \sim Normal(\mu = 3.324, \sigma^2 = 5.306 \times 10^{-4})$ *and* $\theta \sim Normal(\mu = 0.326, \sigma^2 = 4.86 \times 10^{-6})$*.*

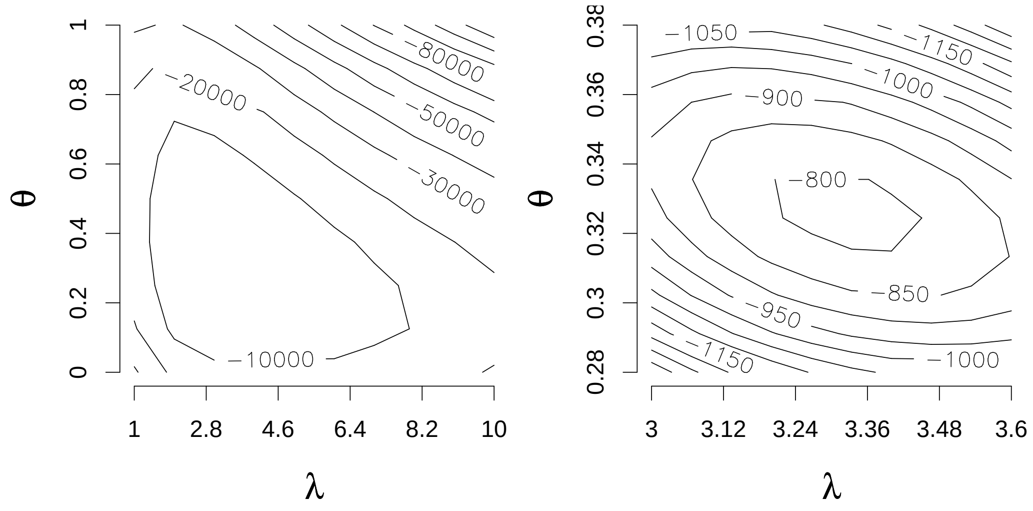


Figure 2.2: Joint Posterior Density for $p(\lambda, \theta|y)$. *The joint density is clearly unimodal. The left plot shows contour levels for a wide range of* λ *and* θ*, while the right plot focuses more closely on the mode.*

$\mathcal{L} = -118.12$. Still, it is unreasonable to assume that there is an, a priori, maximum to the number of friends possible, so we will continue to employ the Poisson/binomial model.[16]

2.5.3 Definition of "Few"

Here, we revisit the definition of 'Few', limiting what the maximum value could be. The intuition here is that when dealing with counts, people will consider anything above a certain small number, say 2, to constitute something more than just a 'Few'; hence, those quantities would be considered as 'Most'. This makes some intuitive sense: say I have ten friends and four of them smoke. Under the initial definitions, I would declare that a 'Few' of my friends smoke. However, it is conceivable that a respondent would instead mark off 'Most'.

At most, 'Few' means ...	$\hat{\lambda}$	$\hat{\theta}$	$\sigma_{\hat{\lambda}}$	$\sigma_{\hat{\theta}}$	$\mathcal{L}(\hat{\lambda}, \hat{\sigma})$
1	3.705	0.226	0.0333	0.00237	−422.10
2	3.690	0.288	0.0313	0.00238	−308.58
3	3.390	0.321	0.0248	0.00222	−715.96
4	3.331	0.326	0.0233	0.00221	−777.79
5	3.325	0.326	0.0231	0.00221	−783.20
6	3.324	0.326	0.0230	0.00221	−783.54
7	3.324	0.326	0.0230	0.00221	−783.56
∞	3.324	0.326	0.0230	0.00221	−783.56

Consequently, 'Most' will simply reflect the converse of a restricted 'Few'; e.g. if $n_{friends} = 6$ and 'Few'$_{max} = 2$, then 'Most' will be 3, 4, or 5. These results advise us to be wary of how respondents map counts to general categories. Indeed, if we limit the definition of 'Few' to 2, we obtain a much better fit than we did under our earlier formulaic definition (under the unrestricted 'Few' we obtained $\mathcal{L} = -783.56$). However, in subsequent analysis, it is not always the case that the fit is improved by altering the definition of 'Few', and since there is no current empirical evidence for restricting what 'Few' means, we will continue to use the original partitioning of the FDCIG categories.

2.5.4 Definition of "Use"

We can take the estimated parameters from Poisson/binomial model and compare the θ probability to the recency indicator variables to see if any would be a better definition of concept of friends' (or self) "use", which for cigarettes was indicated by

[16] The fixed friends approach fails miserably in later analyses, and even with a modification, imposing a binomial on the friend count, it is either inferior or sometimes comparable to the Poisson/binomial model that will employed throughout the rest of this work.

the phrase "how many of your friends *smoke*?". Ideally, if we had a true measure of "use" and had a well-performing model, our θ and this measure should be almost equivalent. Instead, we fit our findings to the empirical data by asking 'how likely is it that our estimate of θ can produce "use" as defined by one of the indicators?':

$$\begin{aligned} L\left(\theta|\frac{n_1}{n}\right) &= \text{Multinom}((n-n_1, n_1)|p=(1-\theta,\theta)) \\ &= \binom{n}{n-n_1, n_1}(1-\theta)^{(n-n_1)} \cdot \theta^{n_1} \end{aligned}$$

where n is the size of the considered population and n_1 is the subset of that population who answered 'Yes' to a particular substance use recency indicator. Again, we employ the multinomial distribution to give us a likelihood of our parameter being appropriate given the data:

Indicator	$n-n_1$	n_1	p_1	$\mathcal{L}$
CIGFLAG	16182	9281	0.364	-87.89
CIGRC3	17564	7899	0.310	-20.45
CIGYR	19456	6007	0.236	-506.84
CIGMON	21248	4215	0.166	-1692.92

Above, n_1 denotes the count of teens who answered "Yes" to the indicator and n is the total sample size. It appears that the three year recency indicator (CIGRC3) best coincides with the estimate of θ, followed by the lifetime use indicator (CIGFLAG). Remember, our estimated θ was 0.326. Furthermore, we can incorporate the fit to empirically observed recency of use directly into the log-likelihood function that estimates λ and θ:

$$\mathcal{L} = \mathcal{L}(\lambda, \theta|\boldsymbol{n}_{\text{FDCIG}}) + \mathcal{L}(\theta|(n-n_1, n_1))$$

And, we get the following adjusted λ and θ for each of the recency indicators:

Indicator	λ	θ	p_1	$\mathcal{L}$
CIGFLAG	3.27	0.340	0.364	-841.14
CIGRC3	3.35	0.321	0.310	-798.57
CIGYR	3.46	0.294	0.236	-1118.94
CIGMON	3.57	0.270	0.166	-1931.69

Recall that the earlier λ was 3.324, but the best fitting value here is 3.35. Again, CIGRC3 appears for now to be the best candidate for cigarette "use". We will keep this issue in mind, as we now turn to estimating parameters for sub-populations of teens.

2.6 Sub-Populations

Given the substantial association between a teen's substance use and the level of use in his or her peer group, we might expect the network parameters for the sub-population of using teens to differ from those of non-using teens. Specifically, the ego-networks of teens who admit to some level of smoking will include more smokers than those ego-networks of teens who claim to have never tried smoking. At first glance, we look at the "ever smoked a cigarette" indicator (CIGFLAG) and split the sample into two sub-populations of those who never used and those you smoked at some point in their lives. The tabulated FDCIG responses for each of these sub-populations are:

y_{CIGFLAG}	None	Few	Most	All	μ_{FDCIG}	prop.
0	8406	6315	1009	185	1.559	0.636
1	1215	4337	2891	692	2.335	0.364

Youths who have smoked at least once will list, on average, a higher proportion of their friends as smokers than those who have not; this observation echoes the findings of the earlier regression model reported in Table 2.1. We expect the estimated θ's to also differ accordingly; though, strictly speaking, this does not have to be the case if for some reason the λ's between the sub-populations greatly differed:

y_{CIGFLAG}	λ	θ	σ_λ	σ_θ	$\mathcal{L}$	n	μ_{smoke}	σ_{smoke}
0	3.476	0.198	0.0393	0.00251	-314.51	16182	0.689	0.0075
1	4.434	0.488	0.0437	0.00313	-55.43	9281	2.163	0.0232
Both	3.824	0.304	-	-	-369.94	-	1.226	-

The fit (i.e. log-likelihood) of this analysis is simply the sum of the log-likelihoods (i.e. product of the likelihoods) of each sub-population, $\mathcal{L}$ = -369.94, and is significantly superior to our earlier plain vanilla analysis which involved no breakdown of the data; there, the $\mathcal{L}$ was -783.56. This improvement in the log-likelihoods is reflected in the narrow standard deviations that surround the λ and θ estimates for both sub-populations; their differences are clearly significant. If we want to account for the differences in the number of parameters, we would compare the respective information criteria. We select the BIC which incorporates the sample size and find that the new model (with BIC = 770) still outperforms the previous one that lacked a population split (with BIC = 1587).[17] Not surprisingly, we see that the rate of use among friends of a smoker (θ = 0.488) is higher than that of friends of non-smokers (θ = 0.198).[18] Accordingly, the difference in the mean number of smoking friends re-

[17] The first model had two parameters, λ and θ, whereas in the last analysis we count y_{CIGFLAG} as an additional parameter; hence, that model has three parameters. Given equal sample sizes, a superior model would need to outperform an inferior model by a factor of Δ(# of parameters) $\cdot$ log(N), that is, log(N) for every additional parameter, and in this case 1 $\cdot$ log(25052) or 10.13.

[18] We employ the term "smoker" loosely; here it of course specifically means one who has ever smoked. This is less awkward than using the terms "initiate" and "non-initiate". While there might be some debate as to whether this definition of smoking is too broad, we will later see that the CIGFLAG indicator is the appropriate proxy for "use" for certain age groups.

flects greater affiliation with similarly behaving friends; smokers have more than three times the number of smoking friends than non-smokers, confirming outside findings of the prominent affiliation between smokers (Urberg et al., 1997).

However, we would not expect, a priori, that a smoker has more friends than a non-smoker, despite the slightly overlapping λ distributions (as indicated by the σ_λ's). But, if we consider that less than half of the teen population are smokers (under any of the recency indicators), it stands to reason that, if smoking is mostly, or even partly, a peer-influenced behavior, those who engage in it will necessarily have more friends.[19] The disparity in the magnitudes of the $\mathcal{L}$'s is partly explained by the sub-population sizes; smaller populations tend to be more easily "fittable". However, as we will see, a worse $\mathcal{L}$ also suggests a less easily definable population, one that is possibly composed of several distinct sub-types, as demonstrated by the superior fit on the population when split by y_{CIGFLAG}. Finally, each of the parametric differences is highly significant, $p < 0.001$.[20]

We now revisit the definition of "use" by comparing the above results to similar decompositions employing the other recency indicator variables: CIGRC3, CIGYR and CIGMON.[21] We find that the $\mathcal{L}$s are -400.88, -478.36 and -572.40, respectively. Again, these results are deceptive, suggesting that partitioning the teen population on CIGFLAG fits best ($\mathcal{L}$ = -369.94) and is perhaps the best definition for "use". As we briefly considered earlier, an alternative approach to fitting includes the overall population-level of "use". If we assume that respondents reside in separate ego-networks, we can assess how good of a "use" variable CIGFLAG is by estimating population-level use from the rate of use in each type of peer group.[22] Specifically, if our "use" decomposition variable (e.g. CIGFLAG) is accurate, then the rates of smoking in each sub-population, weighted by the relative size of the sub-population, should sum to the empirical, population-level rate of smoking:

$$\frac{n_1}{n_0 + n_1} \quad \approx \quad \left(\frac{n_0}{n_0 + n_1}\right)\theta_0 + \left(\frac{n_1}{n_0 + n_1}\right)\theta_1$$

[19] In Appendix B, the friendship data from NSDUH survey years of 1979 and 1982 supports the assertion that smokers have more friends than non-smokers. Urberg et al. (1997) also found that adolescents who tried smoking have more close friends.

[20]

Δparameter	t-statistic	degrees of freedom
$\Delta\lambda$	-1729.25	17386.46
$\Delta\theta$	-7561.35	15818.80
$\Delta\mu_{smoke}$	-5883.45	10255.38

[21] To recap, CIGRC3 means "smoked in the past three years"; CIGYR means "smoked in the past year"; and CIGMON means "smoked in the past month". Naturally, CIGRC3 includes all the respondents who answered 'Yes' to CIGYR and/or CIGMON, and CIGYR includes all the respondents who answered 'Yes' to CIGMON.

[22] This is a reasonable assumption given that the sample of the NSDUH was dispersed over the entire country. Still, it might prove valuable to understand how to estimate best "use" when this assumption does not hold.

$$\begin{aligned} p_{\text{CIGFLAG}} &\approx (1 - p_{\text{CIGFLAG}}) \cdot \theta_0 + p_{\text{CIGFLAG}} \cdot \theta_1 \\ 0.364 &\approx (1 - 0.364) \cdot 0.198 + 0.364 \cdot 0.488 \\ p = 0.364 &\approx q = 0.304 \end{aligned}$$

where n_0 and n_1 are the sizes of the non-use and use sub-populations, respectively When we apply this method of estimating the predicted proportion, q, to each of the indicator variables and compare it to the empirical proportion, p, we obtain:

If "use" is ...	p	q	$\lvert\Delta\rvert$	$\mathcal{L}_i$	$\mathcal{L}_d$	$\mathcal{L}_i + \mathcal{L}_d$
CIGFLAG	0.364	0.304	0.061	-219.49	-369.94	-589.43
CIGRC3	0.310	0.304	0.007	-7.90	-400.88	-408.77
CIGYR	0.236	0.305	0.069	-304.91	-478.36	-783.27
CIGMON	0.166	0.307	0.141	-1342.74	-572.40	-1915.14

where p is the observed proportion of use and q is the predicted proportion calculated.[23] The $\mathcal{L}_i$, the log-likelihood that the decomposition result fits the observed use pattern as specified by the indicator (denoted by the subscript i), was obtained using a 2-parameter multinomial, which is equivalent to a simple binomial; we used the actual weighted counts as the observed data and q as the hypothetical parameter.[24] For example, we obtain the $\mathcal{L}_i$ for the first entry (CIGFLAG):

$$\begin{aligned} \mathcal{L}_i &= \log[\text{Multinom}(n = (n_{\text{CIGFLAG=0}}, n_{\text{CIGFLAG=1}}) | \theta = (1 - q, q))] \\ &= \log\left[\binom{n_{\text{CIGFLAG=0}} + n_{\text{CIGFLAG=1}}}{n_{\text{CIGFLAG=0}} \;\; n_{\text{CIGFLAG=1}}} (1 - q)^{n_{\text{CIGFLAG=0}}} \cdot (q)^{n_{\text{CIGFLAG=1}}}\right] \\ &= \log\left[\binom{25464}{16182 \;\; 9281} (1 - 0.304)^{16182} \cdot (0.304)^{9281}\right] \\ &= -219.49 \end{aligned}$$

$\mathcal{L}_d$ is the log-likelihood from the λ, θ decomposition under the indicator (i.e. similarly computed as all the log-likehood estimates we have seen prior to this section). The

[23] In Appendix C, we provide decomposition (i.e. λ and θ) results for each indicator.

[24] Weighted recency indicator count data; we use the sums at the bottom initially and later the age-based breakdown:

	n_{CIGFLAG}		n_{CIGRC3}		n_{CIGYR}		n_{CIGMON}	
y_{AGE}	No	Yes	No	Yes	No	Yes	No	Yes
12	3580	493	3668	404	3795	277	3903	169
13	3408	921	3546	783	3729	600	3989	339
14	3001	1427	3160	1267	3547	880	3871	556
15	2462	1937	2753	1647	3135	1265	3529	871
16	2018	2115	2353	1780	2765	1368	3113	1020
17	1714	2388	2085	2017	2486	1616	2842	1260
Σ	16182	9281	17565	7898	19456	6007	21248	4215

sum of the two likelihoods gives us a measure of fit taking into account both fit on the ego-network data as well as the estimate for population use as determined by the various recency indicators. The highest log-likelihood is now obtained using CIGRC3 (in which the sum $\mathcal{L}$ is -408.77) suggesting once again that this indicator, rather than CIGFLAG, is the appropriate proxy for "use". Also, the predicted proportion of use (q) is relatively stable across all indicators suggesting an underlying consistency in the recency categories of use and proportions of friends' who use across those categories; however, further investigation is required to conclusively say that this is more than mere coincidence.

Instead of fitting solely on FDCIG and then assessing how well the parameters fit to the indicator data, we can be more precise by again the core log-likelihood function to reflect concurrent fitting on both the ego-network data FDCIG for both sub-populations and the various recency indicators:

Indicator	λ_0	θ_0	λ_1	θ_1	q	$p = n_1/n$	$\mathcal{L}_{di}$
CIGFLAG	3.29	0.219	4.39	0.505	0.324	0.364	-518.77
CIGRC3	3.48	0.212	4.58	0.514	0.306	0.310	-407.92
CIGYR	3.72	0.207	4.69	0.532	0.283	0.236	-691.77
CIGMON	3.91	0.203	4.79	0.567	0.263	0.166	-1528.63

where

$$\begin{aligned}\mathcal{L}_{di} &= \mathcal{L}_{di}(\lambda, \theta | (n_{\text{FDCIG}}), (n_0, n_1)) \\ &= \mathcal{L}_d(\lambda, \theta | (n_{\text{FDCIG}})) + \mathcal{L}_i(\theta | (n_0, n_1))\end{aligned}$$

and p and q are, again, our observed proportion of use and our predicted proportion. We observe that with the exception of θ_0, the parameter estimates vary (perhaps significantly) across the indicators; so, it will matter (for defining smokers' and non-smokers' ego-networks) which indicator we choose to mean "use". And, again, the results support the earlier finding that CIGRC3 (again with the highest $\mathcal{L}$) appears to be the best overall definition of perceived "use".

Given the shift in proportion of smoking with advancing years, it is reasonable to believe that our parameter estimates too will shift with age:

y_{AGE}	λ	θ	σ_λ	σ_θ	$\mathcal{L}$	μ_{smoke}
12	2.763	0.149	0.0788	0.00509	-125.48	0.41
13	3.246	0.219	0.0680	0.00509	-134.94	0.71
14	3.567	0.301	0.0608	0.00507	-113.81	1.07
15	3.670	0.368	0.0571	0.00508	-73.67	1.35
16	3.979	0.406	0.0613	0.00502	-69.61	1.62
17	4.101	0.445	0.0618	0.00495	-56.66	1.83
All	3.555	0.315	—	—	-574.18	1.16

While the fit here (i.e. $\mathcal{L}$) exceeds that of the non-decomposed model, it does not exceed that of the CIGFLAG or CIGRC3 indicator decomposed models suggesting that self-use, rather than age, is the primary determinant of association to peer group smoking; this echoes the findings of the earlier regression model.[25] We now observe that the size of an adolescent's peer group increases with advancing age. While we might be tempted to think that this association is confounded with the increasing prevalence of smoking with age, we will find later that this is not the case.[26] We can again match the predicted proportion of use for each age group with the observed proportion. However in doing so, we make an important assumption: that peer groups are homogeneous in age composition.[27] Despite the apparent rigidity of this assumption, the age difference between adolescent friends tends to be very small.[28] A host of studies have confirmed age homophily, especially among youth, to allow at least an initial assumption of friendship ties being within-age (Verbrugge, 1977; Bott, 1928; Loomis, 1946; Fischer, 1977, 1982; Kirke, 1996; McPherson et al., 2001; Marsden, 1988). The combined likelihood function produces the following fits, per age group:

	$\mathcal{L}_{di}$ for CIG...			
y_{AGE}	FLAG	RC3	YR	MON
12	**-137**	-156	-208	-290
13	**-139**	-151	-197	-349
14	**-121**	**-120**	-198	-384
15	-110	**-78**	-122	-289
16	-139	**-78**	-109	-240
17	-170	**-74**	**-77**	-179
$\Sigma\mathcal{L}_{di}$	-816	**-656**	-911	-1732

We identify the best indicator(s) in each age group with the bold-typed $\mathcal{L}$ values.[29] The results are quite telling: while CIGRC3 remains the best overall indicator of "use" (as expected), the appropriate indicator for "use" varies with age. This makes some sense. Smoking is more stigmatic for younger teens who might consider a friend who smokes only a few times to be a (gasp!) "smoker". With increasing age comes

[25] Still, the age effect on substance use prevalence should not be ignored; across all recency indicators, the level of use increases monotonically with age as shown in Table C.2.

[26] In Appendix B, the separate impact of both age and smoking on the size of the number of friends is shown from empirical close friends data in the 1979 and 1982 years of the NSDUH.

[27] We will maintain this assumption throughout the rest of this report. Even if we relax the assumption and believe a proportion of ties extend to proximal ages, the findings will be only marginally affected due to the increasing trends in λ and θ.

[28] Kirke found the age difference in her data set to be 0.12 years, with s.d. = 1.41 years.

[29] We set the margin of inclusion to be 3 units, which translates to a ratio in likelihood of e^3 or 20.09. Despite this large difference, there is a good chance the true distribution will include the parameters associated with the bold-typed log-likelihoods, depending of course on the level of uncertainty surrounding the estimates.

a lessening of the stigma attached to adolescent smoking (likely due to an overall increase in contact with smokers); hence, the perception of "use" would require a higher frequency of actual use, as captured by the tighter windows of recency.

Next, we decompose on both age and the indicators, starting with CIGFLAG:

y_{CIGFLAG}	y_{AGE}	λ	θ	σ_λ	σ_θ	$\mathcal{L}$	n	μ_{smoke}
0	12	2.923	0.111	0.105	0.005	-81.69	3580	0.33
0	13	3.422	0.150	0.101	0.005	-58.83	3408	0.51
0	14	3.667	0.203	0.094	0.006	-69.68	3001	0.74
0	15	3.841	0.245	0.096	0.006	-56.22	2462	0.94
0	16	4.030	0.271	0.105	0.007	-29.24	2018	1.09
0	17	4.290	0.280	0.119	0.007	-14.53	1714	1.20
1	12	3.529	0.360	0.167	0.016	-16.46	493	1.27
1	13	4.436	0.408	0.145	0.010	-20.59	921	1.81
1	14	4.734	0.456	0.123	0.008	-11.20	1427	2.16
1	15	4.305	0.491	0.093	0.007	-12.16	1927	2.11
1	16	4.643	0.507	0.096	0.006	-18.69	2115	2.36
1	17	4.624	0.537	0.088	0.006	-25.21	2388	2.49
All	-	3.921	0.301	-	$\Sigma\mathcal{L} =$	-414.58	-	1.26

The raw fit, $\Sigma\mathcal{L} = -414.58$, exceeds those produced by models based on CIGRC3 ($\Sigma\mathcal{L} = -424.03$), CIGYR ($\Sigma\mathcal{L} = -474.06$) or CIGMON ($\Sigma\mathcal{L} = -531.89$).[30] However, the earlier CIGFLAG-only decomposed model ($\mathcal{L}$ = -369) we introduced at the beginning of this section outperforms all of these, which might imply that the age break-down is artificial (i.e. the age-based friendship assumption is too rigid). Or, if that is not the case, the worsening of fits is merely coincidental. The results of the age and smoking decomposition are shown graphically as trajectories in Figure 2.3. Both size of peer group and prevalence of friends' smoking are distinct between the sub-populations of adolescents who never smoked and those who have, separated primarily by the prevalence of friends' smoking.

Again, we assess the likelihoods, separately ($\mathcal{L}_d + \mathcal{L}_i$) and jointly ($\mathcal{L}_{di}$), that we obtain from fits to the ego-network response and the proportion of recency indicated, self-reported use:

	CIGFLAG				CIGRC3				CIGYR			
y_{AGE}	$\mathcal{L}_d$	$\mathcal{L}_i$	$\Sigma\mathcal{L}$	$\mathcal{L}_{di}$	$\mathcal{L}_d$	$\mathcal{L}_i$	$\Sigma\mathcal{L}$	$\mathcal{L}_{di}$	$\mathcal{L}_d$	$\mathcal{L}_i$	$\Sigma\mathcal{L}$	$\mathcal{L}_{di}$
12	-98	-10	-109	**-106**	-101	-29	-129	-125	-101	-81	-181	-175
13	-79	-4	-84	**-84**	-82	-17	-99	-91	-89	-65	-154	-137
14	-81	-8	-89	-96	-83	-5	-88	**-87**	-107	-77	-185	-174
15	-68	-42	-111	-122	-69	-6	-75	**-77**	-75	-38	-113	-109
16	-48	-73	-121	-139	-53	-10	-62	**-67**	-55	-33	-87	-82
17	-40	-124	-164	-183	-39	-22	-61	-68	-47	-14	-62	**-59**
$\Sigma\mathcal{L}$				-729				-515				-736

[30] These latter results can be found in the appendix in Tables C.3, C.4, and C.5

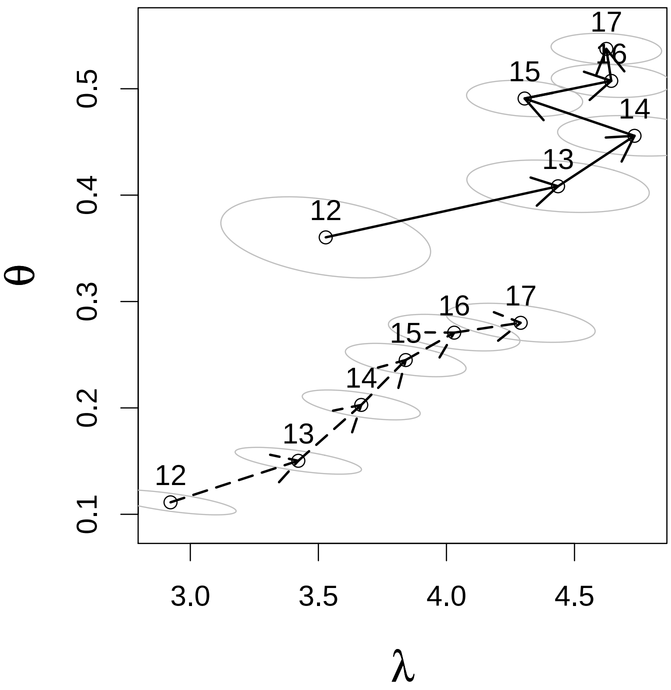


Figure 2.3: Trajectories for Cigarette Use. *Solid line denotes sub-population of respondents who have smoked at least once and dashed line denotes those who never tried. The numerical labels denote age groups. Grey ellipses around each point show the 95% probability regions surrounding each set of standard errors.*

The pattern of age-changing perception of "use" is further clarified, with changes occurring at age 14 and then at age 16. When we pool the parameter estimates that coincide with the best definition of "use" as determined by $\mathcal{L}_{di}$ for each age, we obtain:

	Non-Smokers		Smokers			
y_{AGE}	λ_0	θ_0	λ_1	θ_1	$\mathcal{L}$	Indicator
12	3.05	0.103	3.57	0.348	-106	CIGFLAG
13	3.39	0.153	4.43	0.411	-84	CIGFLAG
14	3.70	0.209	4.84	0.474	-87	CIGRC3
15	3.86	0.257	4.41	0.527	-77	CIGRC3
16	4.00	0.296	4.75	0.542	-67	CIGRC3
17	4.43	0.307	4.88	0.593	-59	CIGYR

Both populations show a steady increase in number of friends and number of friends who smoke. The probability of a friend smoking for a non-smoking 17-year-old approaches that of a smoking 12-year-old, revealing that there is just over a five year separation in the smoking dimension between the worlds of smokers and non-smokers. Oddly, there is a moderate drop in the number of friends for a smoker between the ages of 14 and 15. In fact, this drop is consistent across the other decompositions. While our data does not allow us to infer a reason for this, we can speculate that a change in the school system, when teens enter high school, might be responsible for this shift. Still, with non-smokers, there is no drop in peer group size and only a small increase between these ages.

2.7 Fitting Results to a Linear Model

Instead of merely eyeballing the association between respondents' characteristics (e.g. age and smoking behavior) and their ego-networks, we can directly assess their predictive power on λ and θ, and similarly on μ_{smoke}, by employing a linear fit; we have the necessary point estimates as well as uncertainty bounds for conducting such an analysis. We fit our estimates to the following linear regression models:

$$\begin{aligned}
\lambda &= \beta_0 + \beta_1 \cdot y_{\text{SEX}} + \beta_2 \cdot y_{\text{CIGRC3}} + \beta_3 \cdot y_{\text{AGE}} \\
\theta &= \beta_0 + \beta_1 \cdot y_{\text{SEX}} + \beta_2 \cdot y_{\text{CIGRC3}} + \beta_3 \cdot y_{\text{AGE}} \\
\text{logit}(\theta) &= \beta_0 + \beta_1 \cdot y_{\text{SEX}} + \beta_2 \cdot y_{\text{CIGRC3}} + \beta_3 \cdot y_{\text{AGE}} \\
\mu_m &= \beta_0 + \beta_1 \cdot y_{\text{SEX}} + \beta_2 \cdot y_{\text{CIGRC3}} + \beta_3 \cdot y_{\text{AGE}}
\end{aligned}$$

We choose the CIGRC3 indicator (i.e. smoked in past three years) as it seems to be the best overall definition of cigarette "use".[31] While the reported marginal standard de-

[31] Alternatively, we could also look at a hybrid data set that consists of parameters using the smoking indicator that best suits each age group.

	λ	θ	logit(θ)	μ_{smoke}
Intercept	0.421** (0.113)	-0.309*** (0.014)	-27.710*** (0.519)	-1.945*** (0.064)
Is Male	0.040 (0.025)	-0.015*** (0.003)	-0.321* (0.116)	-0.044* (0.014)
Smoked in Past 3 Yrs	0.849*** (0.029)	0.270*** (0.004)	7.431*** (0.131)	1.447*** (0.016)
Age	0.224*** (0.008)	0.036*** (0.001)	1.344*** (0.036)	0.193*** (0.004)
$\mathcal{L}$	-58.619	-27.855	-116.627	-89.964
adjusted R^2	0.088	0.279	0.203	0.345
p value	0.000	0.000	0.000	0.000

Table 2.3: Fitting Recency of Cigarette Use (CIGRC3) with Simulated Data. *Standard errors appear in parentheses under each coefficient. Coefficient p-values are denoted by: * = $p < 0.05$ ** = $p < 0.01$, *** = $p < 0.001$*

viations, σ_λ, σ_θ, and $\sigma_{\mu_{smoke}}$, give us a sense of the uncertainty around the parameters, it is more precise to employ the covariance matrix, provided by the Newton-Raphson procedure. Still, these do not necessarily express the population level spread around the parameters; the covariance matrix, and the reported marginal σ's, are analogous to the standard errors from a regression model and not the standard deviation of measures obtained from a population. So, we conduct our fit on simulated data, using both the parameters and their uncertainty bounds, as determined by the covariance matrix to create distributions, around each combination of gender, age, and cigarette use, with sample sizes consistent with the actual sub-population sizes.[32]. The dependent measures of Table 2.3 arise from a model employing a fit to the ego-network variable only and does not include the additional fit to the indicator due to difficulties in converging to a stable solutions. We also conduct a regression analysis on logit(θ); this is appropriate when the dependent variable is a proportion or a probability.

In Table 2.3, we observe with no surprise that the respondent's recent smoking is strongly associated with the degree of smoking (both proportion and count) in his or her peer group and also the total number of friends in his or her peer group, corroborating outside reports of a strong affiliation between smokers. We also see that age significantly predicts both size of peer group as well as rate of smoking friends, neither of which are surprising given that friendship circles grow over time, and further, the proportion of smokers increases from age to age, increasing that rate over time. However, predicting the parameter for friends' count, λ is less precise, as demonstrated by the lesser R^2, due to the uncertainty surrounding its estimates. The significant negative coefficient for being male, while puzzling, is corroborated by an

[32] The full data appears in Appendix E under Tables E.1 and E.2

ordinal logistic regression on the ego-network cigarette use (FDCIG) predicted by age and gender.[33] Considering that being male is positively associated with cigarette use of any recency (Table C.2), one explanation is that males are more easily influenced when it comes to smoking; hence, we might observe boys to have fewer smoking friends, depending on how much selection plays a role. Furthermore, we would also see a negative coefficient if it was the case that boys are more likely than girls to continue smoking without the presence of smoking friends. All these observations can be summed up by saying boys take to smoking more easily than girls; this assertion, however, requires further testing.

2.8 Alternative Mixtures

Often, when overdispersion in a Poisson is suspected, the negative-binomial is employed in place of a Poisson, essentially giving λ a gamma prior with two parameters, α and β to specify the mean and the spread.[34] In Appendix B, it is shown that the negative-binomial is potentially the better distribution than the plain vanilla Poisson for raw friendship counts. However, under the few conditions tested here, the negative-binomial appears to be unnecessary:

population	$\lambda \sim \frac{\alpha}{\beta}$	α	β	θ	$\mathcal{L}$
all	3.58	4.06	1.13	0.349	-858.33
$y_{\text{CIGFLAG}=0}$	3.48	2285	657	0.198	-315.36
$y_{\text{CIGFLAG}=1}$	4.43	428	96	0.488	-55.05

While mean friends, $\lambda \sim \frac{\alpha}{\beta}$, is comparable to the non-dispersed, Poisson-only findings, there is actually a loss in the fit as indicated by the lower $\mathcal{L}$. Furthermore, upon looking at sub-populations, the estimation process is often degenerate, suggesting that λ is not overdispersed; the large values for both α and β suggest that the spread around the mean is exceedingly small and, hence, the variance of the distribution approaches that of the Poisson.

Another mixture, in which $n_{friends}$ is modeled as a binomial with a fixed maximum, n_{max}, and also with a variable n_{max} having uniform density, is found to be degenerate

[33] Results of Ordinal Logistic Regression on Friends' Smoking (FDCIG):

Dep. Var.	Coefficient	Std. Error	t value
Age	0.4475	0.00769	58.18
Is Male	-0.0703	0.02456	-2.90

A t value of absolute value greater than 2 usually denotes significance at the $p < 0.01$ level.

[34] The negative-binomial:

$$p(\theta) = \binom{\theta + \alpha - 1}{\alpha - 1} \left(\frac{\beta}{\beta + 1}\right)^{\alpha} \left(\frac{1}{\beta + 1}\right)^{\theta}, \text{ where } \theta = 0, 1, 2, \ldots$$

in many of the sub-population analyses, especially the substance using population. Finally, we tested an alternative modeling of θ as well, treating it also as overdispersed, by employing a beta-binomial (i.e. a beta prior on θ). However, this enhancement is found to overfit the results; convergence cannot be achieved because there are simply too many solutions.[35]

[35] However, these modifications have not yet been tested on any of the joint analyses.

Chapter 3

Generating Distributions of Complete Networks From Ego-Networks

We have so far managed to improve our understanding of an adolescent's peer group, by estimating its size and smoker composition. While it behooves us to continue the analysis at the ego-centric level and further explore the processes involved in substance use, such as initiation as well as changes in peer group composition (and we will examine these in a later chapter), we also need to address, in some way, the interdependence between these ego-networks. That is, ego-networks do not evolve in isolation, and the degree to which changes in one set of peers affect the dynamics in another can often be predicted by their social proximity as well as strength of the connections. Only under extreme conditions will ego-networks alone have as much explanatory power as a complete network: when the dynamic, or behavior, under study, exhibits complete independence between dyads or between ego-networks. In the former case, simple dyadic data would be sufficient, and even the ego-network would be overkill.[1]

Clearly, adolescent substance use is rife with interdependent social dynamics; hence, ego-network-only analysis will offer only a partial depiction of events. Behavior borne, even partly, from influence has a tendency to spread into epidemiological proportions, though tempered when it also induces structural change, concurrently. For example, an adolescent likely initiates in response to the influence of some using friends. In turn, this new user becomes a source of influence for those friends who have yet to sample the substance. Furthermore, if these non-using friends and the new user have additional using friends in common, then the risk of initiation for the non-users will jump a notch as a result of that one friend's transformation. While this dynamic may be indirectly captured in ego-network-only analysis, ideally we want

[1] However, one would be hard pressed to imagine any social process that exhibits this level of independence.

longitudinal, complete network data, which enumerates all existing ties among peers of a discrete population at numerous points in time. However, network researchers do not possess anywhere near the fantastical array of resources needed to collect this kind of data on a large scale. Instead, current studies resort to surveying modestly-sized populations, several hundred at most, with just a few waves of data collection. Hence, we are left with two practical options for studying the role of peer networks in adolescent substance use: either collect a large number of ego-networks with the expectation that the sheer volume of information will permit generalizability or collect few instances of complete networks, sacrificing generalizability for more inter-related data.[2]

Since our data falls under the first category of substance use data, the kinds of inferences we can draw seem limited, at first glance, to isolated ego-networks. However, our data permits us an additional layer of inference. Since we can identify which respondents are users of a particular substance and which ones are non-users,[3] we can essentially link ego-networks by assigning one respondent as the friend of another, matching them according to the type of dyad these two are fullfilling (i.e. user to user, non-user to user, or user to non-user). The goal, here, is not the construction of the actual network from which the data was obtained,[4] but one possible network from which the ego-networks may have been drawn. Ultimately, we would generate a population of hypothetical networks with the hopes of identifying network properties that distinguish users from non-users; if we do not see any significant differences, then either a) such distinctions cannot be uncovered through ego-network matching or b) none exist. This linking of ego-networks is a relatively untested technique, with good reason; the idea of creating distributions of data to supplement inadequate data can feel anathema. Missing data techniques such as bootstrapping and multiple imputation methods are relatively new in both network and general statistical analysis (Schafer, 1997). Still, some research such as Friedman et al. (1997, 1999) have resorted to using demographic and ethnographic data to link unresolvable ego-networks for the purposes of attaining a complete network, using a software package LinkAlyzer specifically designed for this purpose (Tien, 2001).

Implicit in linking ego-networks is the distinction between ties that remain within-group and those that span groups, that is between-group ties. It is easy to envision this process if one considers a hypothetical population of actors defined by only one or two binary properties, such as substance use (yes or no). Building off work by Rapoport (1957, 1958, 1963), Fararo and Sunshine (1964), Fararo and Skvoretz (1984), and Skvoretz et al. (2004) further developed a biased net approach to explore network

[2] In network analysis, the scope of the network need not be vast if the focal behaviors are naturally restricted by context, like an organization or school, or by definition, like kin relationships.

[3] Some caution is appropriate since the definition of 'use' has been shown to vary with age.

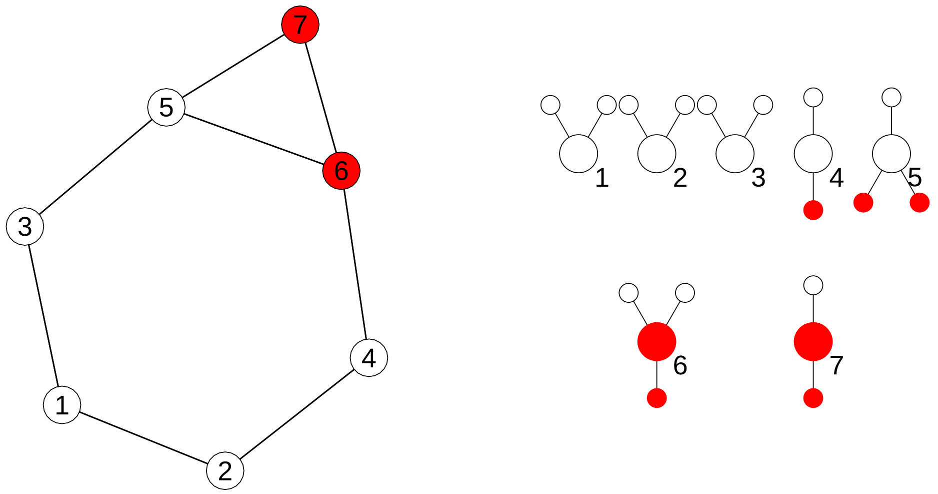
[4] Yet, a network reconstruction is possible if a) we know the ego-networks to have been sampled from a common, discrete population and b) we have sufficiently identifying data that allows for a one-to-one mapping between the ego-networks and a complete network.

structure that is predicted, in part, by tie volume between groups. While homophilic tendencies induce ties to remain within group, containment is generally not absolute; hence, ties across groups ought to observed. This approach assumes a high level of reciprocity in the ties studied; communication or friendship ties are often reciprocal and shown to be so in empirical analysis, though not perfectly. Explicit modeling of reciprocated ties across groups has been further developed by Heckathorn (1997, 2002, 2007) and Salganik and Heckathorn (2004) in their efforts to improve on link-trace sampling with respondent-driven sampling. All these models and the matching process in this chapter convey some of the concepts in Blau's macrosocial structural theories focusing on intergroup relations as emergent from micro-structure (Blau, 1977; Blau and Schwartz, 1984). Beyond static data, some dynamic modeling studies have exploited both the propensity of ties to be formed along homophilic dimensions, while admitting a proportion of ties to be non-homophilic (Carley, 1990, 1991; Zeggelink, 1994, 1995). In Lee (2002, 2004), I conduct some preliminary exploration into the ego-network linking approach, highlighting its complexity, strengths and weaknesses. That line of work offers reasons to be wary of the simplistic nature of the ego-networks analyzed here; the lack of alter-to-alter cross-ties in the NSDUH ego-network data will widen the uncertainty of the matchings.

Perfect matching can only occur when our information allows us to fully identify the alters of each ego-network; both actor and structural properties can serve as identifying information. This is exactly how network analysts assemble a complete network, by using uniquely identifying data, such as full names. Other traits such as gender, age, etc. can also be instrumental when full identifying information is unavailable; and the pattern of linkages between actor categories serve as structural information that can assist in the matching. For instance, if a 16 year-old respondent (a) is linked to a male 17-year-old (b) and a female 15-year-old (c) and reports a link between the two of them, potential candidates for the male 17-year-old (b) would have to report being involved in a triad with a female 15-year-old (c) and a male 16-year-old (a).

Unfortunately, in this work, alter-to-alter ties cannot be modeled since no relevant data were provided by the NSDUH respondents. Instead, we focus solely on actor properties to make the linkages, specifically substance use and age.[5] We start with a simple example to demonstrate the matching algorithm. In the left graph of Figure 3.1, we show a labeled, connected network of seven actors, two of whom are colored; for our purposes, this denotes smoking. The right graph depicts each actor's ego-network. The small circles denote the types of friend(s) that the ego (large circles) reports having. That is, actor 1 reports having exactly two non-smoking friends while actor 6 reports having one smoking friend and two non-smoking friends. The limitations of the ego-network should be easily apparent; without proper identifying information, we cannot know which two egos are actor 1's non-smoking friends. However, since

[5] While networks are strictly limited to within-age groups, future work will include a relaxation of this restriction.



(a) Complete Network

(b) Ego-Networks

Figure 3.1: Example 7-node Network. *On the left, we supply a simple, illustrative example network comprising 7 members. The right figure displays each ego-network, separately. Links directed upwards require recipient egos from the upper, white group while links directed downwards have recipients from the lower, dark-colored group.*

there are only two smokers in this network, these have to be actor 5's two smoking friends.

Our data is even less certain than what was just described. Instead of the exact number of friends and the number of those who use a substance, we only know distributions of ties counts through the inferred λ and θ. The goal of this chapter is to provide a method for generating distributions of networks that comply with the parameters inferred from the friends' smoking (FDCIG) variable, when decomposed with substance use covariates like recency of use.

3.1 Matching Ties

Our primary constraint is the number of ties that cross the distinct groups, the distributions of which must be similar, under the assumption that our system is more or less closed.[6] That is, *the total number of smoking friends that non-smokers, from a given network, report need to equal the total number of non-smoking friends that smokers, in the same network, report.* The distribution of the total ties spanning non-same groups is simply Poisson which converges to a normal distribution for large

[6] This remark alludes to our treatment of age-specific networks. Also, we treat the entire 12–17 year-old population as closed, another assumption we will relax in subsequent writings.

means.[7] That is, the sum of n Poisson samples, with a mean and variance of $\lambda\theta$, yields a distribution with mean $\lambda\theta n$ and variance $\lambda\theta n$. We are primarily concerned about equating the number of friendship ties from n_0 non-users to n_1 users and ties from those users to non-users, which we denote using T_{01} and T_{10}, respectively. The parameters for describing each set of ties are:

Non-User To User Ties	User To Non-User Ties
$\mu_{01} = \lambda_0\theta_0 n_0$	$\mu_{10} = \lambda_1(1-\theta_1)n_1$
$\sigma^2_{01} = \lambda_0\theta_0 n_0$	$\sigma^2_{10} = \lambda_1(1-\theta_1)n_1$
$T_{01} \sim \text{Normal}(\mu_{01}, \sigma^2_{01})$	$T_{10} \sim \text{Normal}(\mu_{10}, \sigma^2_{10})$

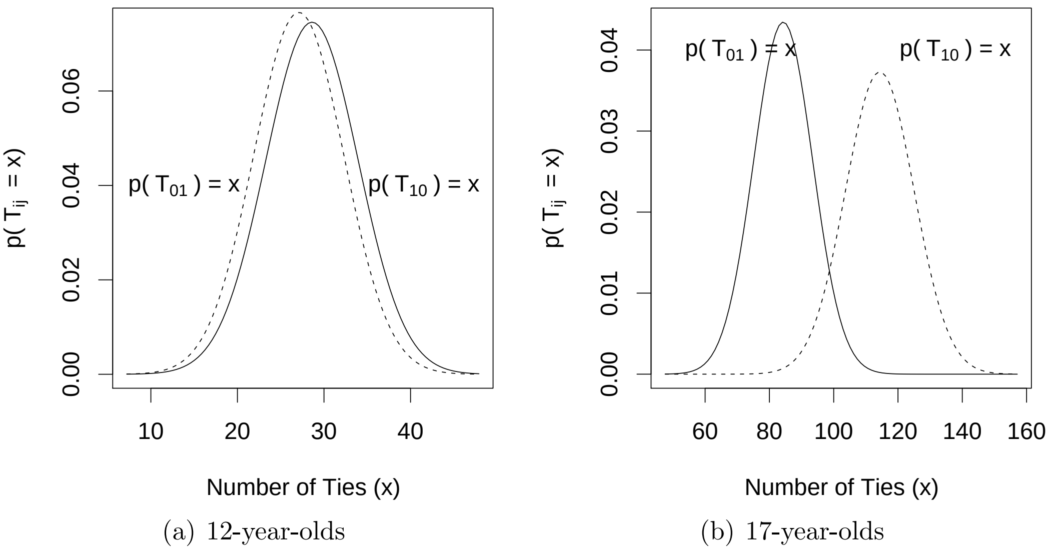
To assess how likely $p(T_{01} = T_{10})$, we simply compute all the ways in which the distributions can be equal and multiply their associated probabilities. For large values of n, these distributions converge to normal distribution, and since we are concerned with only discrete values, we can interpret the density directly as a probability:

$$\begin{aligned}
&p(T_{01} = T_{10}|\mu_0, \sigma_0, \mu_1, \sigma_1) \\
&= \sum_{i=\mu_0-4\sigma_0}^{\mu_1+4\sigma_1} \text{Normal}(i|\mu_0, \sigma_0^2) \cdot \text{Normal}(i|\mu_1, \sigma_1^2) \\
&= \sum_{i=\mu_0-4\sigma_0}^{\mu_1+4\sigma_1} \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{1}{2\sigma_0^2}(i-\mu_0)^2\right) \cdot \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2\sigma_1^2}(i-\mu_1)^2\right)
\end{aligned}$$

We examine a range of four standard deviations below the lower μ and the same above the higher μ in order to cover the range in which either probability is effectively non-zero. Note, the farther apart μ_0 and μ_1 are, the lower our $p(T_{01} = T_{10})$ will be, in which case we deem the parameters inaccurate or inappropriate. In Figure 3.2, we compare the overlapping tie distributions, given a network size of $n = 100$, using the appropriate definition of "use": CIGFLAG for the 12-year-olds and CIGYR for the 17-year-olds. The less-than-ideal overlapping of 17-year-old sub-groups suggests that our calculated tie volumes are somewhat inaccurate. However, we can attempt to adjust for this by incorporating tie volume equivalence into the fitting of ego-network parameter estimates, under the belief that these estimates need to also reflect tie volume consistency.

For each of the recency of use indicators, we report estimates when fitting to both the degree of overlap in the tie count distributions for the two sub-groups and the likelihood that the estimated θ's reflect population level "use"; the log-likelihood is denoted as $\mathcal{L}_{dti}$. For comparison, we also report the earlier estimates obtained from fitting to FDCIG only, with no additional constraints:

[7] If we wanted to be really precise, we would incorporate the covariance between λ and θ but since these are small enough to not overly affect the variance of the tie volume distribution, we will omit them for now.



(a) 12-year-olds

(b) 17-year-olds

Figure 3.2: *Matching Ties for Two Age Groups: on the left, we show the distribution of* T_{01} *and* T_{10} *ties for 100 12-year-olds,* $n_{12} = (12,\ 88)$*; on the right, the distributions for an equivalent number of 17-year-olds,* $n_{17} = (61,\ 39)$*.*

	Tie and Indicator					None				
	λ_0	θ_0	λ_1	θ_1	$\mathcal{L}_{dti}$	λ_0	θ_0	λ_1	θ_1	$\mathcal{L}_d$
y_{CIGFLAG}	3.30	0.219	4.38	0.506	-526	3.48	0.198	4.43	0.488	-370
y_{CIGRC3}	3.48	0.212	4.57	0.515	-413	3.50	0.210	4.58	0.513	-401
y_{CIGYR}	3.72	0.207	4.69	0.532	-695	3.54	0.229	4.66	0.551	-478
y_{CIGMON}	3.90	0.203	4.80	0.567	-1536	3.58	0.247	4.76	0.607	-572

The differences between the sets of results follow an interesting pattern: while the ranges of θ's are more constricted for tie and indicator fitting, the ranges of λ's are wider. Furthermore, the level of friends' use for non-users θ_0 is very narrow under tie and indicator fitting; this coincides with an earlier observation, in the last chapter, where we saw the predicted population level of smoking to be almost constant across all definitions of "use". The log-likehoods are naturally worse due to the additional tie-matching constraint introduced into the likehoood.

Below, we report age-specific tie matching $\mathcal{L}_{dt}$ for each of the indicators as well as fitting both tie matching and indicator $\mathcal{L}_{dti}$, for each indicator:

	$\mathcal{L}_{dt}$ for CIG...				$\mathcal{L}_{dti}$ for CIG...			
y_{AGE}	FLAG	RC3	YR	MON	FLAG	RC3	YR	MON
12	**-101**	-104	-106	-110	**-109**	-128	-179	-260
13	**-84**	**-85**	-93	-113	**-88**	-95	-141	-304
14	**-88**	**-88**	-113	-145	-102	**-92**	-178	-386
15	-78	**-73**	-80	-94	-129	**-81**	-113	-276
16	-63	**-59**	**-59**	-84	-149	**-72**	-85	-229
17	-58	**-46**	-51	-65	-195	-74	**-63**	-163
$\Sigma\mathcal{L}$	-471	**-455**	-501	-610	-773	**-542**	-758	-1618

The pattern of changing definition of "use" we observed earlier continues to hold. Definition of friends' smoking increases in recency from lifetime at ages 12 and 13, to past three year use for ages 14-16, and past year use for age 17. Furthermore, the past three years use recency indicator continues to be the best general definition of friends' "smoking".

3.2 Network Generation

Now that we have the appropriate parameters that resolve ties between sub-populations of substance users and non-users, we can commence generating a complete network. Later, we will generate a population of simulated networks from which we can draw distributions of network-related measures.

Firstly, we decide on the size, N, of the total population of the network. This population reflects a bounded group within which we believe almost all of the ties remain; ties extending beyond will be rare or inconsequential. Accordingly, this population can represent a school, a class, or even a single classroom, depending on what we believe or assume about the boundaries of the teens' social world. Next, we assign a proportion, p, of those teens to be the substance users. Finally, we require the parameters for number of friends for each sub-population, λ_0 and λ_1, and the rate of use within each of those, θ_0 and θ_1. The procedure below details the generation of just a single network; for the purposes of clarity, we will explain the algorithm in segments.

Algorithm 1 Generate Ego-Networks, Part 1

```
Require: $N \gg \lambda_0, \lambda_1$                        ▷ size of population
Require: $p, \theta_0, \theta_1 \in [0,1]$                  ▷ substance using proportions
Require: $\lambda_0, \lambda_1 > 0$                          ▷ mean friends per group
```

The population of N needs to be of sufficient size so as to support matching. For instance, a network with high λ's but low N would result in a non-matchable set of ego-networks; there simply wouldn't be enough egos to support the professed friendship ties.[8]

Algorithm 2 Generate Ego-Networks, Part 2

```
procedure GenEgoNets($N$, $p$, $\lambda_0$, $\lambda_1$, $\theta_0$, $\theta_1$)
    Draw $(M_0, M_1) \sim \text{Multinom}(N, (1-p, p))$        ▷ Sizes of sub-populations
    for $i = 1, \ldots, M_0$ do        ▷ Generate egonets for non-using population
        Draw $n_{0i} \sim \text{Pois}(\lambda_0)$        ▷ # of friends per non-user
        Draw $m_{0i} \sim \text{Binom}(n_{0i}, \theta_0)$        ▷ # of user friends per non-user
    end for
    for $i = 1, \ldots, M_1$ do        ▷ Generate egonets for using population
        Draw $n_{1i} \sim \text{Pois}(\lambda_1)$        ▷ # of friends for user
        Draw $m_{1i} \sim \text{Binom}(n_{1i}, \theta_1)$        ▷ # of user friends for user
    end for
```

Above, we sample the sizes of each sub-population as determined by a multinomial distribution on $1 - p$ and p. For each member of each sub-population, we draw the size of the ego-network (i.e. number of friends) n and from that count, we draw the number of friends who are substance users, m.

Algorithm 3 Generate Ego-Networks, Part 3

```
    $T_{00} \leftarrow \Sigma_{i=1}^{M_0}(n_{0i} - m_{0i})$        ▷ # of non-user ties to non-users
    $T_{01} \leftarrow \Sigma_{i=1}^{M_0} m_{0i}$        ▷ # of non-user ties to users
    $T_{10} \leftarrow \Sigma_{i=1}^{M_1}(n_{1i} - m_{1i})$        ▷ # of user ties to non-users
    $T_{11} \leftarrow \Sigma_{i=1}^{M_1} m_{1i}$        ▷ # of user ties to users
end procedure
Ensure: $T_{00}, T_{11} \bmod 2 = 0$        ▷ Even number of ties within group
Ensure: $T_{01} = T_{10}$        ▷ Cross-group tie counts should match
```

Finally, we tally the tie volume for each category of non-user to non-user, non-user to user, non-use to user, and user to user. An exact matching procedure requires the number of ties within a group to be even; otherwise, we end up with an unmatchable edge. Furthermore, the volume of ties across the sub-populations need to be equal for the same reason. We adjust the simulated sample, if necessary, to achieve these conditions.

Now that we have a sample population of egos and their professed ties to one another, we explore *all* possible matchings. The process basically entails edge resolution (i.e. finding an appropriate friend for everyone who needs one) by selecting the most constrained edge to resolve, one at a time.

[8] The closest the data comes to suggesting the size of a closed network would be the census segment or MSA status (Metropolitan Statistical Area) in which the respondents reside. We might assume the sizes of school in rural areas to be generally lesser than those in urban areas. However, in this report, we will consider N to be a free-parameter.

Algorithm 4 Match Ego-Networks, Part 1

```
Require: $V_0 = \{1, ..., M_0\}$
Require: $V_1 = \{M_0 + 1, ..., N\}$
Require: $E_0 = \{(n_{01}, m_{01}), ..., (n_{0M_0}, m_{0M_0})\}$
Require: $E_1 = \{(n_{1M_0+1}, m_{1M_0+1}), ..., (n_{1N}, m_{1N})\}$
Require: $T_{00}, T_{11} \bmod 2 = 0$    ▷ Even number of ties within group
Require: $T_{01} = T_{10}$    ▷ Fiddle, if necessary, to make # of cross-group ties equal
Require: $\boldsymbol{T} = (T_{00}, T_{01}, T_{11})$    ▷ Tuple of counts for tie types
Require: $\boldsymbol{E} = \emptyset$    ▷ Set of all possible edge matchings, initially empty
```

In order to completely explore the combination space of all possible matches, the procedure below must be recursive. At each nesting level, we provide it with sets of remaining edges, one for non-users (E_0) and another for users (E_1), the current set of matched edges, and finally a running tally of ties remaining in each category ($\boldsymbol{T}$). Finally, we store each unique matching we find in $\boldsymbol{E}$.

Algorithm 5 Match Ego-Networks, Part 2

```
procedure MatchEgoNets($E_0$, $E_1$, $E_{curr}$, $\boldsymbol{T}$)
    if $\boldsymbol{T} = (0, 0, 0)$ then    ▷ All edges have been matches
        if ordered($E_{curr}$) $\notin \boldsymbol{E}$ then    ▷ Ignore duplicates
            $\boldsymbol{E} = \{\boldsymbol{E}, \text{ordered}(E_{curr})\}$    ▷ Store complete matchings
        end if
        return
    end if
```

At the beginning of each procedure call, we check to see if, basically, we are done with the current match. If there remains no more unmatched edges, we can then check our complete matching against the set of those already found; we are interested only in the set of unique matches and not concerned with the number of ways a certain match can be attained. Exact duplicate matchings are discovered due to the inefficiency of the algorithm and has no bearing on the distributional properties of the networks; that is, the number of duplicates we can find for a given matching is a "feature" of the matching algorithm and not the social process that produces networks, in general.[9] Finding duplicates is facilitated by sorting the (i,j) edge tuples on i and then j for both pre-existing matches and newly found ones. Doing this will ensure that we compare matches by the membership composition of each ego-network, and not how each position (i.e. 1st friend, 2nd friend) is filled. Checking for duplicates efficiently

[9] However, if a social process explicitly mirrors our matching process, then duplicates need to be considered. Such a process would involve assigning counts for friends and smoking friends to each member network member and have each member seek friends randomly in order to satisfy these constraints.

is difficult since we cannot know a priori where in the recursion tree they occur; it is often the case they arise in disparate parts of the tree.

Algorithm 6 Match Ego-Networks, Part 3

```
 8:     $(u, v) \leftarrow \text{argmin}(T_{00}, T_{01}, T_{11})$          ▷ Find smallest sized tie type
 9:     if $u = 0 \wedge v = 0$ then                                   ▷ Match within non-user group
10:         $i \leftarrow \text{argmax}(n_{0i})$          ▷ Source node: has most un-matched edges of type $u$
11:         $V_{dst} \leftarrow \{x : n_{0x} - m_{0x} > 0\}$             ▷ Nodes with un-matched within-group edges
12:         $V_{dst} \leftarrow \{V_{dst} \setminus i\}$                 ▷ Remove $i$ as candidate if within group
13:         $\delta \leftarrow (-1, 0)$                          ▷ Note how to decrease count of non-using friends
14:     else if $u = 1 \wedge v = 1$ then                              ▷ Match within user group
15:         $i \leftarrow \text{argmax}(m_{1i})$
16:         $V_{dst} \leftarrow \{x : m_{1x} > 0\}$
17:         $V_{dst} \leftarrow \{V_{dst} \setminus i\}$
18:         $\delta \leftarrow (-1, -1)$                             ▷ Note how to decrease count of using friends
19:     else                                                           ▷ Match across groups
20:         $i \leftarrow \text{argmax}(m_{0i})$
21:         $V_{dst} \leftarrow \{x : n_{1x} - m_{1x} > 0\}$                  ▷ Always match from 0 → 1
22:                                                  ▷ Decreasing count of opposing group member occurs later
23:     end if
```

In general, we select a candidate source edge by giving priority to those that exhibit a higher level of constraint, so as to preclude incomplete matches (i.e. the situation in which the matching procedure is halted and left with unmatchable edges).[10] As such, we select an edge from the type category (u, v) of smallest size. From that type, we select an ego having the most number of edges of that type (lines 10, 15, and 20). We construct a set V_{dst} of target egos, considering only those egos having unmatched edges of the appropriate type. If the matching occurs within a sub-population, we, of course, ignore the source ego from consideration as a target (lines 12 and 17); reflexive friendships (i.e. friendships to self) are irrelevant. Finally, we record which parts of the ego-network tuple (total friends and/or friends who use) need to be later decremented in δ. For instance, in line 13, we decrement only the first part of the tuple, representing just the "number of friends" since we are matching only non-user to non-user edges. In the case of cross-group tie matching, we update differently (ignoring δ, line 22) since the situation entails two separate actions.

While the notation does not explicitly reveal the source of n_{ij} and m_{ij}, it is understood that these values are contained in the sets of remaining edges, E_0 and E_1, of the current call.

[10] The simplest example of a mismatch can happen with a 3-node, uncolored network. One node (#1) has degree two (ties to two other nodes) and the remaining two nodes (#2 and #3) each has degree one. The only possible matching results in #2 and #3 being partnered with #1. However, a matching process, that starts with one of the 1-degree nodes and allows it to match with the other 1-degree node, leaves #3 with unmatchable edges!

Algorithm 7 Match Ego-Networks, Part 4

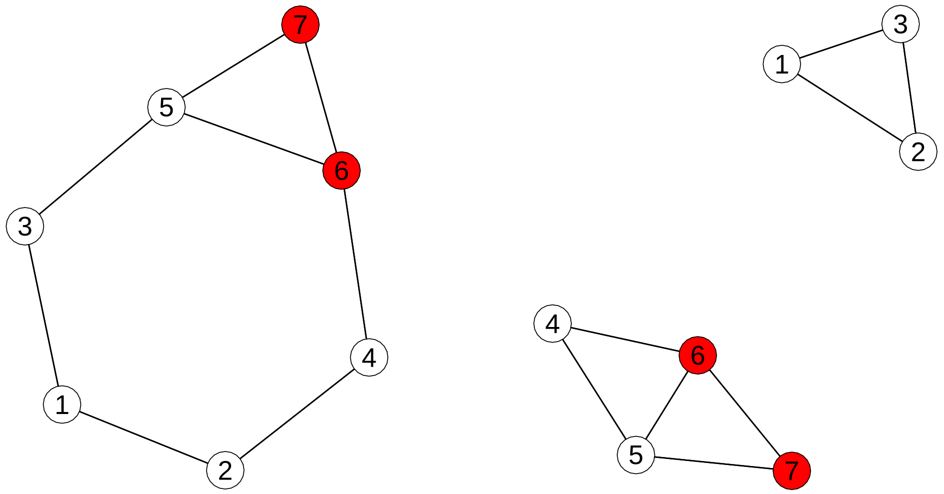
```
    for all $j \in V_{dst}$ do ▷ Delve into matching each dest. candidate one at a time
        if $u = v$ then ▷ Update remaining ties for nodes in same group
            if $i < j$ then ▷ Handle ordering just to keep notation clean
                $E_u^{new} \leftarrow \{..., (n_{ui}, m_{ui}) + \delta, ..., (n_{uj}, m_{uj}) + \delta, ...\}$
            else
                $E_u^{new} \leftarrow \{..., (n_{uj}, m_{uj}) + \delta, ..., (n_{ui}, m_{ui}) + \delta, ...\}$
            end if
        else ▷ Update ties for nodes in different groups
            $E_0^{new} \leftarrow \{..., (n_{0i}, m_{0i}) + (-1, -1), ...\}$ ▷ Update non-user to user ties
            $E_1^{new} \leftarrow \{..., (n_{1j}, m_{1j}) + (-1, 0), ...\}$ ▷ Update user to non-user ties
        end if
        $E_{curr} \leftarrow \{E_{curr}, (i, j)\}$ ▷ Add new match to current set of edges
        MatchEgoNets($E_0^{new}, E_1^{new}, \{E_{curr}, (i, j)\}, \boldsymbol{T} - (i, j)\}$)
                    ▷ Fork a new branch in the matching tree for each candidate
    end for
end procedure
```

Now that we have a candidate set of target egos V_{dst}, we iterate through each one, exploring the space of all possible matches under each edge match. We update the sets of remaining edges by employing the decrement tuple δ on the appropriate sub-population(s). In the case of a cross-group match, we decrement both members of the tuple for the non-user (matching to a using friend means we have to decrement the total number as well) and just the total number of friends for the user; since his or her match is to a non-user, the user-to-user count m_{1j} remains the same. We then call the procedure again with the sets reflecting the current match.

The algorithm is straight-forward and employs no additional optimizations, if any are possible. Often, with large set of nodes, duplicate completed networks are found, which we discard in the final count of unique, labeled structures. However, the set of unlabeled graphs is significantly smaller than the set of labeled graphs due to the lack of uniquely identifying features, which in our case are the two states for substance use and n, m composition of the ego-networks. The 7-node network we presented earlier results in 7 uniquely labeled graphs and 2 unique unlabeled graphs.

In Figure 3.3, we show two of the matched structures, one of which exactly matches the original network. However, it is easy to see how a different labeling can occur, nodes 1 and 2 can be switched around yielding the same unlabeled graph. The right graph shows how the matching process can lead to differing unlabeled structures. The network measures obtained from this graph will significantly differ from those of the original.

For a more complex example, we generate a 100 actor network using the smoker/ non-smoker λ and θ parameters for 12-year-olds and another one for 17-year-olds.



(a) Complete Network

(b) Alternate Network

Figure 3.3: Matching 7-node Network. *On the left, we supply the sample, illustrative 7 person complete network which also happens to be fully connected. On the right, another network that is consistent with the ego-networks (assuming unlabeled nodes) of the network on the left, but happens to be incorrect.*

The generated networks in Figure 3.4 show some clear differences, most of which are expected. The 12-year-old network is less dense (fewer ties) resulting in more disconnected components like isolates and separated dyads.[11]

3.3 Network Measures

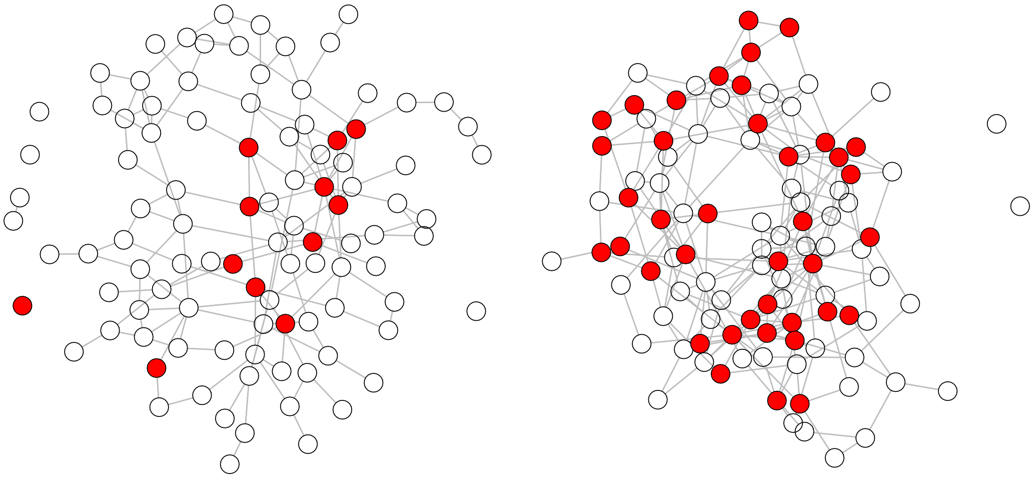
We examine the generated, completed networks with some widely-used network measures; these describe structural locations of individuals or groups relative to other groups. The first measure, betweenness centrality, describes the degree to which a node is embedded in the overall network, by highlighting the degree to which it lies in between other nodes. Individuals with high betweenness centrality scores are often considered to be boundary-spanners or gate-keepers due to their unique position of connecting disparate groups. Betweenness centrality (C_B) is formally defined defined as:

$$C_B(v) = \sum_{s \neq v \neq t \in V s \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st} is the number of shortest paths (i.e. geodesics) from nodes s to t, and $\sigma_{st}(v)$ is the number of those paths that pass through vertex v.[12] This measure may

[11] Sample networks for 13–16 year-olds can be found in Figure D.1.

[12] A geodesic is in fact the shortest path between two vertices.



(a) 12-year-olds

(b) 17-year-olds

Figure 3.4: Sample Matched Networks. *On the left, we display a simulated population of 100 12-year-olds and on the right we have 100 17-year-olds, maintaining the correct proportions of smokers per age group, using age specific definitions of "use": lifetime use (CIGFLAG) and past year use (CIGYR), respectively. The respective proportions of use are 0.12 and 0.39. Dark colored circles represent cigarette smokers.*

be normalized by dividing through by the number of pairs of vertices not including v, which is $(n-1)(n-2)$. The betweenness centrality scores for isolates or dyads is zero as vertices in those structures do not lie in between anyone. We look at betweenness measures for nodes in the main component as well as the entire population (which incurs a bias towards 0 from those nodes not in the main component):

	Betweenness > 0			All Betweenness + 1		
12-year-olds	μ_{C_B}	σ_{C_B}	$\mathcal{L}$	μ_{C_B}	σ_b	$\mathcal{L}$
non-smoker	4.40	1.30	-407.61	3.37	2.19	-490.49
smoker	5.42	0.81	-72.83	4.97	1.68	-82.94

17-year-olds	μ_{C_B}	σ_{C_B}	$\mathcal{L}$	μ_{C_B}	σ_{C_B}	$\mathcal{L}$
non-smoker	4.18	1.07	-317.42	3.86	1.52	-347.76
smoker	4.45	1.08	-231.99	4.47	1.05	-231.80

Above are the results of a lognormal fit to the distributions of betweenness for non-smokers and smokers per age group; the μ and σ fit the log of the original data.[13] Under both categories of betweenness measurements, smokers are more embedded in the social network than non-smokers, though the disparity shrinks with older cohorts. However, the results are not significantly different. In order to conduct a more robust

[13] Employing the lognormal requires us to offset the betweenness measure of the entire network.

analysis, we look now at a broader set of results, summarizing betweenness for just the main component from 100 simulated networks for each age group, we find notable differences between non-smokers and smokers:

	Non-Smokers			Smokers						
Age	$\overline{\mu}_{C_B}$	$\overline{\sigma}_{C_B}$	df	$\overline{\mu}_{C_B}$	$\overline{\sigma}_{C_B}$	df	n_0	n_1	t	p
12	4.567	1.394	211	5.372	0.992	120	88	12	-2.50	**0.023**
13	4.388	1.255	245	4.971	1.097	162	79	21	-2.10	**0.043**
14	4.260	1.257	211	4.747	1.136	171	71	29	-1.89	**0.064**
15	4.325	1.201	283	4.555	1.131	192	63	37	-0.96	0.340
16	4.280	1.185	290	4.408	1.191	241	57	43	-0.53	0.595
17	4.180	1.219	338	4.497	1.148	236	61	39	-1.32	0.192

In assessing the differences, a conservative t-test is employed, in which the degrees of freedoms are dictated by merely the average of the counts of non-user and users. The deviations are computed according to those of imputation-based methods (Rubin, 1987).[14] We find modest significance (below the 0.10 level, indicated by the bold-typed p values) in the ways smokers and non-smokers are embedded in the network with younger smokers being more centrally located than their older counterparts, hence, their higher betweenness scores. However, an increase in both friends' count and smoking friends, as youths in both sub-populations age, reduces the distinctiveness of their ego-networks, as evidenced by the lack of significant differences in this network measure for the older youths.

Closeness centrality (C_C) is modestly related to betweenness; it describes the degree to which vertices are apart (or close) to all other vertices. One way of capturing closeness is by the converse concept "farness" (C_F), which is defined as the mean geodesic distances between a vertex and all other vertices:

$$C_F(v) = \frac{1}{C_C(v)} = \sum_{t \in V \backslash v} \frac{d_G(v,t)}{n-1}$$

where n is the number of vertices $|V|$ and $n \geq 2$. Closeness (C_C) is then merely the reciprocal of farness.

[14] The total variance, T, of an estimate combines the between-sample and within-sample variances:

$$T = W + \frac{K+1}{K} B$$

where K is number of simulations (100) and

$$B = \frac{1}{K-1} \sum_{k=1}^{K} \left(\mu_{C_{B,k}} - \overline{\mu}_{C_B}\right)$$

and

$$W = \frac{1}{K} \sum_{k=1}^{K} \sigma^2_{C_B,k}$$

	Farness				Closeness			
	Main Comp.		All		Main Comp.		All	
12 year olds	μ_{Fc}	σ_{Fc}	μ_F	σ_F	μ_{Cc}	σ_{Cc}	μ_C	σ_C
non-smoker	4.42	0.74	15.30	20.82	0.232	0.037	0.093	0.021
smoker	3.73	0.42	17.10	26.11	0.271	0.029	0.097	0.028

17 year olds	μ_{Fc}	σ_{Fc}	μ_F	σ_F	μ_{Cc}	σ_{Cc}	μ_C	σ_C
non-smoker	3.42	0.48	8.47	17.00	0.298	0.040	0.182	0.035
smoker	3.34	0.38	5.29	0.37	0.303	0.032	0.190	0.013

The difference in closeness centrality for 12 year-old smokers and non-smokers is significant, t-value $= 4.85$, $p < 0.001$. The higher closeness/lesser farness of smokers again indicates their centralized positions in the main component, relative to non-smokers. We can also look at within- and between-group closeness:

	non-smoker		smoker	
12 year-old	μ_{Fc}	σ_{Fc}	μ_{Fc01}	σ_{Fc}
non-smoker	4.50	0.740	4.24	0.981
smoker	3.90	0.414	2.73	0.657

17 year-old	μ_{Fc}	σ_{Fc}	μ_{Fc01}	σ_{Fc}
non-smoker	3.43	0.492	3.49	0.518
smoker	3.46	0.450	3.24	0.340

There is more differentiation in structure between 12-year-old smokers and non-smokers, while this dissipates with increasing age, which is expected given how increasingly more of the population smokes, thereby lessening the distinction between smokers and non-smokers. Interestingly, the closeness of 12 year-old smokers to non-smokers is not the same as that of smokers to non-smokers, which is possible given the disparate sizes of the sub-populations. We can further look at these distinctions in both farness/closeness for each age group drawing our measures from a distribution of 100 simulated networks, as was done for betweenness earlier:

Farness:

	Non-Smokers			Smokers						
Age	μ_{F_c}	σ_{F_c}	df	μ_{F_c}	σ_{F_c}	df	n_0	n_1	t	p
12	4.604	0.856	123	4.139	0.809	111	88	12	1.85	**0.084**
13	3.995	0.633	153	3.680	0.587	132	79	21	2.15	**0.039**
14	3.727	0.557	165	3.499	0.513	144	71	29	1.97	**0.054**
15	3.626	0.530	162	3.526	0.527	154	63	37	0.91	0.365
16	3.457	0.478	135	3.443	0.507	140	57	43	0.14	0.887
17	3.482	0.508	147	3.309	0.444	145	61	39	1.80	**0.075**

Closeness:

	Non-Smokers			Smokers						
Age	μ_{C_c}	σ_{C_c}	df	μ_{C_c}	σ_{C_c}	df	n_0	n_1	t	p
12	0.224	0.036	1021	0.249	0.037	400	88	12	-2.20	**0.045**
13	0.256	0.037	2086	0.278	0.038	982	79	21	-2.34	**0.026**
14	0.274	0.037	2461	0.291	0.038	1273	71	29	-2.13	**0.038**
15	0.281	0.037	2194	0.289	0.038	1565	63	37	-1.03	0.307
16	0.294	0.036	976	0.296	0.039	1117	57	43	-0.23	0.815
17	0.293	0.038	1366	0.307	0.037	1094	61	39	-1.90	**0.061**

Networks become more dense as youths age and form more and more friendship ties with one another; hence, the closeness between everyone will increase, as we see in the above results. Smokers and non-smokers remain distinct in their closeness or reachability to others until about age 15, at which point smokers and non-smokers mix in a such a fashion as to homogenize their friendship ties. Recalling that peak smoking initiation occurs just around that age, we would hypothesize that these first time users' networks are also distinct from those of non-users in such a way as to facilitate influence forces. As smokers and non-smokers networks become more entwined after the age of 14, smoking influence becomes diluted; hence we also should see a sudden drop in initiation rates. We explore this association later in this chapter.

An alternative definition of closeness exponentially weights the geodesic, giving little weight to disconnected vertices (Dangalchev, 2006):

$$C_C(v) = \sum_{t \in V \setminus v} 2^{-d_G(v,t)}$$

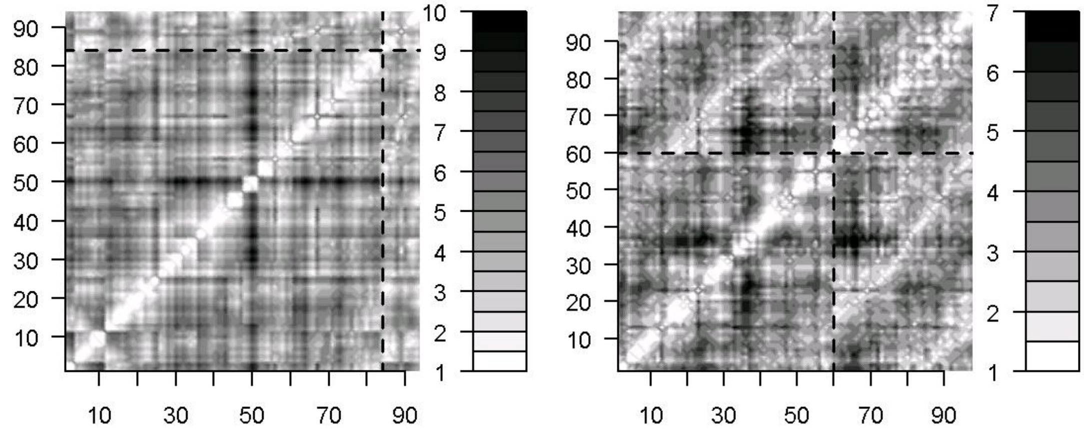
	Closeness, All	
12 year olds	μ_{gc}	σ_{gc}
non-smoker	7.10	3.17
smoker	9.71	3.87

17 year olds	μ_{gc}	σ_{gc}
non-smoker	11.95	3.84
smoker	13.02	2.50

Here, the higher closeness values indicate actual closeness. While the differences within age-groups are not significant, the smokers' closeness are consistently higher than those for non-smokers, under this alternative closeness measure.

	Non-Smokers			Smokers						
Age	μ_{C_c}	σ_{C_c}	*df*	μ_{C_c}	σ_{C_c}	*df*	n_0	n_1	t	p
12	7.716	3.578	114	9.797	4.711	103	88	12	-1.47	0.165
13	10.082	3.767	112	11.741	4.622	105	79	21	-1.52	0.141
14	11.349	3.758	115	12.817	4.143	108	71	29	-1.65	0.105
15	11.941	3.708	116	12.474	4.101	112	63	37	-0.65	0.518
16	12.980	3.855	109	13.074	4.075	110	57	43	-0.12	0.907
17	12.820	3.969	110	14.079	3.907	108	61	39	-1.56	0.122

Again, we generate 100 sample networks, collate the results, and conservatively t-test the differences for Dangalchev's closeness. While the differences do not achieve significance, even at the 0.10 level, the p-values still suggest a the shift in structure from age 14 to 15.



(a) 12-year-olds

(b) 17-year-olds

Figure 3.5: Closeness Between Nodes. *The geodesics (shortest paths) between all non-isolated nodes (in the main component) to all others appear in shades of gray; lighter color indicates higher closeness. The dashed lines denote the sub-population partition between non-smokers, left and below the partitions, and smokers, to the right and above the partitions. The axes differ slightly due to there being more isolates in the 12-year-old sub-population. The graph is undirected hence the distances and plot are symmetric.*

In Figure 3.5, the features of a core/periphery network structure (Borgatti and Everett, 1999; Boyd et al., 2006) are shown to be prominent for the network of 12 year-olds; the dark streaks denote members who are essentially outliers, far from almost everyone. The more dense 17 year-old network displays a more complex structure of multiple clusters (i.e. the streams of white), some of which overlap and some of which do not (i.e. the dark patches). Still, the non-user component of this network does seem to exhibit core/periphery features.

Watts' clustering coefficient describes the degree to which a network displays qualities of a small-world network, generally comprising loosely connected dense clusters (Watts and Strogatz, 1998; Watts, 1999a,b). At the core of this measure is the degree to which an ego's alters are connected divided by the number of possible connections among them. For a graph with vertices $V = v_1, v_2, \ldots, v_n$ and a set of edges E where e_{ij} denotes an edge between vertices v_i and v_j, a neighborhood for a vertex v_i is defined as:

$$N_i = \{v_j\} : e_{ij} \vee e_{ji} \in E$$

and k_i is the degree of vertex v_i or $|N_i|$. Then, the clustering coefficient of vertex v_i is:

$$C_W(v_i) = \frac{|\{e_{jk}\}|}{k_i(k_i - 1)} : v_j, v_k \in N_i, e_{jk} \in E$$

The clustering coefficient of the entire graph is simply the mean of the measure for each vertex:

$$C_W = \frac{1}{n} \sum_{i=1}^{n} C_W(v_i)$$

We compute the clustering coefficients for each of the simulated networks:

12-year-olds	μ_{C_W}	σ_{C_W}	$p(\mu_{C_W} < X)$	density
non-smoker	0.124	0.203	0.00	—
smoker	0.009	0.022	0.90	—
all	0.111	0.194	0.00	0.0299

17-year-olds	μ_{C_W}	σ_{C_W}	$p(\mu_{C_W} < X)$	density
non-smoker	0.108	0.144	0.00	—
smoker	0.107	0.140	0.00	—
all	0.107	0.142	0.00	0.0477

Both networks as a whole (i.e. the 'all' rows) exhibit a strong degree of clustering, more so than random; the significance values p show how the measures compare to those obtained from distributions of Bernoulli random graphs with identical densities.[15] What is surprising is the near-significance of the 12-year-old smoker; the measure suggests that really young smokers exist in less cohesive clusters than the norm. In a sense, they are still outsiders, despite their apparent embededdness by being connected to the popular peers. Ennett and Bauman (1993) find that smokers are more likely to be 'fringe' members or 'isolates' of the network, instead of 'liaisons', which is not quite consistent with what we find in our analysis. We obtain

[15] The assessment of the significance of graph-level measures by employing Bernoulli random graphs to generate a null hypothesis distribution is explored in depth in Anderson et al. (1999).

more robust measures of age and substance use specific clustering coefficients from the distribution of generated networks:[16]

	Non-Smokers			Smokers						
Age	μ_{C_W}	σ_{C_W}	df	μ_{C_W}	σ_{C_W}	df	n_0	n_1	t	p
12	0.174	0.268	549	0.033	0.064	210	88	12	4.15	0.0001
13	0.188	0.248	534	0.058	0.104	283	79	21	3.62	0.0005
14	0.178	0.232	535	0.086	0.136	355	71	29	2.47	0.0156
15	0.158	0.209	474	0.081	0.137	386	63	37	2.22	0.0286
16	0.144	0.191	424	0.101	0.163	428	57	43	1.22	0.2236
17	0.161	0.208	448	0.129	0.155	308	61	39	0.90	0.3708

While earlier we saw early age smokers have higher closeness than their non-smoking peers suggesting that they are more central in their respective networks. However, the lower densities of their ego-networks, as indicated by the lower clustering coefficients, suggest, more firmly here, that while smokers are central, they are not tightly embedded and instead exist interstitially between more cohesive structures such as cliques, supporting the observation that they are 'liaisons' and not isolates or on the fringe.

Again, comparing the Watts' clustering coefficients from the generated network to those coefficients measured from Bernoulli random graphs gives us a sense of how much non-random social structuring occurs in sub-populations:

	Everyone		Non-Smokers		Smokers		
Age	μ_{C_W}	p	μ_{C_W}	p	μ_{C_W}	p	density
12	0.157	0.0001	0.174	0.0000	0.033	0.4987	0.0305
13	0.160	0.0002	0.188	0.0001	0.058	0.2754	0.0370
14	0.151	0.0002	0.178	0.0000	0.086	0.0827	0.0408
15	0.129	0.0017	0.158	0.0003	0.081	0.1332	0.0416
16	0.125	0.0030	0.144	0.0086	0.101	0.0540	0.0444
17	0.149	0.0007	0.161	0.0010	0.129	0.0148	0.0463

And in fact, we again find that younger smokers exist in less structured local networks. Yet due to their high embeddedness as evidenced by their high betweenness scores,

[16] The clustering coefficient is distributed between 0 and 1 and upon observation, a beta distribution seems appropriate to model the sampled measures. We report shape parameters, α and β, for the distributions surrounding each of the reported measures:

	Everyone		Non-Smokers		Smokers	
Age	α_{C_W}	β_{C_W}	α_{C_W}	β_{C_W}	α_{C_W}	β_{C_W}
12	13.88	74.32	12.58	59.55	0.65	19.00
13	16.13	84.44	14.80	64.06	2.04	33.41
14	18.51	104.05	14.98	69.29	4.26	45.56
15	13.75	92.72	11.83	63.21	3.66	41.90
16	15.54	108.45	9.35	55.58	5.92	52.96
17	15.21	87.16	11.63	60.51	7.09	48.06

their local networks are less dense. Still, these observations lend support to the claim that smoking is more stigmatic for younger teens; these young smokers may be linked to the central social groups, but are not deeply embedded in them. With older teens, we find a reversal of this trend; smoking teens now exist in structured networks and non-smoking teens less so, in comparison. One explanation for this observation that warrants further investigation is that, as teens age, smoking becomes more accepted and more normative, while anti-smoking sentiments becomes frowned upon.

As mentioned earlier, there is a noticeable shift in how all the reviewed network measures, that highlight structural differences between smokers and non-smokers, shift around the age of 15. Specifically, the differences in mean centralities significantly maintain from age 12 to 14 and suddenly vanish from age 15 or 16. It was suggested that the networks of smokers and non-smokers, despite their parameters continuing to be visibly distinct, begin to merge, as indicated by the increasing tie volume between the user and non-user groups. This shift apparently, but not surprisingly, has a behavioral consequence:

prop. initiating = $p(y^{t+1}_{\text{CIGFLAG}} = 1)$ $- p(y^{t}_{\text{CIGFLAG}} = 1)$	Age	Differences in $\overline{C}$ between smokers and non-smokers; t-test significance:		
		$p_{\Delta\overline{C}_B}$	$p_{\Delta\overline{C}_C}$	$p_{\Delta\overline{C}_W}$
0.092	12	0.023	0.045	0.000
0.109	13	0.043	0.026	0.001
0.118	14	0.064	0.038	0.016
0.071	15	0.340	0.307	0.029
0.071	16	0.595	0.815	0.224
	17	0.192	0.061	0.371

In the above table, the shift in significance in the centrality and clustering *differences* between smokers and non-smokers corresponds to a sudden shift in the rates of initiation. Just after the age of 15, the initiation rate drops from 11.8% of the population to just 7.1%, and concomitantly, the differences in the mean betweenness and closeness centralities cease to be significant. Watts' clustering coefficient follows suit at age 16. This parallel shift suggests that the enmeshing of smokers and non-smokers into increasingly heterogeneous social groups, then, dffuses much of the smoking influence; hence, we see the drop in initiation rates with increasing age.

Earlier, it was suggested that the ego-network matching endeavor might be rendered useless if we found no significant structural differences. Not only do we demonstrate the opposite, but we also find a pattern of structural differences (and, for later ages, structural similarities) that mirror and explain the changing inertia of influence. These findings highlight the suitability of ego-network matching in inferring complete networks from the NSDUH data and warrant further efforts into the matching methodology.

Chapter 4

Joint Poly-Substance Analysis

4.1 Alcohol and Marijuana: Marginal Analysis

So far we have focused solely on cigarette smoking. We now introduce analysis on the remaining two substances, alcohol and marijuana, for which NSDUH respondents provided ego-network data. Population proportions of recent use of all three substances are:

Used in ...	Cigarettes	Alcohol	Marijuana
Lifetime	0.364	0.401	0.179
Past Three Years	0.310	0.390	0.174
Past Year	0.236	0.329	0.141
Past Month	0.166	0.178	0.077

Not surprisingly, marijuana use remains far below the other two substances, most likely due to its illicit status. However, while differences in proportions for cigarette smoking between the recency categories are evenly spaced, those for alcohol and marijuana are not: the proportion for lifetime and past three year use for these are almost identical.[1] We give the empirical data on alcohol and marijuana recency

[1] A breakdown of the proportions by age suggests that after initiation, some level of use of alcohol and marijuana will continue while, individuals are more likely to cease smoking cigarettes:

	CIG...				ALC...				MRJ...			
Age	FLAG	RC3	YR	MON	FLAG	RC3	YR	MON	FLAG	RC3	YR	MON
12	2	3	3	4	1	3	4	3	0	0	1	1
13	3	4	6	8	1	5	10	7	0	1	2	2
14	4	9	7	13	1	5	15	13	0	2	4	6
15	7	9	9	20	1	6	18	23	0	3	8	10
16	8	10	8	25	1	8	22	28	1	6	11	12
17	9	10	9	31	1	9	22	34	1	8	12	16

The data in this table reflects the a breakdown of the Recency variable introduced in the preliminary logistic regressions in Table 2.1, in which the overlap between the categories is removed (e.g. CIGRC3 now exclusively means "smoked in the past three years *but* not in the past year or month"). The indicators the same logistic regression treatment as we did for tobacco (back in Table 2.1) and find similar results: that while all covariates significantly predict self use, friends' use of the substance is the primary predictor.[2] However, gender and adults' use is more predictive of marijuana use than it is for smoking or drinking alcohol, with coefficients that are roughly twice in magnitude and more significant. Again, the illicit status of marijuana offers the likely explanation: boys are more likely to engage in illegal behavior than girls and, furthermore, some contact with adult's who use marijuana would precede initiation given the difficulties in obtaining that substance.

We conduct a preliminary Poisson/binomial fit on each of the ego-network variables for alcohol and marijuana (FDALC and FDMJ):

	λ	θ	μ	σ_λ	σ_θ	σ_μ	$\mathcal{L}$	n
Tobacco	3.32	0.326	1.09	0.0230	0.00221	0.00800	-783.56	25052
Alcohol	2.91	0.345	1.01	0.0203	0.00238	0.00756	-1092.05	24959
Marijuana	2.60	0.234	0.61	0.0232	0.00237	0.00562	-1131.26	24949

The estimates are potentially problematic because of the significantly differing rates of peer network size, λ; this inconsistency underscores some inexactness in our mixture approach to estimating the parameters. However, the large negative log-likelihoods for alcohol and marijuana suggest the differences might be explained under a model that estimates for sub-populations. It seems the case that a relatively bad fit is due to a heterogeous population; recall earlier, our fit on friends' smoking behavior was vastly improved (i.e. the log-likelihood increased) when we explored different sub-populations for the friends' smoking response (FDCIG). Later, we will find this inconsistency corrected when we estimate the parameters jointly.

If we merge the likelihood function for alcohol and marijuana to include a fit to the population or age-population prevalence of recent use as well as a fit to reciprocating ties, as we did for tobacco use in the last chapter, we can obtain preliminary definitions of their "use":

reported quantities are rounded percentages of each age population; the rows will not sum to 100 because we omit the percentages of individuals who have never tried the substance. The increasing CIGFLAG implies that more smokers return to a state of relative non-smoking (beyond three years) more so than those who have used alcohol or marijuana: the populations associated with their exclusive FLAG indicators do not increase. Note, recent initiates do not contribute to the FLAG category, but instead will fall into either the past year (YR) or past month (MON) categories. The slightly later ages of initiation for alcohol and marijuana would only partly account for the pattern we are seeing here.

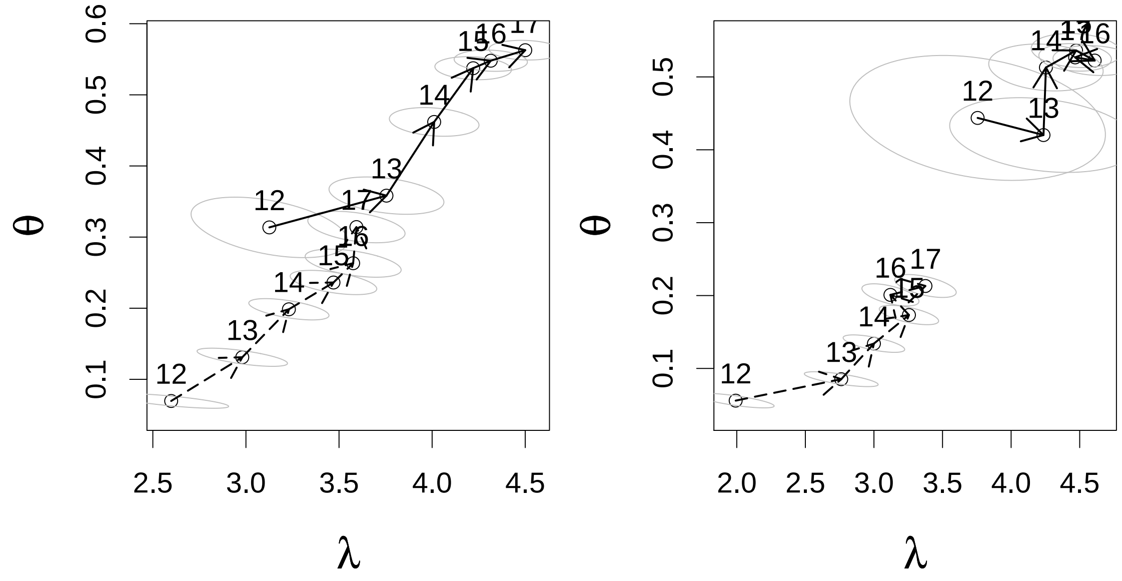
[2] Results are reported in Tables F.1 and F.2 in Appendix F.

	$\mathcal{L}_{dt}$ for ALC...				$\mathcal{L}_{dti}$ for ALC...			
y_{AGE}	FLAG	RC3	YR	MON	FLAG	RC3	YR	MON
12	**-69**	**-68**	**-69**	-90	-74	**-71**	-89	-199
13	**-154**	**-152**	-169	-187	-171	**-164**	-177	-361
14	**-123**	-128	**-125**	-178	-148	-143	**-131**	-409
15	**-94**	**-93**	**-95**	-131	-164	-152	**-107**	-308
16	-72	-68	**-63**	-80	-214	-190	**-90**	-248
17	-79	-76	-66	**-61**	-264	-249	**-113**	-181
$\Sigma\mathcal{L}$	-591	**-585**	**-588**	-726	-1035	-970	**-707**	-1706

$\mathcal{L}_{dt}$ reports the fits on the ego-network (FDALC) and matching ties between users and non-users. $\mathcal{L}_{dti}$ extends that two-fold fit to include a fit to the sub-population prevalence of recent use. Again, the bold-typed log-likelihoods highlight the best fits per age group. Inclusion of population prevalence into the likelihood ($\mathcal{L}_{dti}$) clarifies the pattern of "use" to mean recent within three years, for 12 and 13 year-olds, and past year use for 14-17 year-olds. This clearly differs from the "use" pattern of cigarette smoking, which graduated from lifetime use (CIGFLAG) to recent use within three years (CIGRC3). The overall best indicator of "drinking alcohol" is past year use (ALCYR). Since prevalence of alcohol consumption is higher than tobacco use in our population, the corresponding definition of "use" is expected to be associated with a more recent indicator.

	$\mathcal{L}_{dt}$ for MRJ...				$\mathcal{L}_{dti}$ for MRJ...			
y_{AGE}	FLAG	RC3	YR	MON	FLAG	RC3	YR	MON
12	**-78**	**-78**	**-79**	-100	**-127**	**-129**	-143	-224
13	**-114**	-119	-124	-139	**-163**	-170	-194	-282
14	**-108**	**-109**	-116	-160	**-147**	-150	-200	-384
15	**-88**	**-87**	-103	-139	**-100**	**-102**	-151	-388
16	**-92**	**-91**	**-93**	-142	**-96**	**-96**	-123	-412
17	**-57**	-61	-64	-96	**-68**	**-68**	-87	-329
$\Sigma\mathcal{L}$	**-536**	-545	-580	-776	**-701**	-715	-898	-2019

With marijuana use, we find a more liberal definition of "use", one that leans towards any level of use (MRJFLAG). Again, this is not surprising; since marijuana use is far less prevalent than tobacco or cigarette use, we would expect the appropriate indicator to be more inclusive than that of tobacco or alcohol. In later analysis, when we need select non-age-specific indicators of use, we will accordingly employ the indicators for past three years' use of cigarettes (CIGRC3), past year use of alcohol (ALCYR), and lifetime use of marijuana (MRJYR). In Figure 4.1, the age trajectories between at-least-once-users and non-users are shown to be distinct, especially for marijuana use. These substances, along with cigarette smoking, have dissimilar connotations in society, which perhaps mirror the magnitudes in the differences among these trajectories. However, it is impossible to tell at this point, for marijuana use, whether we are seeing teens with using friends rapidly initiating or simply having



(a) Lifetime Alcohol Use

(b) Lifetime Marijuana Use

Figure 4.1: Trajectories of Alcohol and Marijuana. *The solid line denotes changes in the λ and θ parameters for at-least-once-used adolescents and the dashed line for never-tried adolescents. The gray ellipses denote the 95% probability region around the standard errors.*

stronger homophilic tendencies towards peers who also have used at some point; this issue is addressed in the next chapter. Given that alcohol consumption and cigarette smoking is more prevalent in our culture than marijuana use, we naturally see more overlap in their trajectories.

4.2 Joint Analysis of Two Substances

Since each respondent provided ego-network data on all three substances, we can estimate λ and θ jointly. For the sake of clarity, we will discuss the joint analysis of two substances at a time, and then, in the following section, perform our analysis on all three substances at once. We first present the joint data on friends' usage of cigarettes and alcohol in Table 4.1. The data suggests a possible pattern of co-substance use among friends; respondents are more likely to report similar levels friends' use for both substances than not. We can confirm this with the contingency data comprising each of the four categories of joint use:

Friends' Smoke	Friends' Drinking (FDALC)							
	weighted frequency				%			
(FDCIG)	None	Few	Most	All	None	Few	Most	All
None	8035	1284	188	66	32.3	5.2	0.8	0.3
Few	2225	6442	1636	257	8.9	25.9	6.6	1.0
Most	230	1089	2130	432	0.9	4.4	8.6	1.7
All	42	130	220	478	0.2	0.5	0.9	1.9

Table 4.1: Joint Friends' Use of Cigarettes and Alcohol. *In order to improve readability, the proportions have been converted into percentages.*

		Drank Alcohol in Past Year? (ALCYR)			
		weighted freq		%	
		No	Yes	No	Yes
Smoked in Past Three Years?	No	14634	2931	57.5	11.5
(CIGRC3)	Yes	2441	5458	9.6	21.4

And, there is indeed a significant co-use pattern in our population; the χ^2 for the table is an extreme 6775, with $p < 0.001$. The marginal levels of use are 31.0% for cigarette smoking and 32.9% for alcohol consumption. Even with a more restrictive definition of alcohol consumption, its perceived prevalence of "use" is higher than that of cigarette smoking.

With the four categories of tobacco/alcohol use, the set of friend types now expands as well to four categories. That is, the single parameter θ for friends' use of a single substance now expands to four parameters for each combination of use/non-use:

$$\theta_{00} + \theta_{01} + \theta_{10} + \theta_{11} = 1$$

where $\theta_{ij} = p(z_{\text{CIGRC3}} = i, z_{\text{ALCYR}} = j)$ and z is use behavior of some hypothetical friend. p and θ express the probability that a friend smokes, drinks, or does both. The marginal probabilities of use among friends are:

$$p(z_{\text{CIGRC3}} = 1) = \theta_{10} + \theta_{11} = \theta_{1\cdot}$$

and

$$p(z_{\text{ALCYR}} = 1) = \theta_{01} + \theta_{11} = \theta_{\cdot 1}$$

The likelihood is now expressed as $\mathcal{L}(\boldsymbol{\theta} = (\theta_{00}, \theta_{10}, \theta_{01}, \theta_{11})|\boldsymbol{n}_{\text{FD}})$, where $\boldsymbol{n}_{\text{FD}}$ contains all sixteen joint proportions from Table 4.1. Deriving it follows the same strategy as that for smoking; however, the addition of a single substance increasingly complicates the procedure. We will need to explain the likelihood in multiple parts. Furthermore, the focal parameter is $\boldsymbol{\theta}$ and not the prior distribution for λ, so we will assume a fixed $n_{friends}$ for now, which we will refer to as just n.

The first and easiest part of the joint distribution is assessing the probability that the respondent claims one or both of the friends' use level is 'None'. Given a predetermined number of friends, n, and a hypothetical rate of friends' use, $\boldsymbol{\theta}$ $= \{\theta_{00}, \theta_{10}, \theta_{01}, \theta_{11}\}$ for two (unspecified) substances, and the friends' use responses $x_0, x_1 \in \{$'None', 'Few', 'Most', 'All'$\}$; if we want to know the likelihood of either x_0 = 'None' or x_1 = 'None', then:

$$p(y_{\text{FDCIG}} = x_0, y_{\text{FDALC}} = x_1 | n, \boldsymbol{\theta}) =$$

$$\begin{cases} p(n_{00} = n, \qquad n_{ij} = 0, n_{ji} = 0, n_{11} = 0) & \text{if } x_i = \text{`None'} \\ \displaystyle\sum_{k=1}^{\lfloor n/2 \rfloor} p(n_{00} = n-k, n_{ij} = 0, n_{ji} = k, n_{11} = 0) & \text{if } x_i = \text{`Few'} \\ \displaystyle\sum_{k=\lfloor n/2 \rfloor+1}^{n-1} p(n_{00} = n-k, n_{ij} = 0, n_{ji} = k, n_{11} = 0) & \text{if } x_i = \text{`Most'} \\ p(n_{00} = 0, \qquad n_{ij} = 0, n_{ji} = n, n_{11} = 0) & \text{if } x_i = \text{`All'} \end{cases}$$

where

$$(i,j) = \begin{cases} (1,0) & \text{if } x_0 = \text{`None'} \\ (0,1) & \text{if } x_1 = \text{`None'} \end{cases}$$

which expands to the following multinomial probabilities:

$$\begin{cases} \dbinom{n}{n,0,0,0} \theta_{00}^{n}\, \theta_{ij}^{0}\, \theta_{ji}^{0}\, \theta_{11}^{0} & \text{if } x_i = \text{`None'} \\ \displaystyle\sum_{k=1}^{\lfloor n/2 \rfloor} \dbinom{n}{n-k,0,k,0} \theta_{00}^{(n-k)}\, \theta_{ij}^{0}\, \theta_{ji}^{k}\, \theta_{11}^{0} & \text{if } x_i = \text{`Few'} \\ \displaystyle\sum_{k=\lfloor n/2 \rfloor+1}^{n-1} \dbinom{n}{n-k,0,k,0} \theta_{00}^{(n-k)}\, \theta_{ij}^{0}\, \theta_{ji}^{k}\, \theta_{11}^{0} & \text{if } x_i = \text{`Most'} \\ \dbinom{n}{0,0,n,0} \theta_{00}^{0}\, \theta_{ij}^{0}\, \theta_{ji}^{n}\, \theta_{11}^{0} & \text{if } x_i = \text{`All'} \end{cases}$$

When one of the friends' use variables is 'None', we expect its marginal sum, n_{ij} + n_{11}, to be zero. Subsequently, which use categories the friends fall into will be completely determined by the non-'None' variable, n_{ji}. If the second variable is also 'None', then we predict for all n friends falling into n_{00}.

If we want the probability for either x_0 = 'All' or x_1 = 'All', then:

$$p(y_{\text{FDCIG}} = x_0, y_{\text{FDALC}} = x_1 | n, \boldsymbol{\theta}) =$$

$$\begin{cases} p(n_{00} = 0, n_{ij} = n, \qquad n_{ji} = 0, n_{11} = 0) & \text{if } x_i = \text{`None'} \\ \displaystyle\sum_{k=1}^{\lfloor n/2 \rfloor} p(n_{00} = 0, n_{ij} = n-k, n_{ji} = 0, n_{11} = k) & \text{if } x_i = \text{`Few'} \\ \displaystyle\sum_{k=\lfloor n/2 \rfloor+1}^{n-1} p(n_{00} = 0, n_{ij} = n-k, n_{ji} = 0, n_{11} = k) & \text{if } x_i = \text{`Most'} \\ p(n_{00} = 0, n_{ij} = 0, \qquad n_{ji} = 0, n_{11} = n) & \text{if } x_i = \text{`All'} \end{cases}$$

where

$$(i,j) = \begin{cases} (1,0) & \text{if } x_0 = \text{`All'} \\ (0,1) & \text{if } x_1 = \text{`All'} \end{cases}$$

which expands to:

$$\begin{cases} \binom{n}{0,n,0,0} \theta_{00}^{0}\, \theta_{ij}^{n}\, \theta_{ji}^{0}\, \theta_{11}^{0} & \text{if } x_i = \text{`None'} \\ \sum_{k=1}^{\lfloor n/2 \rfloor} \binom{n}{0,n-k,0,k} \theta_{00}^{0}\, \theta_{ij}^{(n-k)}\, \theta_{ji}^{0}\, \theta_{11}^{k} & \text{if } x_i = \text{`Few'} \\ \sum_{k=\lfloor n/2 \rfloor+1}^{n-1} \binom{n}{0,n-k,0,k} \theta_{00}^{0}\, \theta_{ij}^{(n-k)}\, \theta_{ji}^{0}\, \theta_{11}^{k} & \text{if } x_i = \text{`Most'} \\ \binom{n}{0,0,0,n} \theta_{00}^{0}\, \theta_{ij}^{0}\, \theta_{ji}^{0}\, \theta_{11}^{n} & \text{if } x_i = \text{`All'} \end{cases}$$

Here, we predict for one of the friends' use variables being 'All'. In this case, the marginal sum, $n_{ij} + n_{11}$ needs to equal the entire set of friends n. The degree of use of the second variable is then determined by n_{11} which expresses both substances. n_{ji} which expresses the second substance is disqualified because it contradicts the complete presence of the first substance.

If both $x_0 =$ 'Few' and $x_1 =$ 'Few', then:

$$\begin{aligned} & p(y_{\text{FDCIG}} = \text{`Few'}, y_{\text{FDALC}} = \text{`Few'} | n, \boldsymbol{\theta}) \\ &= \sum_{i=0}^{\lfloor n/2 \rfloor} \sum_{j=\mathcal{I}_0(i)}^{\lfloor n/2 \rfloor - i} \sum_{k=\mathcal{I}_0(i)}^{\lfloor n/2 \rfloor - i} p(n_{00} = n - i - j - k, n_{10} = j, n_{01} = k, n_{11} = i) \\ &= \sum_{i=0}^{\lfloor n/2 \rfloor} \sum_{j=\mathcal{I}_0(i)}^{\lfloor n/2 \rfloor - i} \sum_{k=\mathcal{I}_0(i)}^{\lfloor n/2 \rfloor - i} \binom{n}{(n-i-j-k), j, k, i} \theta_{00}^{(n-i-j-k)}\, \theta_{10}^{j}\, \theta_{01}^{k}\, \theta_{11}^{i} \end{aligned}$$

where

$$\mathcal{I}_0(i) = \begin{cases} 1 & \text{if } i = 0 \\ 0 & \text{otherwise} \end{cases}$$

The $y_{\text{FDCIG}} =$ 'Few' and $y_{\text{FDALC}} =$ 'Few' denotes of a mixture of friends in n_{01}, n_{10}, and n_{11}, or just n_{11} alone; it does not necessarily mean the respondent has a few friends who both smoke cigarettes and drink alcohol. For instance, if a teen has one friend who only smokes and another who only drinks, the response would still be 'Few' and 'Few'. The indicator function $\mathcal{I}$ reflects the two minimum conditions: 1) a respondent has no friends who use both substances, in which case s/he needs to have at least two friends, who use different substances, 2) a respondent has at least one friend who uses both substances; this is also a minimum condition, as the respondent need not have any more friends using either substance to qualify for this combination category.

If both x_0 = 'Most' and x_1 = 'Most', then:

$$\begin{aligned}
&p(y_{\text{FDCIG}} = \text{'Most'}, y_{\text{FDALC}} = \text{'Most'}|n, \boldsymbol{\theta}) \\
&\quad = \sum_{i=0}^{\lceil n/2 \rceil -1} \sum_{j=\mathcal{I}_0(i)}^{\lceil n/2 \rceil -i-1} \sum_{k=\mathcal{I}_0(i)}^{\lceil n/2 \rceil -i-j-1} p(n_{00} = i, n_{10} = j, n_{01} = k, n_{11} = n-i-j-k) \\
&\quad = \sum_{i=0}^{\lceil n/2 \rceil -1} \sum_{j=\mathcal{I}_0(i)}^{\lceil n/2 \rceil -i-1} \sum_{k=\mathcal{I}_0(i)}^{\lceil n/2 \rceil -i-j-1} \binom{n}{i, j, k, n-i-j-k} \theta_{00}^{i}\, \theta_{10}^{j}\, \theta_{01}^{k}\, \theta_{11}^{(n-i-j-k)}
\end{aligned}$$

where

$$\mathcal{I}_0(i) = \begin{cases} 1 & \text{if } i = 0 \\ 0 & \text{otherwise} \end{cases}$$

The pivotal constraint for the 'Most','Most' category is the number of completely non-using friends, n_{00}. If there is at least one completely non-using friend, $n_{00} > 0$, then the rest of the friends can be all dual substance users. However, if $n_{00} = 0$, then in order to satisfy this category, the respondent needs to have again at least two friends who each use only one of the substances, but different ones.

If either x_0 = 'Few' and x_1 = 'Most', or x_0 = 'Most' and x_1 = 'Few', then:

$$\begin{aligned}
&p(y_{\text{FDCIG}} = x_0, y_{\text{FDALC}} = x_1|n, \boldsymbol{\theta}) \\
&\quad = \sum_{i=0}^{\lfloor n/2 \rfloor} \sum_{j=\lfloor n/2 \rfloor +1-i}^{n-1-i} \sum_{k=\mathcal{I}_0(i)}^{min(\lfloor n/2 \rfloor -i, n-i-j)} p(n_{00} = n-i-j-k, n_{uv} = k, n_{vu} = j, n_{11} = i) \\
&\quad = \sum_{i=0}^{\lfloor n/2 \rfloor} \sum_{j=\lfloor n/2 \rfloor +1-i}^{n-1-i} \sum_{k=\mathcal{I}_0(i)}^{min(\lfloor n/2 \rfloor -i, n-i-j)} \binom{n}{n-i-j-k, k, j, i} \theta_{00}^{i}\, \theta_{uv}^{j}\, \theta_{vu}^{k}\, \theta_{11}^{(n-i-j-k)}
\end{aligned}$$

where

$$(u, v) = \begin{cases} (1, 0) & \text{if } x_0 = \text{'Few' and } x_1 = \text{'Most'} \\ (0, 1) & \text{if } x_0 = \text{'Most' and } x_1 = \text{'Few'} \end{cases}$$

and

$$\mathcal{I}_0(i) = \begin{cases} 1 & \text{if } i = 0 \\ 0 & \text{otherwise} \end{cases}$$

and

$$min(x, y) = \begin{cases} x & \text{if } x < y \\ y & \text{otherwise} \end{cases}$$

The primary constraint is friends who use both substances, n_{11}, which determines the allowable range for the substance 'Most' friends use (j). The 'Few' number of friends who use the other substance (k) is then defined by both i and j, the maximum of which is the remaining number of friends up to $\lfloor n/2 \rfloor$.

Now that we can obtain the probabilities, $\boldsymbol{p}_{\text{FD}}$, associated for all sixteen possible combinations of y_{FDCIG} and y_{FDALC}, we can express the log-likelihood (for four parameters in $\boldsymbol{\theta}$):

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{n}_{\text{FD}}, n) &= \log\left[\text{Multinomial}(\boldsymbol{n}_{\text{FD}}|\boldsymbol{p}_{\text{FD}}, n)\right] \\ &= \log\left[\left(\sum_{i=\text{None}}^{\text{All}} \sum_{j=\text{None}}^{\text{All}} n_{ij}\right) \cdot \left(\prod_{i=\text{None}}^{\text{All}} \prod_{j=\text{None}}^{\text{All}} \frac{1}{n_{ij}!} \cdot p_{ij}^{n_{ij}}\right)\right] \end{aligned}$$

where $p_{ij} = p(y_{\text{FDCIG}} = i, y_{\text{FDALC}} = j|n, \boldsymbol{\theta})$ and $\boldsymbol{n}_{\text{FD}}$ contains counts of all sixteen possible combinations of joint responses to y_{FDCIG} and y_{FDALC} responses.

For practical computation, we will resort to a less complicated procedure: enumerate all possible combinations and then assess both the category to which the combination belongs and the associated probability:

n_{00}	n_{10}	n_{01}	n_{11}	z_{FDCIG}	z_{FDALC}	$q = \text{Multinomial}(\boldsymbol{n}|\boldsymbol{\theta})$
3				None	None	0.1315
2	1			Few	None	0.0705
1	2			Most	None	0.0126
	3			All	None	0.0007
2		1		None	Few	0.0985
1	1	1		Few	Few	0.0352
2			1	Few	Few	0.2123
	2	1		Most	Few	0.0031
1	1		1	Most	Few	0.0759
	2		1	All	Few	0.0068
1		2		None	Most	0.0246
	1	2		Few	Most	0.0044
1		1	1	Few	Most	0.1060
	1	1	1	Most	Most	0.0189
1			2	Most	Most	0.1143
	1		2	All	Most	0.0204
		3		None	All	0.0020
		2	1	Few	All	0.0132
		1	2	Most	All	0.0285
			3	All	All	0.0205

where $\boldsymbol{n} = (n_{00}, n_{10}, n_{01}, n_{11})$ (i.e. a particular row combination), and $\boldsymbol{\theta}$ = (0.575, 0.096, 0.115, 0.214) (i.e. the self-reported proportions of our adolescent population based on past three year smoking (CIGRC3) and and past year alcohol consumption (ALCYR)). To obtain the $\boldsymbol{p}_{\text{FD}}$ probabilities, we collate the q probabilities for each FD pair from the above table:

$$p_{ij} = \sum_{k=1}^{n_{combos}} \mathcal{I}(i, j, k) \cdot q_k$$

where n_{combos} is the number of enumerated combinations (i.e. 24 combinations for $n = 3$), k points to a particular row,

$$\mathcal{I}(i,j,k) = \begin{cases} 1 & \text{if } z_{k,\text{FDCIG}} = i \text{ and } z_{k,\text{FDALC}} = j \\ 0 & \text{otherwise} \end{cases}$$

and $i \in \{\text{None, Few, Most, All}\}$ and $j \in \{\text{None, Few, Most, All}\}$. Now that we can quickly compute the given the likelihood given a fixed n friends, we include the Poisson prior on n:

$$\begin{aligned} p_{ij} &= \sum_{n=0}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \left(\sum_{k=1}^{\binom{n+3}{n}} \mathcal{I}(i,j,k) \cdot q_k \right) \\ \mathcal{L}(\lambda, \boldsymbol{\theta}) &= \log\left[\text{Multinomial}(\boldsymbol{n}_{\text{FD}} | \boldsymbol{p}_{\text{FD}}, n)\right] \end{aligned}$$

where the number of combinations per n, $n_{combos}|n$, is $\binom{n+3}{n}$. For practical purposes, we compute for values of n friends which give noticeable probabilities, $> 1 \times 10^{-6}$; limiting the sum to five times our projected, hypothetical λ (i.e. 5λ) effectively accomplishes this, given that our λ is generally less than 6.

Alternatively, we can compress the notation, by summing each of the sixteen probabilities $\boldsymbol{p}_{FD}$ at index (u, v), where $u, v \in \{\text{'None', 'Few', 'Most', 'All'}\}$; in the tuple notation for $\boldsymbol{p}_{FD}$, the zero padding indicates summing for only the probability associated with index (u, v):

$$\boldsymbol{p}_{\text{FD}}(\lambda, \boldsymbol{\theta}) = \sum_{n=0}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \left(\sum_{i=0}^{n} \sum_{j=0}^{n} \sum_{k=0}^{n} (0, \ldots, 0, p_{uv}(i,j,k,\boldsymbol{\theta}), 0, \ldots, 0) \right)$$

where

$$\begin{aligned} p_{uv}(n,i,j,k,\boldsymbol{\theta}) &= \text{Multinomial}((i,j,k,n-i-j-k)|\boldsymbol{\theta}) \\ &= \binom{n}{i,j,k,n-i-j-k} \cdot \theta_{00}^{i} \cdot \theta_{10}^{j} \cdot \theta_{01}^{k} \cdot \theta_{11}^{(n-i-j-k)} \end{aligned}$$

and

$$u = \begin{cases} \text{'None'} & \text{if } i + k = n \\ \text{'Few'} & \text{if } 1 \le n - i - k \le \lfloor n/2 \rfloor \\ \text{'Most'} & \text{if } \lfloor n/2 \rfloor + 1 \le n - i - k \le n - 1 \\ \text{'All'} & \text{if } i + k = 0 \end{cases}$$

and

$$v = \begin{cases} \text{'None'} & \text{if } i + j = n \\ \text{'Few'} & \text{if } 1 \le n - i - j \le \lfloor n/2 \rfloor \\ \text{'Most'} & \text{if } \lfloor n/2 \rfloor + 1 \le n - i - j \le n - 1 \\ \text{'All'} & \text{if } i + j = 0 \end{cases}$$

As before, the probability of use for each of the sixteen states of friends' use $\boldsymbol{p}_{\mathrm{FD}}$ is employed in a multinomial with the sixteen categories of empirical counts $\boldsymbol{n}_{\mathrm{FD}}$ to assess the log-likelihood $\mathcal{L}(\boldsymbol{n}_{\mathrm{FD}}|\lambda, \boldsymbol{\theta}) = \log[\mathrm{Multinomial}(\boldsymbol{n}_{\mathrm{FD}}|\boldsymbol{p}_{\mathrm{FD}})]$.

4.3 Results from Joint Two Substance Analysis

We derive parameter estimates for each joint pairings; from the three substances, we have three unique, unordered pairs:

x_0	x_1	λ	θ_{00}	θ_{10}	θ_{01}	θ_{11}	μ_0	μ_1	$\mathcal{L}$
CIG	ALC	3.12	0.567	0.092	0.091	0.250	1.07	1.07	-2145.88
RC3	YR	3.17	0.575	0.096	0.115	0.214			-2387.92
CIG	MRJ	3.08	0.627	0.162	0.034	0.177	1.04	0.65	-2503.48
RC3	FLAG	3.21	0.659	0.162	0.031	0.148			-2673.61
ALC	MRJ	2.83	0.620	0.157	0.026	0.197	1.00	0.63	-2267.77
YR	FLAG	3.16	0.641	0.180	0.030	0.149			-2712.06

The first line, of each pair of lines, reports the basic Poisson/binomial parameter fit, with no divisions in the population.[3] The second line provides two pieces of information. In the x_0 and x_1 columns, the line displays the recency of use indicator employed for the fits in both lines of the pair. Next, the estimated parameter in the second line is just λ, because we fix $\boldsymbol{\theta}$ to the population-level "use" proportions, where "use" is indicated by the recency indicator types; that is, we want to know how well the population-level prevalence alone predicts the friends' use ego-network. As expected, there is more consistency in the mean number of friends λ as well as the marginals on the mean number of using friends, indicated under μ_0 and μ_1. Furthermore, we find that the population-level proportions alone are insufficient proxies for $\boldsymbol{\theta}$; this suggests that there exists distinct sub-populations which display vastly different ego-network substance use properties.

For the purposes of verification, we look at sub-populations according to school grade level:

[3] The standard errors surrounding the parameters of the first line in each pair:

x_0	x_1	σ_λ	$\sigma_{\theta_{00}}$	$\sigma_{\theta_{10}}$	$\sigma_{\theta_{01}}$	$\sigma_{\theta_{11}}$	σ_{μ_0}	σ_{μ_1}
CIG	ALC	0.0188	0.0025	0.0014	0.0014	0.0020	0.0075	0.0074
CIG	MRJ	0.0199	0.0025	0.0017	0.0009	0.0017	0.0075	0.0057
ALC	MRJ	0.0186	0.0025	0.0017	0.0008	0.0019	0.0073	0.0056

Grade	λ	θ_{-}	θ_{c}	θ_{a}	θ_{ca}	$\mathcal{L}$	n	μ_{any}	σ_{any}
6	2.76	0.783	0.085	0.031	0.101	-307.21	3879	0.87	0.021
7	3.08	0.681	0.096	0.053	0.170	-343.10	4357	1.51	0.027
8	3.41	0.563	0.101	0.084	0.252	-337.79	4609	2.35	0.036
9	3.61	0.477	0.094	0.125	0.303	-251.14	4232	2.99	0.044

Each user type is represented by the following character tokens: '–' refers to non-use, 'c' refers to smoking only, 'a' refers to alcohol consumption only, and 'ca' refers to use of both. Despite the age dispersion for each grade level, key parameters such as λ and θ_{ca} increase monotonically with grade level. These results are compared to those of Simons-Morton and Chen (2006), who in 1996 followed a cohort of 2453 6th graders into the 9th grade; these students were sampled from seven middle-schools in a suburban Maryland school district, and the time frame overlaps the NSDUH survey years for our data. Respondents were asked 'how many of your five closest friends smoke/drink?'; as in the NSDUH, the definition of 'use' (i.e. smoking or drinking) is left unspecified. The above μ_{any} mirrors the measure of friends' use employed in that study, in which reports the count of friends who smoke added to the count of friends who drink; hence, friends who use both substances are tallied twice. The plot in Figure 4.2 directly verifies the accuracy of our μ_{any} and indirectly verifies the friends count parameter λ and the level of friends' use in the $\boldsymbol{\theta}$'s. Our mean number of substance using friends as a function of education level corresponds very closely to the same measure of their study.

We briefly report one more set of results before moving on to the next section where we jointly estimate parameters for all three substances.[4] The sub-populations in the following tables are defined by each of the four combinations of cigarette and alcohol use, defined by past three year use and past year use, respectively. Each sub-table is denoted by the additional constraints on the likelihood function: none (i.e. meaning we fit only on the joint friends' use data), indicator fitting (i.e. include a fit to the overall population level proportions of CIGRC3 and ALCYR), tie matching (i.e. include a fit for count of ties between each user type to every other user type with the exception of its own kind).[5]

[4] Age-specific estimates for all joint pair combinations can be found in Table F.6

[5] The size of the sample population for the tie matching fit is somewhat arbitrary, yet should represent the size of a connected network. We select a size of $n = 100$, which is a) a realistic size for a network component and b) admits enough tie overlapping between each sub-population without straining the fit on the joint friends' use.

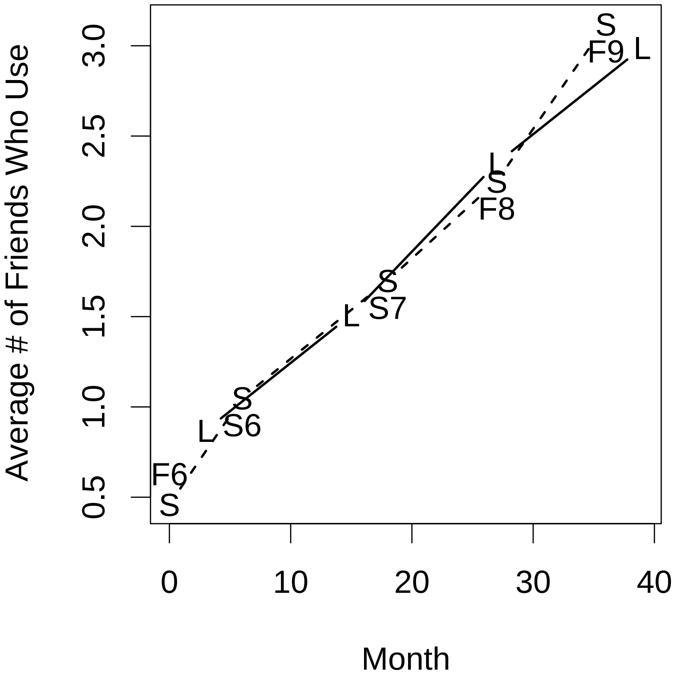


Figure 4.2: Average Friends' Use. *The dashed 'S' line denotes data from Simon-Morton's 1996 study, that looks at a cohort from early fall of the 6th grade (F6 or month 0) to fall of the 9th grade (F9 or month 36), with three additional surveys in between; the line plots the number of friends who smoke and/or drink alcohol. In that study, friends who use both substances are counted twice for this plot. The solid 'L' line denotes equivalent average counts of friends who use from our joint analysis on tobacco and alcohol; a grade year is considered to span from June to June so point estimates are placed at mid-grade months, i.e. November. Due to the large sample sizes for both data sources, confidence intervals are not visible.*

Augmented Likelihoods

	none					indicator only				
use	λ	θ_{-}	θ_{c}	θ_{a}	θ_{ca}	λ	θ_{-}	θ_{c}	θ_{a}	θ_{ca}
-	3.17	75.3	6.8	5.0	12.9	3.12	74.4	7.0	5.5	13.1
c	4.03	51.2	19.0	7.0	22.7	4.02	50.0	19.7	8.1	22.2
a	4.20	47.6	7.5	20.3	24.6	4.19	46.0	8.1	22.1	23.8
ca	4.63	30.0	11.8	14.6	43.7	4.63	28.8	12.5	16.1	42.6
	$\mathcal{L}_d = -1319$					$\mathcal{L}_{di} = -1392$				

	ties only					ties and indicator				
use	λ	θ_{-}	θ_{c}	θ_{a}	θ_{ca}	λ	θ_{-}	θ_{c}	θ_{a}	θ_{ca}
-	3.17	75.2	6.8	5.0	13.0	3.12	74.3	7.1	5.6	13.1
c	4.03	51.2	18.5	7.3	23.0	4.01	49.9	19.1	8.5	22.5
a	4.01	47.6	7.3	19.7	25.4	4.00	46.1	7.9	21.4	24.6
ca	4.44	30.0	11.8	14.7	43.6	4.45	28.7	12.6	16.2	42.4
	$\mathcal{L}_{dt} = -1861$					$\mathcal{L}_{dti} = -1935$				

Parameter estimates for each row are percentages, for the sake of readability. The four likelihoods produce similar θ's, differing at most by a few percentage points. In all of these, we find an interesting pattern of affiliation that implies separate, yet overlapping, substance use cultures. For instance, respondents who only smoke ("smokers") are more likely to report having friends who also only smoke more so than any of the other type of substance user; the "ties and indicator" table entry at (c,θ_{c}) is 19.1%, highest in its column. Similarly, respondents who only drink alcohol ("drinkers") are also more likely to have only alcohol-drink friends than any others; (a,θ_{a}) = 21.4%, also the highest in its column. Both of these demonstrate a possible homophilic selection mechanism along specific substances, and not just general substance use. While some of this symmetry is observable directly from the FDCIG and FDALC data, it is more prominent in these findings.

However, the differences in mean friends λ for the non-tie matching likelihoods complicate our earlier assertion that users of a less prevalent substance would have more friends. We would expect smokers to have more friends than drinkers, since cigarette use is less prevalent than alcohol consumption: $p_{\mathrm{CIGRC3}} < p_{\mathrm{ALCYR}}$. However, this is not the case. What might account for this is the difference in their levels of affiliation to other substance users and consequently to non-users. Smokers show higher rates of affiliation to the majority population of non-users more than drinkers, while conversely drinkers show a higher affiliation to other drinkers, as well as drinker/smokers more than smokers to smokers. This inconsistency suggests an additional mechanism such as alcohol consumption being associated with a higher level of social activity, which happens to be the case at least anecdotally: social gatherings of youth and adults alike often include alcohol consumption.

4.4 Joint Analysis of Three Substances

We now turn to joint analysis of all three substances. Here, we report the tabulated responses of friends' use (again as percentages of the total number of responses):

	y_{FDMJ} = None				y_{FDMJ} = Few			
	y_{FDALC} =				y_{FDALC} =			
y_{FDCIG}	None	Few	Most	All	None	Few	Most	All
None	31.92	4.03	0.39	0.09	0.39	1.05	0.21	0.08
Few	7.68	10.79	1.50	0.21	1.14	14.45	3.76	0.34
Most	0.70	0.76	0.65	0.11	0.13	2.85	3.01	0.70
All	0.08	0.10	0.07	0.08	0.03	0.30	0.29	0.16

	y_{FDMJ} = Most				y_{FDMJ} = All			
	y_{FDALC} =				y_{FDALC} =			
y_{FDCIG}	None	Few	Most	All	None	Few	Most	All
None	0.01	0.08	0.15	0.04	0.02	0.00	0.01	0.06
Few	0.12	0.56	1.26	0.26	0.02	0.03	0.06	0.23
Most	0.08	0.68	4.71	0.45	0.01	0.07	0.19	0.47
All	0.04	0.10	0.35	0.36	0.02	0.03	0.17	1.32

The mass of responses for each quadrant gathers at the diagonal entries (e.g. y_{FDCIG} = 'Few', y_{FDALC} = 'Few', y_{FDMJ} = 'Few') again pointing to co-substance using friends. If we look at the proportions of use for each combination of past three year smoking (CIGRC3), past year drinking (ALCYR), and lifetime use of marijuana (MRJFLAG), we find some of the co-substance use pattern to hold:

		Never Tried Marijuana		Have Tried Marijuana	
		Drank Alcohol in Past Year?			
		No	Yes	No	Yes
Smoked in the Past	No	56.4	9.5	1.1	2.0
Three Years?	Yes	7.7	8.5	1.9	12.9

Marijuana use itself is strongly associated with co-substance use. On the left the use percentages are roughly similar (7.7%-9.5%), while on the right, when a respondent has tried marijuana at least once in his or her lifetime, there is higher concomitant recent use of both cigarettes and alcohol (12.9%). Also we note the use of just marijuana is a rarity (1.1%) as is the use of just alcohol and marijuana (2.0%).

The eight categories of co-substance use for three substances correspond to now eight categories for $\boldsymbol{\theta}$:

$$\theta_{ijk} \in \{\theta_{000}, \theta_{100}, \theta_{010}, \theta_{110}, \theta_{001}, \theta_{101}, \theta_{011}, \theta_{111}\}$$

	# of Substances		
n	One	Two	Three
0	1	1	1
1	2	4	8
2	3	10	36
3	4	20	120
4	5	35	330
5	6	56	792
6	7	84	1,716
7	8	120	3,432
8	9	165	6,435
9	10	220	11,440
10	11	286	19,448
11	12	364	31,824
12	13	455	50,388
13	14	560	77,520
14	15	680	116,280
15	16	816	170,544
16	17	969	245,157
17	18	1,140	346,104
18	19	1,330	480,700
19	20	1,540	657,800
20	21	1,771	888,030
Σ	231	10,626	3,108,105

Table 4.2: Combination Space of Types of Substance Using Friends. *Each entry contains the number of ways a set of n friends can fall into 2, 4, or 8 (i.e. 2^1, 2^2, 2^3) substance using categories. The number of combinations per n equals $\binom{n+m-1}{n}$ where m is the number of substance using states; if d is the number of substances then the number of substance usingstates is $m = 2^d$.*

where $i = 1$ indicates past three years smoking, $j = 1$ indicates past year drinking, and $k = 1$ indicates ever having used marijuana. So, for example, θ_{010} refers to the probability that a friend only drinks alcohol.

Table 4.2 gives us a sense of how much more difficult it is to analyze three substances jointly. The total number of combinations, required for each probability estimate, expands exponentially per added substance. As expressed earlier, the number of combinations for two substances given an assumed number of friends n is $\binom{n+3}{n}$. This can be generalized to $\binom{n+m-1}{n}$, or equivalently $\binom{n+m-1}{m-1}$, where m is the number possible substance use states, 4 for two substances and 8 for three substances; this is the general expression for the number of ways, or combinations, n items can fall into m categories or bins. To make matters more difficult, the Newton-Raphson requires several hundred estimations per iteration for estimating parameters for three substances jointly. The number of parameters to estimate grows from 1 (for λ) + 1 (for 2 θ states - 1) = 2 in the 1-drug model, to $1+(4-1) = 5$ in the 2-drug model, and $1 + (8 - 1) = 8$ in the 3-drug model, and consequently the number of partial second derivatives to compute increases quadratically from $2^2 = 4$ to $4^2 = 16$ to $8^2 = 64$. Estimating a single set of parameters takes a few seconds to a few minutes for one and two substances. The estimation of parameters covering three substances requires about 2 hours of computation time on the fastest (3+GHz) servers at our disposal! For indicator-inclusive or tie-matched likelihoods, we need 16 parameters to cover all eight sub-populations; this means 16^2 or 256 partial second derivatives are computed per estimation step. Due to limitations in available computing power, we have not performed these latter two estimations as of this time.

The joint population-level estimates for all three substances are:

Friends' Use

λ	θ_{-}	θ_{c}	θ_{a}	θ_{ca}	θ_{m}	θ_{cm}	θ_{am}	θ_{cam}	θ_{c}	θ_{a}	θ_{m}	$\mathcal{L}$
3.03	55.1	7.9	6.7	8.7	0.8	1.7	2.8	16.2	34.5	34.5	21.5	-3627
2.94	50.5	7.9	11.3	12.3	0.3	1.1	1.4	15.1				-5141

Adults' Use

λ	θ_{-}	θ_{c}	θ_{a}	θ_{ca}	θ_{m}	θ_{cm}	θ_{am}	θ_{cam}	θ_{c}	θ_{a}	θ_{m}	$\mathcal{L}$
4.22	36.4	11.9	19.0	23.7	0.5	0.7	0.8	6.9	43.2	50.5	9.9	-3737

The θ's are reported as percentages. As with the joint two substance analysis, we subscript the θ's with character tokens indicative of a substance combination: '$-$' = No Substance Use, 'c' = smoking only, 'a' = drinking only, 'm' = marijuana use only, 'ca' = smoking and drinking only, etc. The first line reports estimates of both λ and θ's while the second line reports a fit on λ using the population *lifetime* use rates (across all substances) as our θ's; as expected, we achieve a superior fit when we employ our assumed definitions of "use". The ego-network findings again mirror the bimodal pattern of self-use: teens tend to either use all three or abstain from all three. Outside of those two categories, we see how certain substances seem more likely to be used in tandem with others: marijuana in particular is far less likely to be used on its own, according to both respondent data as well as the inferred parameter θ_{m}. These patterns do not hold quite so closely for relationships with substance using *adults*; there appears to be far less similar use and far less affiliation with marijuana using adults.[6]

Next, we infer parameter estimates per age group and obtain:

y_{AGE}	λ	θ_{-}	θ_{c}	θ_{a}	θ_{ca}	θ_{m}	θ_{cm}	θ_{am}	θ_{cam}	$\mathcal{L}$	n
12	2.54	81.1	7.5	2.0	3.8	0.3	0.9	0.5	3.9	-521	4072
13	3.00	71.9	7.5	3.4	7.4	0.6	1.2	0.8	7.2	-542	4329
14	3.26	59.0	8.5	6.4	8.2	0.8	1.7	2.2	13.2	-665	4427
15	3.44	49.8	6.9	8.2	9.6	1.2	1.6	3.0	19.7	-620	4400
16	3.66	42.1	9.3	10.1	9.3	1.6	1.9	4.8	20.9	-665	4133
17	3.84	38.7	7.6	10.0	10.7	0.9	2.9	5.0	24.2	-561	4102
$\lvert\Delta\theta\rvert$		42.4	1.8	8.1	6.9	1.3	2.0	4.5	20.3	-3573	

The values in the $|\Delta\theta|$ row, with the exception of $\Sigma\mathcal{L}$ report the differences between the maximum and minimum percentages of each column, reflecting the change of representation in the type of substance using friends. While we expect the proportion of friends who use all three substances (cam) to be larger, in general, than other types of using friends, the fact that it dwarfs other categories is surprising — in how quickly it accumulates over time as well as its magnitude at age 17, given that the proportion of respondents who used all three substances ('cam'), 12.9% (shown earlier in this section), is not that much greater than the proportion of respondents expressing other combinations of use, particularly 'c', 'a' and 'cm', the range of which is 7.7%-9.5%. Friendship ties apparently gravitate towards those who are 'cam' poly-users; hence, we should expect 'cam' users to have the highest count of friends, and this is confirmed in the next table. However, initiation patterns do not mirror an increase in concurrent poly-substance initiation,[7] suggesting this migration describes current users' behavior.

We also notice decreasing rates of single substance users, 'c', 'a', and 'm' between the ages of 16 and 17. While the migration towards poly-use would likely account for some of the decrease, it is also likely that with cigarettes 'c', prior smokers have ceased as we pointed out earlier in footnote [1] of this chapter, and returned to the non-use '–' category; θ_{-} drops only 3.4% between 16 and 17 while the drops at every other

[6] The respondents' data for joint adults' use can be found in Table F.5

[7] The percentages of respondent initiates for all three substances per age groups is:

Age	12	13	14	15	16	17
% init.	0.51	0.53	0.81	0.39	0.15	0.07

While this measure does not capture the total proportion of initiates per age, due to the broad time interval in which this data was collected, this curvilinear pattern in which poly-use initiation peaks at age 14 and drops off quickly, is representative of the pattern for the true initiation rates.

age step is appreciably more. Finally, we see a more even progression of increasing λ, roughly an increase of 0.2 friends per year, which seems more reasonable than a progression with uneven jumps, given our large, national sample.

We estimate parameters for each type of respondent:

	# of										
self	Friends	Friends' Use									
use	λ	θ_-	θ_{c}	θ_{a}	θ_{ca}	θ_{m}	θ_{cm}	θ_{am}	θ_{cam}	$\mathcal{L}$	n
–	3.11	75.1	6.0	3.9	5.9	0.5	0.9	1.0	6.7	-1250	14357
c	3.87	52.5	17.0	5.1	9.2	0.4	2.5	1.6	11.7	-342	1960
a	4.16	49.5	6.6	16.7	10.5	0.6	1.1	2.9	12.1	-271	2426
ca	4.35	36.4	10.5	13.4	17.5	0.9	1.0	2.4	17.8	-326	2167
m	4.02	41.2	5.5	4.5	5.3	6.2	4.9	9.6	22.8	-102	277
cm	4.36	39.8	13.4	3.8	4.2	4.0	4.6	4.8	25.3	-140	480
am	4.32	31.3	4.7	9.3	7.6	3.8	2.3	16.3	24.7	-141	505
cam	4.81	22.8	8.3	6.9	7.9	1.9	4.2	7.6	40.4	-370	3291
$\Sigma\mathcal{L}$										-2942	

Consistent with earlier analysis on cigarettes, the use-based decomposition outperforms the age-based decomposition once again suggesting friends' use is more associated with self-use than age. Use-based homophily continues to be present, though the evidence is not as stark. When we review the parameter estimates for respondents of use categories 'c', 'a', 'ca', 'cm', and 'am', excluding non-use '–' and poly-use 'cam', we find the third highest friends' use category is the same as the respondents' of that same category, with the exceptions for marijuana users. Marijuana use of course is associated with poly-use of all three; θ_{cam} for all self-use categories that include 'm' dominates as the primary friends' use category. However, for non-'m' use categories, 'c', 'a', and 'ca', the specific combination of self-use is reflected in the pool of friends, which again suggests specialized substance use homophily. Also, due to the paucity of certain categories of users, namely 'm' and 'cm', respondents claiming these combinations of use likely have difficulty finding identically using friends; hence, we see greater affiliation with overlapping categories, instead. However, some caution is warranted in making these claims. Since our data is self-reported, there remains the possibility that respondents were biased in believing their friends used similar substances to themselves; still, the break in the pattern with marijuana users suggests that this not entirely the case.

Now, we obtain poly-substance parameter estimates on also adults' substance use:

	# of									
self	Adults	Adults' Use								
use	λ_a	θ_{-}	θ_{c}	θ_{a}	θ_{ca}	θ_{m}	θ_{cm}	θ_{am}	θ_{cam}	$\mathcal{L}$
–	4.16	44.1	11.2	17.4	22.7	0.3	0.5	0.5	3.3	-2177
c	4.34	34.0	17.7	17.1	24.0	0.8	0.5	0.1	5.8	-355
a	4.55	29.5	9.5	26.2	26.6	0.4	0.6	0.7	6.4	-373
ca	4.93	26.3	12.3	23.8	28.3	0.7	0.8	0.6	7.3	-371
m	4.43	33.9	7.7	12.9	21.8	1.4	2.1	3.5	16.6	-115
cm	4.47	30.8	14.1	14.0	20.2	0.7	1.6	2.3	16.4	-168
am	4.47	22.1	13.0	22.0	19.3	3.0	0.8	2.3	17.5	-182
cam	5.01	22.6	12.7	18.8	20.5	1.3	2.1	3.1	18.8	-404
$\Sigma\mathcal{L}$										-4146

Firstly, we notice that the λ estimate for acquaintanceship with adults is generally higher than for adolescents' friends. This is not surprising given that 'knowing' expresses a less restrictive affiliation than friendship. However, the lack of specificity in 'knowing' makes it difficult to ascertain the specific types of adults considered by the respondents: parents, teachers, friends of parents, older siblings, etc. Still, we see, as we did with friends' count, the level of association to adults increases monotonically with age.[8] The pattern of association between adolescents and adults is interestingly different than that between adolescents and their friends. There is less substance use homophily and more dominance of affiliation to alcohol consuming and both alcohol consuming and smoking adults, which should not be surprising, since both of these substances are not stigmatic in the adult world and have a prevalence and visibility far more than marijuana. Furthermore, adolescents generally do not choose the adults with whom they associate; hence, we would see far less homophilic tendencies.

[8]

λ's Estimated from Friends' and Adults' Use of

	Tobacco		Alcohol		Marijuana	
Age	λ	λ_a	λ	λ_a	λ	λ_a
12	2.76	4.68	2.43	3.60	1.94	2.33
13	3.25	4.86	2.89	3.81	2.60	2.49
14	3.57	5.10	3.14	4.15	2.69	2.77
15	3.67	5.09	3.38	4.18	2.94	3.03
16	3.98	5.20	3.61	4.23	3.05	3.36
17	4.10	5.19	3.88	4.46	3.25	3.25

Chapter 5

Prelude to a Dynamic Model

Dynamic modeling represents both the Holy Grail and bane of network analysis. In theory, the approach attempts to account for subtle dependencies and feedback effects not captured by static or cross-sectional inference. However, difficulties in gathering temporal network data and the lack of adequate methodologies have until recently muted previous attempts to explain the interplay between adolescents' peer networks and their substance use. The analyses in this chapter attempt to bridge the two perspectives by employing the cross-sectional data of the NSDUH to inform the development of a dynamic model, by isolating age transition dynamics. Specifically, we focus on how friendship networks grow in the number of substance using friends through influence (in the form of friends' initiating) and selection (by distinguishing between the degree to which users and non-users select using friends). For these analyses, we employ on 'ever having used' a substance, or lifetime use, as the definition of 'use'. Despite the implications of this assumption — that is, we cannot model users who revert to non-use status — we find striking differences between how users and non-users networks evolve from age 12 to 17.

5.1 Risk of Initiation

In order to understand change in substance use dynamics, we focus on the initiation stage. Initiates are respondents whose current age (at the time they participated in the survey) equals that of their age of first use (i.e. CIGTRY, ALCTRY, or MJAGE). The friends' use data for initiates of each substance appears as follows:

	raw				prop.			
Friends' Use	None	Few	Most	All	None	Few	Most	All
Tobacco	87	394	149	21	0.134	0.605	0.229	0.032
Alcohol	228	633	342	79	0.178	0.494	0.267	0.062
Marijuana	63	344	151	35	0.106	0.580	0.255	0.059

Again, these data can be deceptive. For instance, we might be inclined to think that that first time smokers have more smoking friends than do first time alcohol drinkers have drinking friends due to a higher proportion of first time alcohol expressing 'None' of their friends use. However, λ and θ parameter estimation gives us a more accurate way to describe prevalence rates:

Parameter Estimates for Substance Use Initiation

Substance	λ_i	θ_i	σ_{λ_i}	σ_{θ_i}	$\mathcal{L}$	n_i	μ_i	σ_{μ_i}
Tobacco	5.16	0.406	0.205	0.011	-10.18	651	2.10	0.090
Alcohol	4.21	0.446	0.113	0.009	-20.79	1282	1.88	0.056
Marijuana	4.97	0.445	0.202	0.012	-8.95	593	2.21	0.096

The estimates tell us that, while the average initiating smoker indeed has more smoking friends than first time drinkers have drinking friends, in fact, the proportions are reverse. And, when the parameters are considered jointly, we confirm that the prevalence of raw number of using friends is least for first alcohol consumers; the differences across all three substances are significant.[1] For notation, we use lower-case i to indicate parameters inferred directly from first time user data. We will later use I to denote to denote subsequent inferences from the primary initiation estimates. If we examine parameter estimates for first time smokers in each age group, we obtain:[2]

Parameter Estimates for Tobacco Use Initiation

Age	λ_i	θ_i	σ_{λ_i}	σ_{θ_i}	$\mathcal{L}$	n_i	μ_i	σ_{μ_i}
12	4.23	0.416	0.477	0.036	-6.12	75	1.76	0.220
13	5.23	0.376	0.488	0.025	-8.56	123	1.96	0.197
14	6.21	0.420	0.590	0.023	-6.40	129	2.60	0.264
15	5.24	0.407	0.479	0.025	-7.38	123	2.13	0.215
16	5.02	0.467	0.467	0.027	-8.05	111	2.34	0.231
17	5.24	0.336	0.588	0.030	-6.12	89	1.75	0.215

There is considerable overlap in both parameters across the age groups. T-test significance, at the $p < 0.05$ level, in the difference between the mean number of smoking friends at initiation, μ_i, occurs between age pairs (12,14), (12,13), (14,15), (12,16), (12,13), (14,17), and (16,17); that is, there is no noticeable pattern. In general, smoking initiation occurs in the presence of between 1.75 and 2.60 smoking friends (minimum and maximum μ_i). With alcohol, the range is 1.44-2.41 and, with marijuana, it is 1.51-3.09. These ranges are consistent with the pooled population estimates of mean using friends: the ordering, from least to most, goes alcohol, tobacco, and marijuana. Right now, these findings suggest that the circumstances of initiation are similar across all age groups. However, this claim is conditional on a

[1] Relevant t-statistics: t_{ca} = -57.02, t_{cm} = -20.79, and t_{am} = -62.38. Joint initiation analysis will not be conducted due to the sparseness of the data.

[2] Similar results for first time alcohol and marijuana users are Tables G.1 and G.2.

respondent already having initiated. If we wanted to know the likelihood of initiation for any given youth, we would need to consider the population of those who did not initiate.

We can use these initiation parameters to ascertain what the risk of initiation is for a youth having a particular quantity of substance using friends. If over a given year there are n_i initiates and n_0 non-users, then given m smoking friends, the probability of initiating p_I is:

$$p_I(m) = \frac{n_i|m}{n_i|m + n_0|m}$$

Altering the equation to express each sub-population as proportions of the entire non-using population prior to the initiation of those n_i adolescents:

$$\begin{aligned} p_I(m) &\approx p(y^{t+1}_{\text{CIGFLAG}} = 1 | y^{t}_{\text{CIGFLAG}} = 0, m) \\ &= \frac{p(m|\lambda_i, \theta_i) \cdot p_i}{p(m|\lambda_i, \theta_i) \cdot p_i + p(m|\lambda_0, \theta_0) \cdot (1 - p_i)} \end{aligned}$$

where p_i is the proportion of the non-using population (and, later, age-specific non-using populations) that initiated, λ_i and θ_i are estimated from the initiating sub-population, and λ_0 and θ_0 are estimated from the non-using sub-population. While, in essence, we are trying to measure the probability of a transition from never having used a substance to trying it with the lifetime indicator, the first line expresses an approximation because we treat the data as simultaneously measured at interval sampling points, when in fact, the dynamics of initiation are much more fluid; not everyone initiates exactly on the same day of the year. We attempt to account for this sampling issue by using changes to the FLAG indicator between age categories as an estimate for the proportion of the population that initiates; the proportion of the population that indicated initiation in the survey cannot be used because the number of initiates per age is severely undersampled.[3] We assume that distribution of using friends for the pre-initiation population can be described by weighting and summing

[3] Given a respondent who initiates at a certain age, if we assume a uniform distribution for the time between a respondent's birth date and the NSDUH interview date and a uniform distribution for the date of initiation, the "first try cigarettes" response item will only capture 25% of the respondents; for instance, roughly half of those who initiate between 12 and 13 will do so while they are still 12-year-olds and the NSDUH will pick up only half of those initiates while who are still 12 and miss the half of that group because they have not initiated at the time of the interview, but will do so before their 13th birthday. We can attempt to correct the NSDUH estimate; however, we cannot be assured that either distribution is uniform. Still, the data compares closely enough to obviate some, not all, concern:

	12	13	14	15	16	17
$p_i^t/0.25$	0.075	0.116	0.118	0.114	0.108	0.090
$p^{t+1}(y_{\text{CIGFLAG}} = 1)$ $- p^t(y_{\text{CIGFLAG}} = 1)$	0.092	0.109	0.118	0.071	0.071	

the coinciding distributions for the non-using population and the initiation population. For each category m substance using friends, we can compute the probability of initiation (e.g. m for tobacco is equivalent n_{smoke}):

Risk of Substance Use Initiation

	Number of Substance Using Friends, m						
	0	1	2	3	4	5	6
Tobacco	0.018	0.054	0.148	0.345	0.615	0.829	0.936
Alcohol	0.032	0.103	0.283	0.576	0.824	0.941	0.982
Marijuana	0.008	0.041	0.193	0.573	0.883	0.977	0.996

As expected, the risk of initiation increases per using friend; it would be extremely surprising if this was not the case! This generally corroborates other findings on impact of peer influence on substance use initiation; it is less likely that within the allotted time frame (between the last birthday and interview date) that a substance-inclined adolescent sought new friends in such a manner as to produce the distributions observed above. The pattern is also interesting: a sudden increase in risk when the number of using friends m reaches 2 or 3. These numbers identify which adolescents would be most affected by changes in the network structure that might alter the number of using friends and also suggest the conditions under which initiation might spread more quickly: if all non-using adolescents had only one using friend, the speed of initiation would differ than if just few of them had an abundant number using friends; this assertion warrants future investigation. For alcohol and marijuana, the leap in risk is more pronounced when the number of using friend increases from one to two; this suggests that these two substances might have different social properties that make initiation more difficult to resist. These patterns are readily apparent in Figure 5.1.

While earlier we saw evidence that pointed to similar ego-networks for all initiates across all age-groups, here we find that in fact age protects against influence and shifts the risk curve to the right:[4]

Risk of Tobacco Use Initiation

	Number of Smoking Friends, $m = n_{smoke}$						
Age	0	1	2	3	4	5	6
12	0.026	0.124	0.434	0.806	0.957	0.992	0.998
13	0.022	0.079	0.248	0.558	0.829	0.949	0.986
14	0.013	0.043	0.136	0.356	0.660	0.872	0.960
15	0.020	0.045	0.096	0.194	0.352	0.551	0.735
16	0.017	0.036	0.074	0.147	0.270	0.443	0.631

A 16-year-old with five smoking friends incurs the same risk of initiation as a 12-year-old with only two smoking friends. This pattern holds for alcohol and marijuana.[5]

[4] This finding is corroborated by initiation analysis in Appendix A and host of outside research: initiation into substance use drops after a peak age of roughly 14 or 15.

[5] Similar results for alcohol and marijuana are reported in Tables G.3 and G.4.

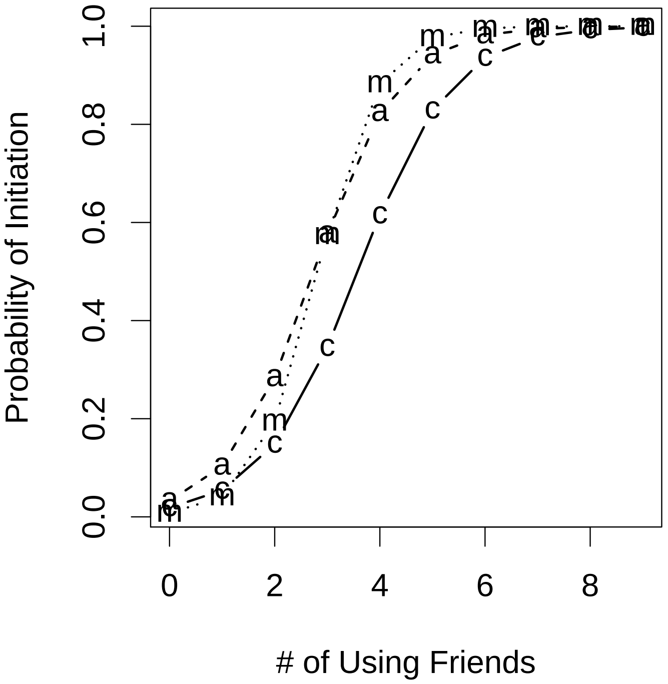


Figure 5.1: Risk of Initiation. *Solid line 'c' denotes probability of trying smoking as a function of smoking friends; dashed line 'a' denotes risk of alcohol initiation; and dotted line 'm', risk of marijuana initiation.*

We can even look at the joint distribution and examine the risk per pair of number of total friends n and number of those friends who use m. In Figure 5.2, we display the joint posterior distribution for 12-year-old initiates.[6] We can see in both the graph and its source data (relegated to the Appendix) that the influence effects become diluted when the number of total friends increases holding for a fixed number of using friends (i.e. number of non-using friends increases).

5.2 Change in Network

5.2.1 Risk of Initiation II

We can now extend our earlier computation of risk of initiation in order to infer how ego-network composition changes from age to age for each group of non-users and users; the following equations apply to each group separately so we do not identify them by a subscript, in order to maintain readability:

$$\hat{p_i^t} = p_i^t \cdot s^t$$

[6] The data for this figure appears in Table G.5.

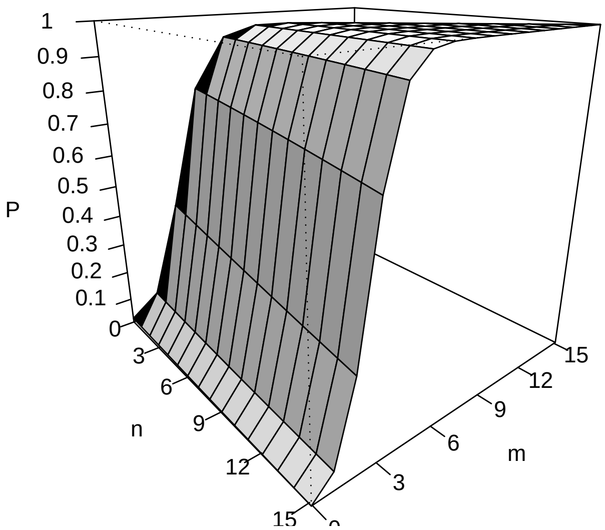


Figure 5.2: Joint Cigarette Initiation

$$
\begin{aligned}
p_I^t(n,m) &= \frac{p(n,m|\lambda_i^t,\theta_i^t)\cdot\hat{p}_i^t}{p(n,m|\lambda_i^t,\theta_i^t)\cdot\hat{p}_i^t+p(n,m|\lambda_0^t,\theta_0^t)\cdot(1-\hat{p}_i^t)}\\
q_{I0}^t &= \sum_{n,m}\left(p_I^t(n,m)\cdot\frac{p(n,m|n\neq m,\lambda_0^t,\theta_0^t)}{p(n\neq m|\theta_0^t,\lambda_0^t)}\right)\\
q_{I1}^t &= \sum_{n,m}\left(p_I^t(n,m)\cdot\frac{p(n,m|n>0,m>0,\lambda_1^t,\theta_1^t)}{p(n>0,m>0|\lambda_1^t,\theta_1^t)}\right)
\end{aligned}
$$

We continue to use the joint distribution for risk of initiation $p_I^t(n,m)$ for each age group, denoted t, while q_{I0} and q_{I1} refer the scalar probabilities of initiation for a friend of a non-smoker and a friend of a smoker, respectively. A non-user will have a joint n, m friend distribution as defined by $p(n,m|\lambda_0,\theta_0)$. However, any friend of this non-smoker must have at least one friend who does not smoke; hence, the diagonal $n=m$ of the joint distribution is ignored giving us: $p(n,m|n\neq m,\lambda_0,\theta_0)$. The friend of a smoker will have at least one friend who does smoke, hence, the first row and column of the joint distribution is ignored: $n>0$ and $m>0$. These edited distributions are appropriately normalized; for instance $p(n>0,m>0|\lambda_1^t,\theta_1^t)$ is the sum probability for the subsection of the joint probability and equals $\sum_{n,m}p(n,m|n>$

$0, m > 0, \lambda_1^t, \lambda_1^t)$.

Furthermore, we find that the proportion of initiation within each age groups requires modification when used as weights for joint distributions on n and m. This is not surprising given that the parameters for initiates were estimated without constraint and not concurrently fitted to another measure (as we did for earlier models when we fitted to recency indicators and tie volume). Furthermore, given the smaller sample sizes for the initiates' sub-population, it is not surprising that the estimates will be moderately biased; in this case, they over-estimate the prevalence of using friends. We correct for this by estimating a scale factor, s^t, for the age-specific initiation rate p_i^t that allows q_{I0} and q_{I1} to sum the original p_i^t when weighted by the respective proportions of use in its sub-population. We display estimates for all three substances in Table 5.1. σ_s and $\mathcal{L}_s$ refer to how well the scaling factor s^t applied to p_i^t results in a fit to the proportion of the population which initiated:

$$\mathcal{L}_s(s^t|\lambda_i^t, \theta_i^t, \lambda_0^t, \theta_0^t, \lambda_1^t, \theta_1^t, (n_0^t, n_1^t)) = \text{Multinom}(\lfloor(1-p_i^t, p_i^t)\cdot n^t\rfloor | \frac{(n_0^t, n_1^t)}{n^t} \cdot (q_{I0}^t, q_{I1}^t), s^t)$$

where $n^t = n_0^t + n_1^t$. As we expected, the risk of initiation for friends of non-users is far less than that of friends of users. The overall initiation risks reflect the prevalence for each substance; for instance, friends of non-drinkers still have a higher initiation rate than say friends of smokers or friends of marijuana users.

5.2.2 Transition Parameters

We can now try to estimate some selection parameters, specifically, the degree to which non-users and users will retain or drop friends who use and when acquiring new friends, the degree to which they select for those who use. We propose the following set of change equations that describe what occurs between each advancing year, t, where t is age of cohort its range is [12,17]. Again, this set of computations can apply to each of the non-using and using sub-populations or both; hence we omit the subscript denoting user or non-user sub-population:

$$\begin{array}{rcll}
\theta_i & = & q_I & \text{(prob. of any friend initiating)} \\
\mu_m & = & \lambda^t \theta^t & \text{(\# of friends who are users)} \\
\mu_i & = & \lambda^t(1-\theta^t)\theta_i & \text{(\# of non-using friends who initiate)} \\
\mu_r & = & \phi(\mu_m + \mu_i) & \text{(\# of using friends who are removed)} \\
\mu_\lambda & = & \lambda^{t+1} - (\lambda^t - \mu_r) & \text{(\# of new friends)} \\
 & \text{or,} & \text{equivalently,} & \\
\lambda^{t+1} - \lambda^t & = & \mu_\lambda - \mu_r & \text{(same as last eq., change in friends)} \\
\mu_s & = & \omega \mu_\lambda & \text{(\# of new friends who use)} \\
\lambda^{t+1}\theta^{t+1} & \approx & \mu_m + \mu_i - \mu_r + \mu_s & \text{(\# of using friends at } t+1\text{)}
\end{array}$$

Tobacco Initiation Parameters

Age	q_{I0}	q_{I1}	p_i	s	σ_s	$\mathcal{L}_s$
12	0.077	0.201	0.092	0.78	0.00163	-3.82
13	0.092	0.175	0.109	0.69	0.00099	-3.93
14	0.099	0.159	0.118	0.60	0.00066	-3.98
15	0.061	0.085	0.071	0.50	0.00078	-3.75
16	0.060	0.081	0.071	0.45	0.00067	-3.71
17	0.065	0.075	0.071	0.47	0.00074	-3.71
μ_I	0.075	0.129	0.087			

Alcohol Initiation Parameters

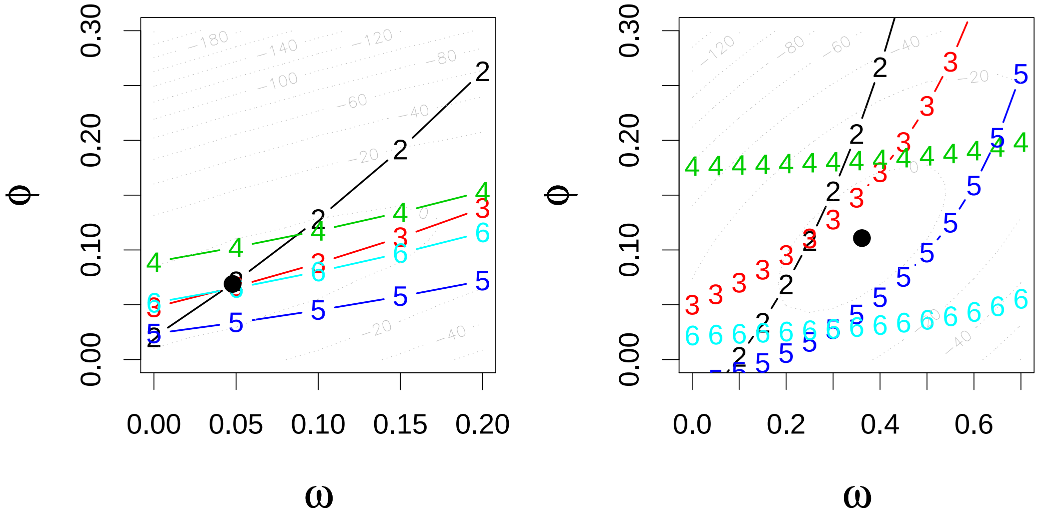
Age	q_{I0}	q_{I1}	p_i	s	σ_s	$\mathcal{L}_s$
12	0.098	0.246	0.115	0.81	0.00138	-3.92
13	0.104	0.196	0.124	0.70	0.00087	-3.99
14	0.120	0.190	0.142	0.60	0.00053	-4.06
15	0.085	0.117	0.099	0.48	0.00049	-3.90
16	0.057	0.078	0.068	0.38	0.00051	-3.69
17	0.062	0.071	0.068	0.40	0.00056	-3.69
μ_I	0.083	0.165	0.103			

Marijuana Initiation Parameters

Age	q_{I0}	q_{I1}	p_i	s	σ_s	$\mathcal{L}_s$
12	0.025	0.070	0.030	0.82	0.00562	-3.30
13	0.060	0.118	0.072	0.80	0.00205	-3.74
14	0.083	0.134	0.099	0.75	0.00125	-3.90
15	0.063	0.088	0.074	0.69	0.00140	-3.77
16	0.055	0.075	0.065	0.63	0.00143	-3.68
17	0.060	0.069	0.065	0.67	0.00160	-3.67
μ_I	0.057	0.144	0.068			

Table 5.1: Risk of Initiation for Friends of Users and Non-Users. *For each age group, the probability of initiation for friends of a non-user q_{I0} and friends of users q_{I1} are reported along with the proportion of initiation p_i and the required scaling factor s.*

Our unknown key parameters are the selection parameters: ϕ for the proportion of using friends that are dropped and ω for the probability that a new friend will be a user. The knowns are the ego-network composition parameters for the age groups for which we want to describe transition, λ^t, θ^t, λ^{t+1}, θ^{t+1}, and θ_i, the probability that a current friend will initiate, which is equivalent to the q_I^t values from Table 5.1. In the last equation, we know how many friends we have at $t+1$, so if we assume to know the rate of using friends we selected out via ϕ, then ω is a known. Hence, there really is only one unknown in these equations (for each age group). Note, it would be more realistic to allow dropping of both using and non-using friends; however, the data does not allow us to include this additional degree of freedom.[7] This restricts interpretation on the results, but we can still make comparative assessments about the degrees of selection among using and non-using respondents. In fact, a simpler version of the model would omit the dropping of any friends, except for the few instances when λ drops between ages. However, our risk of initiation parameters often result in too many friends who initiate between ages; they exceed what we would expect from the parameters of the subsequent age. We could simply categorize these overshoots as mis-fits and allow the parameter estimation to reflect this. Or, we introduce a dropping parameter and achieve a modestly better fit.



(a) non-user, $y_{\text{CIGFLAG}} = 0$

(b) user, $y_{\text{CIGFLAG}} = 1$

Figure 5.3: Age Specific Curves for Selecting Cigarette Smoking Friends. *The ϕ expresses the proportion of using friends that are dropped and ω expresses the probability that a new friend is a user. The plot is overlaid in light gray with the contours of the log-likelihood fits on the centroid parameter coordinate.*

[7] So far I have not managed to converge on solutions for each sub-population (detailed later) when I include dropping of any kind of friend, but this might change with future attempts and improvements.

In Figure 5.3, we display the values of ϕ and ω involved in the transition for each age group, denoted by age mod 10 (e.g. 12-year-olds is the '2' curve) to the next age, and for each type of respondent, non-smoker ($y_{\text{CIGFLAG}} = 0$) and smoker ($y_{\text{CIGFLAG}} = 1$). For each age transition, the equations contain only one unknown variable; hence, the solutions for dropping and selecting using friends, ϕ and ω, for each age transition lie on a line.

So, for example, if we believe the rate that 12-year-olds (in turning 13) acquire new using friends among all their new friends to be $\omega = 0.15$, then the rate of their dropping using friends would have to be $\phi \approx 0.19$. That is, the point $(\omega = 0.15, \phi \approx 0.19)$ lies on the '2' curve for non-smokers, $y_{\text{CIGFLAG}} = 0$. The age-specific curves flatten for age transitions of older respondents, which reflects the increasing prevalence of smokers; youths gain more and more smoking friends. The curves for users also exhibit, more or less, a similar flattening. We can use the intersections of these curves to triangulate specific estimates of ϕ and ω. For instance, if we only had data for 12-14 year-olds, we would settle on the intersections of the '2' and '3' curves to specify the selection parameters. For non-users, the curves intersect at roughly $(\omega, \phi) \approx (0.05, 0.07)$ and for users, $(0.25, 0.10)$. Not surprisingly, users are selecting more users as friends while non-users appear to be losing using friends, but this reduction is countered by an initiation rate for friends of non-users (i.e. q_{I0} from Table 5.1) that exceeds the dropping rate; hence, we will still see an increase in the number of using friends even among friends of non-users, which verifies the increases in the age-specific λ and θ parameters for non-users.

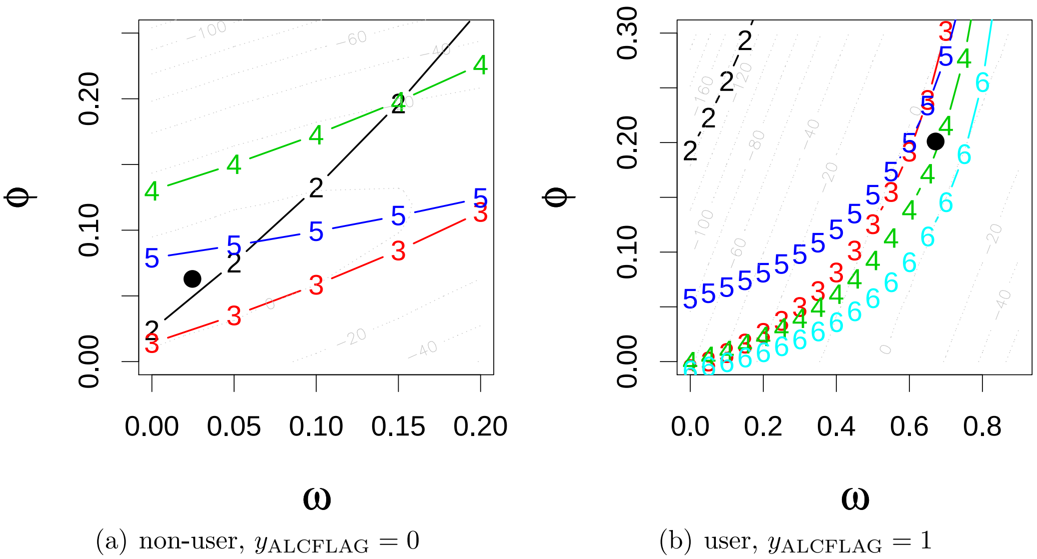
Instead if we wanted a single pair of estimates to describe the dynamic process of friend selection, we would find the centroid of these curves. We fit to the distribution surrounding the number of using friends μ_m^{t+1} and σ_m^{t+1} implicitly expressed by each curve:

$$\begin{aligned}
\psi_{t+1} &= \mu_m + \mu_i - \mu_r + \mu_s \\
&\approx \lambda^{t+1}\theta^{t+1} \\
\mathcal{L}(\phi, \omega | \boldsymbol{\lambda}, \boldsymbol{\theta}, \boldsymbol{\theta}_i) &= \sum_{t=12}^{16} \log[\text{Normal}(\psi^{t+1} | \mu_m^{t+1}, \sigma_m^{t+1})]
\end{aligned}$$

and we obtain the following estimates; note, the mean initiation rate for each subpopulation across all age-groups, θ_i, though not an estimated parameter, is reported here again for informative purposes:

respondent	θ_i	ϕ	ω	σ_ϕ	σ_ω	t_ϕ	p_ϕ	t_ω	p_ω
never smoked	0.075	**0.069**	0.048	0.0209	0.0389	3.30	0.001	1.23	0.217
smoked	0.129	**0.111**	**0.361**	0.0212	0.0562	5.24	0.000	6.42	0.000

Significant parameter estimates are bold-typed. As we suspected, the initiation rate for non-smokers exceeds that of dropping using friends; hence, the prevalence of smoking even among non-using respondents will climb, but not nearly as quickly as it does for using respondents whose selection parameter for new friends is nine times that of non-users!



(a) non-user, $y_{\text{ALCFLAG}} = 0$

(b) user, $y_{\text{ALCFLAG}} = 1$

Figure 5.4: Age Specific Curves for Selecting Alcohol Consuming Friends. *Again, ϕ expresses the proportion of alcohol drinking friends who are dropped and ω is the probability that a new friend drinks alcohol.*

We proceed to look at estimates for ϕ and ω for alcohol and marijuana. In Figure 5.4, we display the appropriate curves. However, the curve for 12-year-old adolescents who have consumed alcohol ($y_{\text{ALCFLAG}} = 1$) does not intersect any of the age curve; so, we omit it from the calculation of the centroid.

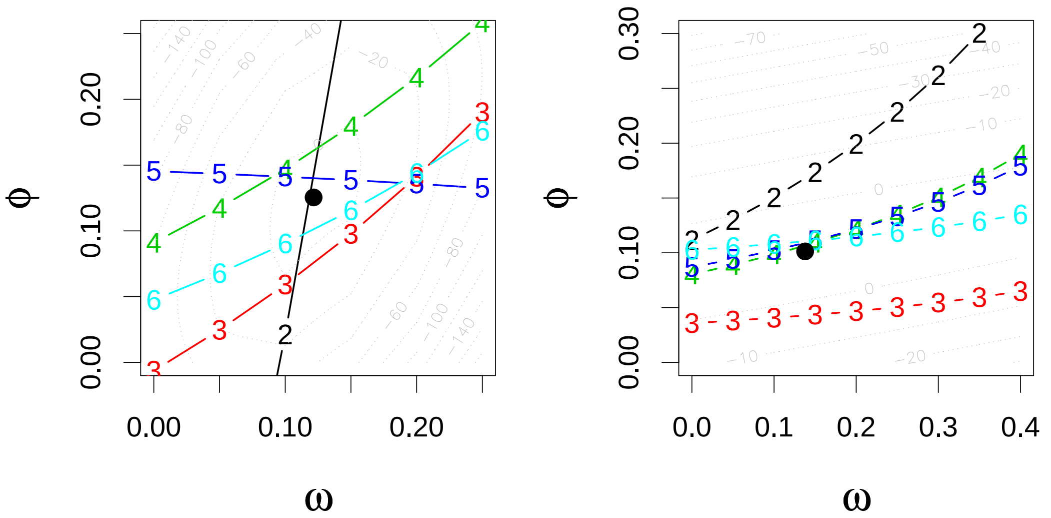
respondent	θ_i	ϕ	ω	σ_ϕ	σ_ω	t_ϕ	p_ϕ	t_ω	p_ω
never drank alc	0.083	**0.063**	0.025	0.021	0.040	3.00	0.003	0.63	0.532
drank alc	0.165	0.201	**0.672**	0.271	0.318	0.74	0.742	2.11	0.035

As with tobacco use, we find non-drinkers to drop more drinking friends than gain and the reverse with drinkers. However, in comparison to smoking, there is more activity with drinkers in the rate of their friends' initiation as well as their network change parameters.

We finally look at marijuana use. In Figure 5.5, as with alcohol drinkers, we have to ignore the 12-year-old curve for marijuana users since it does not intersect any other curse within the [0,1] range of ω.

respondent	θ_i	ϕ	ω	σ_ϕ	σ_ω	t_ϕ	p_ϕ	t_ω	p_ω
never used mrj	0.057	**0.125**	**0.122**	0.016	0.011	7.81	0.000	11.09	0.000
used mrj	0.144	**0.101**	0.138	0.038	0.252	2.84	0.004	0.55	0.583

The network change activity is far more sedate than it is for tobacco or alcohol. In fact, there appears to be an even exchange for non-users. The relative differences in θ_i, ϕ, and ω also speak to the degree in which respondents are embedded



(a) non-user, $y_{\text{MRJFLAG}} = 0$

(b) user, $y_{\text{MRJFLAG}} = 1$

Figure 5.5: Age-Specific Curves for Selecting Marijuana Using Friends. *ϕ is the proportion of marijuana using friends who are dropped and ω is the probability for a gained friend being a marijuana user.*

5.3 Comparing Selection/Influence Parameters to Outside Research

Pearson et al. (2006) examined similar trends in substance use and adolescent networks. Their panel social network data were collected in the Teenage Friends and Lifestyles Study (Michell and Amos, 1997; Pearson and Michell, 2000; Pearson and West, 2003) and comprised both friendship network and substance use data for a cohort of 160 students in the West of Scotland. Selection and influence parameters, which they call homophily and assimilation, were estimated using a co-evolutionary software package called SIENA (Snijders et al., 2005; Steglich et al., 2006b). Their estimated parameters are surprisingly similar to ours.

In Table 5.2, we see there is some similarity between the selection/homophily (p_h vs. $\overline{\omega}$) and influence/assimilation parameters (p_a vs p_I) despite the substantial differences in the data and methodology: Scottish vs. American adolescents, panel cohort vs. cross-sectional age groups, age range of 13-15 vs. 12-17, and sample size of 160 vs. 25052. The SIENA model reports log-odds ratios which we converted into probabilities, p_h and p_a (for homophily and assimilation respectively), to facilitate comparison. While the ordering of the Pearson/Snijders' homophily parameter and the our composite selection parameter, $\overline{\omega}$, does not quite match across substances, the magnitudes come close, with the exception of marijuana/cannabis use. The assimilation parameter, however, does indeed reflect our risk of initiation in their orderings and, also, magnitudes for tobacco and alcohol, but only when compared to the risk

	Selection Parameters			Influence Parameters		
	Pearson/	...		Pearson/	...	
	Snijders'	converted	Lee'	Snijders'	converted	Lee's
	log-odds	to	selection	log-odds	to	initiation
	homophily	p_h	$\overline{\omega}$	assimilation	p_a	$p_I(m = 4)$
smoking	0.42	0.603	0.597	0.39	0.596	0.615
alcohol	0.96	0.723	0.692	1.63	0.836	0.824
marijuana	0.18	0.644	0.509	3.54	0.972	0.883

Table 5.2: Comparing Selection/Initiation Parameters with Pearson/Snijders' Homophily/Assimilation. *Pearson/Snijders' homophily and assimilation (i.e. selection and influence, respectively) parameters (both log-odds and converted probabilities, p_h and p_a) are compared to analogous probability parameters. $\overline{\omega}$ is weighted, composite probability of selecting (both, in and out) similar friends, $[\{\phi_0 + (1 - \phi_1)\} \cdot 0.5 + (1 - \omega_0) + \omega_1]/3$, wherein selecting in is given higher priority than selecting out. $p_I(m = 4)$ is shorthand for $p(y^{t+1}_{x\text{FLAG}} = 1 | y^{t}_{x\text{FLAG}} = 0, m = 4)$ where $x \in \{\text{CIG, ALC, MRJ}\}$ and denotes the probability of initiation a substance given exactly four using friends.*

of initiation under the specific condition of having four using friends; given this restrictive condition and also because we do not model how influence effects cessation of use, the similarity for assimilation/influence is considered weak.

A more precise comparison of assimilation and initiation parameters is possible. In the SIENA model employed by Pearson and Snijders, changes in respondents' substance use behavior was modeled by fitting to log-odds coefficients to the panel data: β_0 for a tendency to use (independent of peer influence) and β_1 for tendency to behave similarly as their friends. We calculate the initiation marginal distribution on m for 13-15 year-old respondents in the NSDUH and estimate similar β coefficients:

		Pearson/Snijders'		Lee's		
		log odds		$\text{logit}(p_I(m)) \approx$		
	tendency	assimilation	difference	$\beta_0 + \beta_1 \cdot m - \beta_1 \cdot (n - m)$		
	β_0	β_1	$\beta_1 - \beta_0$	β_0	β_1	$\beta_1 - \beta_0$
smoking	-3.36	0.39	3.75	-1.82	0.72	2.54
alcohol	0.25	1.63	1.38	-1.38	0.67	2.05
marijuana	-1.02	3.54	4.56	-2.33	1.28	3.61

The β coefficients are noticeably different in absolute magnitudes, which is expected given that we are comparing general use in the Pearson study with initiation of NSDUH respondents. Still, the quantities exhibit similar patterns. For instance, the tendency to drink alcohol (independent of influence) is higher than the tendency to smoke for both Scottish and NSDUH adolescents. Also for both samples, influence is far more responsible for marijuana use than for the other two substances. The

	Hall/Valente's	Lee's Smoking Initiation		
	Smoking Friends	$\vert n<12$,	$\vert n<8$,	$\vert n<10$,
	(6th Grade)	$y_{\mathrm{AGE}} = \cdot$	$y_{\mathrm{AGE}} = 12$	$y_{\mathrm{AGE}} = 13$
	vs. Smoking (7th)	(n<10, n<14)	(n<7, n<9)	(n<8, n<11)
Odds-Ratio	27.05	26.98	26.29	25.46
95% C.I.	(4.53 - 161.40)	(4.50 - 175.48)	(2.13 - 109.20)	(6.61 - 247.35)

Table 5.3: Comparing Smoking Results to Hall/Valente. *Hall/Valente report odds-ratio fit between proportion of friends who smoke in the 6th grade and a binary indicator of self-smoking in the 7th. A respondent is considered a smoker if s/he has ever smoked a cigarette. Odds-ratio between proportion of NSDUH friends smoking and probability of initiation into smoking is directly drawn from the risk of initiation joint distributions. Definition of 'smoking' is identical to that of Hall/Valente.*

NSDUH estimates for alcohol use suggest that alcohol initiation occurs less from peer influence. This finding is not corroborated by Pearson/Snijders, where influence coefficient for drink alcohol is stronger than it is for smoking; we might consider cultural differences to account for some of this disparity. When we compare the magnitudes of tendency to use to assimilation (or influence) for each of the substances, we find a consistent ordering in how much influence exceeds the tendencies to use, with alcohol being the least and marijuana the greatest. This ordering corroborates the earlier findings on selection, ω, in which selection forces accounted for the clustering of alcohol drinkers more than it did for smoking and marijuana use. Selection was the least prominent for marijuana use; the converse is demonstrated here: influence is strongest for marijuana initiation in the NSDUH and also strongest for marijuana use in the Pearson adolescents.

Hall and Valente (2007) explores the influence/selection mechanism, by following 880 6th graders in six, Los Angeles County middle schools into the 7th grade and measuring the extent of smoking among the respondents' five best friends and changes in the respondents' own use. The extent of smoking in each ego-network is described as 'Selecting Smokers'; that is, the proportion of friends whom the respondent listed as being smokers (first at 6th grade and then at 7th grade). Their definition of smoking is identical to the CIGFLAG variable: ever having smoked (lifetime use) qualifies a respondent as a 'smoker'. Their reported result, shown in Table 5.3, is an odds-ratio and modestly matches similar ratios obtained from our joint risk of initiation distribution, but only under the subset of the distribution in which parameters are likely to predict; extreme parts of the distribution such as having fifteen friends with all fifteen smoking biases the calculation of an accurate estimate and, so the distribution is truncated in order to maintain symmetry. Instead of confidence intervals, we can provide other truncations of the distribution that produce an interval of estimation that closely mirrors the C.I. of Hall and Valente's empirical results. The mean of odds-ratios produced for the joint distribution from the risk of initiation for 13 year-olds most closely matches the results of Hall and Valente.

	Hall/Valente's	Lee's Smoking Initiation		
	$Y_{Smoking(7th)} \sim$	$\text{logit}(p_I(n, m, y_{\text{AGE}} = \ldots)) \sim \beta_0 + \beta_1 \text{logit}(m/n)$		
	$\beta_0 + \beta_1 X_{Friends(6th)}$	all ages	12 year-olds	13 year-olds
β_1	0.29	0.1681	0.1719	0.1622
$n, m < 16$		0.2394	0.4277	0.3105

In addition to odds-ratios, Hall and Valente supply beta coefficients from a structural equation model appropriately employing logistic regressions. But here, we see that the β for predicting the 12 year-olds' risk of initiation is closest to that of Hall and Valente's 6th graders smoking in the 7th grade; but all three quantities are clearly off. However, when we lift the restriction on the joint distribution, the β for 13 year-olds almost matches the Hall/Valente β. All these findings should be regarded with some reservation since it was necessary to employ an unbiasing of the joint distributions. Still, it is reasonable to ignore regions of the joint distribution that are not likely to occur and would shift the estimates towards the absurd.

5.4 Sequencing of Covariates

Thus far, we have furthered our knowledge into how adolescents' use behavior and their peer networks co-evolve. However, these dynamics cannot be directly altered by any realistic means; school administrators cannot simply tell students to not smoke or drink and expect them to comply (especially given the strength of both cultural and peer influence) nor can administrators exert control over teens' friendship affiliations. While the enforcement of punitive measures is a modest deterrent, it does not appear to significantly impact use levels, given persistent and widespread substance use among adolescents. Instead, studies seeking to improve intervention and prevention approaches focus on indirect predictors such as youths' attitudes towards substance use by others and also their relationships with their parents (Simons-Morton, 2002). However, it is essential to empirically confirm that the chain of causality is such that these indirect covariates precede the changes in self-use and/or friends' use. If, instead, it is the case that changes in these covariates are the product of the changes in adolescents' use behavior and peer group composition — for example, an adolescent's relationship with his or her parents is strained following the youth's initiation into some substance — then these covariates will be largely useless in informing prevention and intrevention strategies. In this section, we attempt to ascertain the temporal ordering of these events. This process entails comparing the changes in population-level estimates of self-use, peer group composition, to changes in the aforementioned indirect covariates. Admittedly, the analysis at this point is amateur; while a roughly similar technique exists (i.e. latent growth curve analysis), it has not been found to produce estimates of latencies.

We start by looking at an example set of some short time series which are highly inter-correlated. In Figure 5.6, we have three series sampled at unit time points, $t \in$

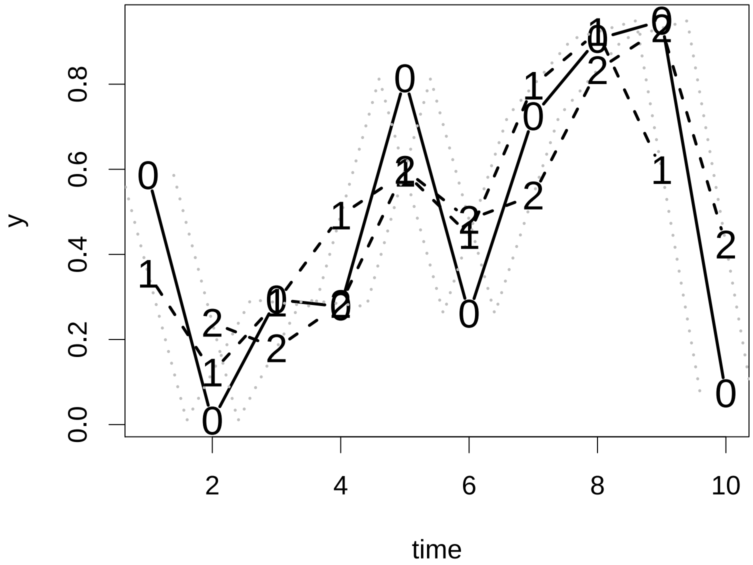


Figure 5.6: Example of Lagged Series. *Two series #1 and #2 are highly correlated with series #0, and perfectly correlated when accounting for the lag.*

[1,10]. We can see that series #1 and #2 are highly correlated with series #0. We can compute lagged correlations which involve lags of unit intervals; we would discover that the original unlagged comparisons fit #0 best. However, we would not uncover the fact that #1 and #2 are perfectly correlated with #0 but with fractional lags; #1 precedes #0 by 0.4 (or lag = -0.4) and #2 follows #0 also by 0.4 (or lag = +0.4). The fractionally lagged series are marked by the gray, dashed lines.

If we assume that changes to these series occur continuously, and not immediately at the sampling points in time, *and* we are also justified in suspecting that the compared series are more similar than they appear, we can attempt to infer the extent of the hypothetical fractional lag by linearly fitting the target series (i.e. #1 or #2) to an alternate series which both correlates better to the source series (i.e. #0) but with a fractional lag *and*, when sampled at the unit intervals, appears identical to the original target series. We look for a fractionally lagged series that minimizes the deviations from both its fit to the primary series and sampled points of the observed series. We employ the following covariates in this analysis:

Covariate	Definition
TALKPAR	# parents can you talk to about important issues[8] $\in \{0 = \text{None}, 1 = \text{One}, 2 = \text{Two}\}$
FDCIG	How many of your friends smoke? $\in \{1 = \text{None}, 2 = \text{Few}, 3 = \text{Most}, 4 = \text{All}\}$
CIGFLAG	Have you ever smoked? $\in \{0 = \text{No}, 1 = \text{Yes}\}$
YEFPKCIG	How do you feel about your friends smoking 1 pack a day? $\in \{1 = \text{Neutral}, 2 = \text{Disapprove}, 3 = \text{Strongly Disapprove}\}$
YEGPKCIG	How do you feel about someone your own age smoking 1 pack a day? $\in \{1 = \text{Neutral}, 2 = \text{Disapprove}, 3 = \text{Strongly Disapprove}\}$

Table H.1 reports that the network, parental attachment, and attitudes toward use covariates, as well as analogous ones for alcohol and marijuana, are found to be significant predictors of ever having used a substance. The table below displays the means (s.d.'s will be used but not shown) for the relevant covariates for each age group:

Age	TALKPAR	FDCIG	CIGFLAG	YEFPKCIG	YEGPKCIG
12	1.320	1.34	0.114	2.68	2.71
13	1.210	1.56	0.216	2.52	2.55
14	1.100	1.78	0.327	2.37	2.42
15	1.050	1.96	0.437	2.28	2.33
16	0.997	2.09	0.515	2.17	2.23
17	1.000	2.17	0.583	2.10	2.22

The high correlations that already exist between the means of these covariates might tempt us to bypass this entire analysis. Instead, we will find the results to be surprisingly compelling. We describe the lagged linear fit of one covariate to another, which is projected to be improved by the unknown lag:

$$\begin{array}{lll} \text{original fit:} & Y & \sim \quad a_0 + a_1 \cdot X \\ \text{lagged fit:} & Y_{lag=x} & \sim \quad b_0 + b_1 \cdot X \end{array}$$

where $Y_{lag=x}$ is that data we would see if there exists a lag of x years. A positive lag x would suggest that X precedes Y and a negative lag, the converse: Y precedes X. Using the Newton-Raphson for fits around a series of points with normal errors, we search for the best lag x and updated coefficients b_0 and b_1 (i.e. the model with the

[8] The TALKPAR variable is constructed from the TALKMOM and TALKDAD indicators, which indicate whether or not a teen discusses serious problems with a parent; this variable is modestly problematic in that there was no way to control for how many parents are available to the teen. Other candidate covariates for parental involvement contained in the NSDUH include behaviors that explicitly originate from the parent: how often the parents tells the respondent how proud they are, how often they tell the respondent to do homework, etc. These will be considered in future work.

lag which fits better than the original if one exists). During fitting, we were required to artificially constrict the standard error around the population mean in order to achieve convergence. The new lagged curve $y^l|lag$ at points x^l is computed using the slope and intercept of the predicting curve y and sampling intervals x:

$$\begin{aligned}
\boldsymbol{x}^l &= \boldsymbol{x} + lag \\
\boldsymbol{m} &= \frac{(\beta_0 + \beta_1 \cdot (y_2, \ldots, y_n)) - (\beta_0 + \beta_1 \cdot (y_1, \ldots, y_{n-1}))}{(x^l_2, \ldots, x^l_n) - (x^l_1, \ldots, x^l_{n-1})} \\
\boldsymbol{b} &= (\beta_0 + \beta_1 \cdot (y_1, \ldots, y_{n-1})) - \boldsymbol{m} \cdot (x^l_1, \ldots, x^l_{n-1}) \\
\boldsymbol{y}^l &= \begin{cases} (\mathrm{NA}, (x_2, \ldots, x_n) \cdot \boldsymbol{m} + \boldsymbol{b})) & \text{if } lag > 0 \\ ((x_1, \ldots, x_{n-1}) \cdot \boldsymbol{m} + \boldsymbol{b}, \mathrm{NA}) & \text{otherwise} \end{cases} \\
\mathcal{L}(lag|\boldsymbol{x}, \boldsymbol{y}, \beta_0, \beta_1) &= \sum_{i=1}^{n} \log[\mathrm{Normal}(y^l_i|y_i, s^2)]
\end{aligned}$$

where s^2 is the constricted standard error, usually by a factor of 0.01. The analysis comprises fitting all pairs of covariates, in both directions. With the five covariates, we end up with 20 pairs, enumerated in Table 5.4. The log-likelihood of the original fit $\mathcal{L}_0$ is compared with the fit incurred with the lagged series $\mathcal{L}_1$; they are for the most part, comparable.

Looking at the bold-typed results, 1. and 2., we see that TALKPAR with some lag is predicted by FDCIG. The lag for TALKPAR ∼ FDCIG is -1.050, meaning a change in TALKPAR *precedes* a change in FDCIG by about a year. When we predict in the opposite direction, we see a positive lag implying that FDCIG *follows* TALKPAR by almost a year, which is almost consistent with the first finding. The implication is that an adolescent's peer group composition will change about a year after a change in the level of confidance with his or her parents.

If our data contained strong or perfect orderings, we would obtain not only consistent signs in the lags for each pair of covariates, but also perfect nesting of lags when certain events occur within the time horizon of others.[9] So, if we continue with just the bold-typed results, and jump to 9. and 10. we see that CIGFLAG follows FDCIG by about 0.8 years. If we wanted to make a causal statement, which we have some allowance to do given our earlier findings, we can say that the time between a change in an adolescent's peer network composition and when his or her risk of initiation also changes is just about 8/10ths of a year. Finally, we complete this mini-sequencing with 3. and 4., where we see that the lag between TALKPAR and CIGFLAG as either 1.86 or 1.47 years. CIGFLAG follows TALKPAR by over year and a half. So now we have a consistent ordering, using the average lag for each pair:

$$\begin{gathered}
\text{TALKPAR} \rightarrow 0.9 \text{ yr} \rightarrow \text{FDCIG} \rightarrow 0.775 \text{ yr} \rightarrow \text{CIGFLAG} \\
\text{TALKPAR} \longrightarrow\longrightarrow\longrightarrow 1.665 \text{ yrs} \longrightarrow\longrightarrow\longrightarrow \text{CIGFLAG}
\end{gathered}$$

[9] Some level of nesting is inevitable and expected given this procedure looks for the best correlations.

Y	$\sim$	X	a_0	a_1	lag=x	b_0	b_1	$\mathcal{L}_1$	$\mathcal{L}_0$
1.TALKPAR	$\sim$	**FDCIG**	1.83	-0.40	-1.05	2.17	-0.54	-3.42	-4.11
2.FDCIG	$\sim$	**TALKPAR**	4.55	-2.46	0.85	4.14	-1.98	-3.26	-3.67
3.TALKPAR	$\sim$	**CIGFLAG**	1.36	-0.69	-1.86	1.61	-0.97	-3.42	-4.12
4.CIGFLAG	$\sim$	**TALKPAR**	1.88	-1.36	1.47	1.53	-0.96	-0.85	-0.65
5.TALKPAR	$\sim$	YEFPKCIG	-0.23	0.57	-0.20	-0.45	0.66	-3.42	-4.11
6.YEFPKCIG	$\sim$	TALKPAR	0.47	1.69	1.02	0.84	1.27	-3.60	-4.07
7.TALKPAR	$\sim$	YEGPKCIG	-0.47	0.66	-0.40	-0.72	0.77	-3.42	-4.11
8.YEGPKCIG	$\sim$	TALKPAR	0.74	1.50	0.45	0.90	1.32	-3.53	-3.97
9.FDCIG	$\sim$	**CIGFLAG**	1.17	1.77	-0.76	0.96	2.00	-2.97	-3.66
10.CIGFLAG	$\sim$	**FDCIG**	-0.66	0.56	0.80	-0.49	0.51	-0.85	-0.63
11.FDCIG	$\sim$	YEFPKCIG	5.21	-1.44	-0.39	5.48	-1.58	-2.97	-3.66
12.YEFPKCIG	$\sim$	FDCIG	3.61	-0.69	-0.22	3.66	-0.70	-3.28	-4.06
13.FDCIG	$\sim$	YEGPKCIG	5.78	-1.64	0.44	5.55	-1.52	-3.26	-3.66
14.YEGPKCIG	$\sim$	FDCIG	3.51	-0.61	-0.37	3.67	-0.67	-3.22	-3.97
15.CIGFLAG	$\sim$	YEFPKCIG	2.27	-0.81	0.70	2.22	-0.76	-0.85	-0.63
16.YEFPKCIG	$\sim$	CIGFLAG	2.80	-1.22	-1.14	3.00	-1.40	-3.28	-4.07
17.CIGFLAG	$\sim$	YEGPKCIG	2.58	-0.92	0.83	2.39	-0.81	-0.85	-0.63
18.YEGPKCIG	$\sim$	CIGFLAG	2.80	-1.06	-1.17	3.01	-1.31	-3.22	-3.98
19.YEFPKCIG	$\sim$	YEGPKCIG	-0.38	1.14	0.60	-0.12	1.00	-3.60	-4.07
20.YEGPKCIG	$\sim$	YEFPKCIG	0.36	0.87	-0.28	0.13	0.98	-3.22	-3.97

Table 5.4: Lagged Fits. *The α coefficients are the intercept and slope for each fit of the means, while the β coefficients arise from a lagged fit. The 'lag=x' is key: the ideal correlation between Y and X is achieved when Y follows X by duration of lag=x. If this lag is negative, then X follows Y. Both directions are examined and are expected to produce slightly differing results due to the covariates being on different scales. The bold-typed pairs are explained first in the text.*

Clearly, these are additive. The upper segment yields 0.9 + .775 = 1.675 yrs, which closely matches the lag predicted between end point covariates in the lower segment: 1.665. The lags from Table 5.4 are depicted graphically in Figure 5.7. The arrows represent the precedence ordering between pairs of covariates and we observe a fairly consistent ordering. A change in number of parents to whom the teen talks (TALKPAR) precedes all other activity, and immediately "affects" attitudes towards general frequent smoking (YEGPKCIG). This in turn precedes both a change in the ego-network composition (FDCIG) as well perception of friends' frequent smoking (YEFPKCIG); we are unsure of the ordering of these last two covariates. Finally, the changes in these attitudes are hypothesized to precede a change in respondents' smoking (CIGFLAG). While we can suppose the effects might occur in reverse order (e.g. a teen's smoking behavior strains the relationship with his or her parents), the analysis suggests that the preponderance of events occurs as stated: changes in teens' attachment to their parents precedes their involvement with substance using friends which precedes their initiation. In Figure 5.8, the FDCIG variable is replaced by the μ_{smoke}, using λ and θ for each age group to generate both the mean and the standard deviation around that mean. The ordering becomes more linear and the duration between a change in the number of using friends and actual smoking initiation has

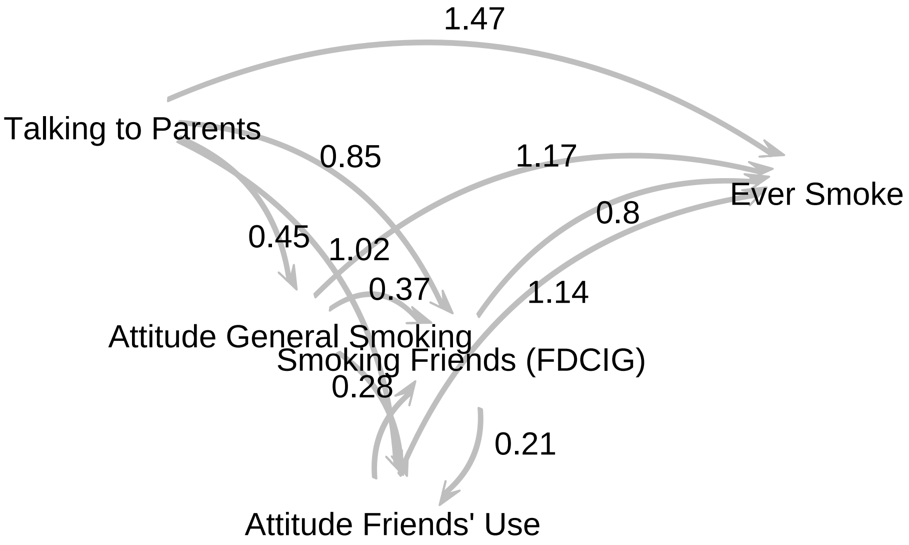


Figure 5.7: Ordering Covariates. *Covariates are spaced according to an MDS output using the lags to represent distances. Due to the some imprecision in the calculations of the MDS, we also display the actual lags, by the length of the arrows and also their numerical values as labels to the arrows.*

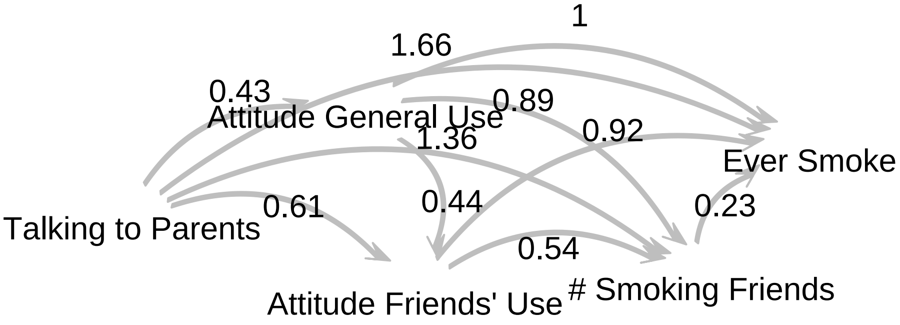


Figure 5.8: Ordering of Covariates II. *Here, friends' use* FDCIG *is replaced by* μ_{smoke}*, the mean number of smoking friends per cohort.*

dropped from 0.80 years to 0.22 years, while the overall time from the change in parental relations to initiation has increased from 1.47 to 1.66.

Recent work (Simons-Morton, 2007) employs latent growth curve analyses on longitudinal data and similarly finds that the negative over-time relationship between parenting practices and adolescent substance use was mediated by the number of substance-using friends. While the methods employed in that research can assess significance in the similarities between the slopes of covariate curves, the simplistic method here does not. On the other hand, those analyses do not report specific durations of latencies like the ones reported here.

Chapter 6

Conclusion

6.1 Summary of Findings

In this dissertation, I executed a series of probability models on partial ego-network, cross-sectional substance use data and extracted estimates of peer use, ultimately, to quantify the selection and influence mechanisms believed to account for much of the behavioral and relational dynamics observed in adolescent substance use. The results have detailed the extent to which influence plays a role, in the form of probabilities of initiation which increase non-linearly with each additional substance using friend; at a critical point, usually two friends, the risk climbs, considerably. However, homophilic, selection forces complement influence in evolving peer groups leading to yet higher levels of substance use homogeneity. In some cases, for instance with smoking, these forces are roughly equal, while with other substances, they are not: evidence shows influence plays a greater role in alcohol and marijuana initiation.

Despite the lack of longitudinal data, these analyses produced some results that confirm those from studies that do employ longitudinal data; this was achieved largely due to the sufficient sample size of the NSDUH and a clear pattern of increasing peer group size and substance use that maintains across periods. However, these findings warrant a more formal treatment of how data from cross-sections might offer solutions similar to those derived from better data. Still, this work can be seen as an argument for valuing cross-sectional data of large samples in order to model events or behavior that have a monotonic relationship with age. Furthermore, the starkly indirect network measures, obtained from the simulated complete networks, mirrored the sudden shift of initiation patterns around the age students enter high school. While we already know initiation to be dependent on local influence forces, the network findings suggest that a more overarching structural dynamic is responsible for how and why initiation drops suddenly around the age of 14.

The methodological steps in this dissertation, some of which are relatively novel and contributive to network analysis as well as adolescent substance use, can be summarized as follows:

- This research set about to add to our understanding adolescent networks using a novel approach. The Poisson/binomial/multinomial mixture has proven useful in inferring distributions of both peer group sizes and the extent of substance use in those groups from ordinal categories of proportions that implicitly cover the [0,1] interval. The estimates easily confirm the homophilic claims of prior research: substance use is a dimension along which peer groups can be distinguished. Furthermore, the joint poly-substance use analysis points to group affiliation by specific combinations of substances, and not just by the category of generic substance use. However, homophilic tendencies are strained for use combinations that are rare, and hence, not many respondents can find peers who share the same combinations. Alternatively, any similarity in rare combinations can be due to influence; this research did not examine this possibility. Furthermore, due to the largely self-reporting nature of the survey, some future consideration ought to be given to the possibility of bias in thinking friends' use is similar to one's own. Also, we were able to verify some of the joint tobacco and alcohol results with those of a recent work that employs longitudinal data and employed ego-network substance use items similar to those of the NSDUH.

- This decomposition of ordinal data required a specification of what a friend "using" means to an adolescent. The evidence suggests that perceived substance "use" among friends varies with age. As population level of use becomes more common, the definition becomes more restrictive. Specifically, the more prevalent substances corresponded to more recent levels of actual use to qualify for perceived "use". Marijuana use is more stigmatic and less commonly used by adolescents than tobacco or alcohol, hence, any experience with this substance qualifies the peer as a "user".

- The estimates provided distributions of in-group and out-group ties between sub-populations of users and non-users which are necessary in the construction an algorithm that generates distributions of complete networks. From these distributions, graph level measures were calculated demonstrating that this matching process produces human-like networks, ones that have specific properties not found in random graphs. While the increasing densities from one age group to the next could be a source of network cacophony, it turns out that older adolescent networks exhibit more complex structures.

- The estimates also allow for a calculation of the risk of initiation as a function of the number of substance using friends. Armed with the knowledge that ego-network parameters increase monotonically over the ages, I inferred transitions from one age to the next and, combined with initiation rates, deduced estimates descriptive of the influence and selection mechanisms. The influence/selection parameters were found to partly coincide with those of two other studies examining peer networks and the dynamics of substance use over time.

- In anticipation of the next step in this research, a final piece of analysis explored the possibility of a consistent ordering of events, including substance use initiation and changes to potentially causative factors like attitudes towards others' use as well as attachment to parents. The findings show not only a consistent ordering, but a specific latency in between changes in states. The latencies can inform intervention and prevention programs of the pattern and ordering of changes in an adolescent's life that will precede initiation, with each change increasingly raising the risk.

6.2 Limitations

Although the use of partial, cross-sectional network data in this work was a conscious choice, its limitations and strengths warrant some comment:

- While the ego-network data is limited to only two of the survey years, the mass of the survey data covers almost two decades, so it remains possible to adjust for some of the age, period, and cohort effects. Some detailed analysis might reveal that even the slight differences between the ego-networks collected in 1998 and 1999 might serve as a source for extrapolating changes to network parameters not just as a function as age but also period and cohort. Furthermore, this ego-network data is grossly incomplete in that it offers no way of deducing the strength and duration of the specified friendships. Still, this problem is one of limited data collection resources and is shared by other studies which collect large networks.

- My analysis primarily considered time as occurring in discrete intervals, in the form of the respondent age groups. This would be appropriate if the survey was administered to all respondents within the same time frame, say a week, and all respondents had the same birthday. Clearly this is not the case and warrants some investigation into relaxing this assumption. At the least, some form of smoothing would slightly alter the trajectories of growth in peer network size and number of using friends.

- While the use of an ordinal scale for describing peer substance use seems inferior to exact counts of friends and friends who use, it is obvious that, unless additional steps are taken to induce sufficient contemplation on the friends' use response, exact counts will exhibit a noticeable pattern of rounding. Furthermore, any claims to whether this unrestricted ordinal scale is more appropriate to exact, but truncated network data, as contained in many of the substance use studies including Add Health, can be laid to rest with further statistical and/or simulation analysis. While close relationships, such as best friends, can be a strong source of influence, weaker relationships can be just as instrumental in the diffusion of behavior (Granovetter, 1973) and can be more likely captured by with a broad survey response item, like the NSDUH friends' use, than a specific one, such as set of best friends.

- The dropping parameter ϕ was simplified to model only dropping of using friends, due to limitations in the data. Instead of simply omitting this component, we can introduce it in the form of sensitivity tests on hypothetical sets of values.

- Often, in friendship-related studies, a distinction is made between close, or best, friends and generic friends; some research has shown such a distinction to be relevant to understanding how substance use influence occurs (Kirke, 1996; Urberg et al., 1997). The NSDUH unfortunately does not allow us to make this distinction, thereby limiting the findings in this research.

- While some distinction was made in levels of use, between someone who use at least once and someone who uses on a more regular basis, the transition analysis did not consider these gradations in use levels. The data will however permit some investigation into changing levels of use from one age to the next, and this avenue will be explored in future writings.

6.3 Future Research

- The original goal of this dissertation was a *dynamic* model of adolescent networks and substance use behavior. While some of the work reported here offers parameters to inform dynamism, the analyses were not executed on populations, real or synthetic, such as the ones generated by the matching algorithm. The next phase of the research will infuse dynamism into distributions of matched networks and we will observe these adolescents' networks change and grow from ages 12 to 17.

- Besides fine-tuning the techniques employed here, by delving deeper into the mechanisms that seem to allow the probability model to work, I can make use of another data set which I ignored throughout this research process, partly because in some respects the data is not as suitable for the methods presented here and partly because of its prior unavailability. However, the sheer size of Monitoring the Future data obligates me now to take a closer look at its potential for informing this line of work.

- The treatment of cross-sectional data has its limitations, especially when employed in the manner that I did. While the investigation of interactions between age, cohort, and period was suspended due to the lack of fully informative data, and also to simplify the presentation of these methods, at this stage, it behooves us to incorporate those effects where ever possible; for instance in the lag analysis, we can simply omit any under-sampled data items like friends' use and instead include responses to the other items spanning additional survey years. Furthermore, if we consider the distributions of friendship sizes to be a constant across periods, we can simply apply the λ estimates as well as the relative θ differences between using and non-using groups into other periods adjusting the θ's to the population (and sub-population) level prevalence rates.

- An additional ordinal use variable asked respondents 'how many students do you know use [smoke, drink alcohol, or use marijuana or hash]?' Preliminary estimates showed both λ and θ to be significantly higher that those of friends' use, which is what we would expect. These estimates can be used in constructing a separate network of weaker ties which can be merged with the friendship network in constructing a network of affiliations weak and strong ties.

- Ego-network linking was performed with only two groups in mind. However, inter-group ties between multiple substance-using groups, such as those examined in joint substance analysis, can be inferred. In fact, Heckathorn (2007) extends his inter-group tie analysis to multiple groups. While the manner in which ties are inferred in this work differs from Heckathorn's, his work demonstrates that such analysis is tractable with our data.

- Also, the ego-network matching process would surely benefit from any priors parameters we might glean from pre-existing complete networks of other studies. For instance, we might learn that the degree of clustering among the using and non-using sub-populations tend be restricted within a certain interval. We can update our generation process to reflect this evidence.

Appendix A

Analysis of Initiation Age and Subsequent Risk of Persistent Use

We explore the sequence of initiation into cigarette use, alcohol consumption and marijuana use, and subsequently, assess the effect of early initiation on the persistence of substance use in adulthood. The following steps detail how we estimate the initiation ages:

1. We estimate the initiation proportion by averaging the probability of initiation. Due to observed recency effects in which older respondents recall their initiation age progressively later, we only consider data from respondents who initiated only a few years prior to their current age. For example, for cigarette use:

   $$p(y_{\text{CIGTRY}} = x) = \frac{\sum_{i=x+1}^{x+5} p(y_{\text{CIGTRY}} = x | y_{\text{AGE}} = i)}{4}$$

   The data we employ here appear in Table A.1 at the end of this appendix chapter.

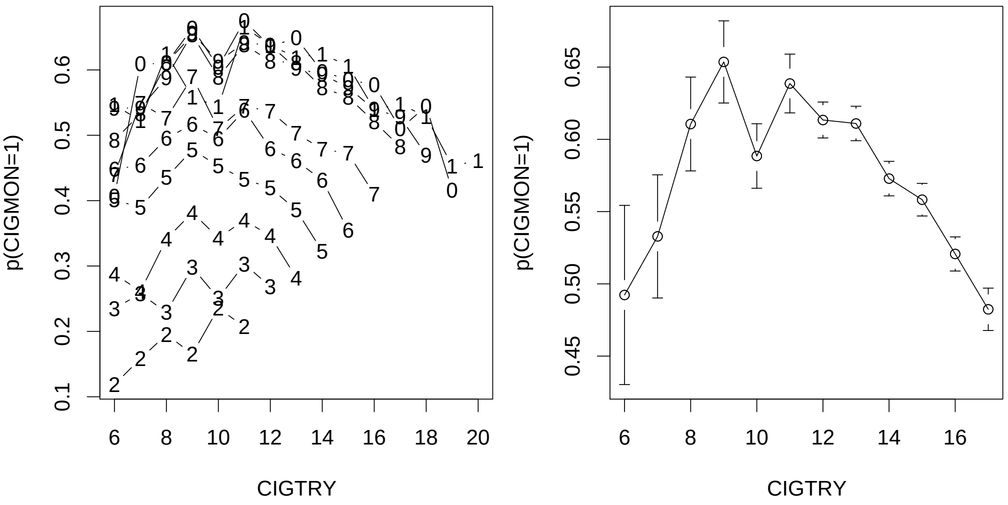
2. Since teens are oversampled, results from these age groups are likely to be have less deviation than later age ranges. As such, the estimated distribution fits consider the variance around the probability estimate from each age group:

   $$\sigma_{ij}^2 = p_{ij} \cdot (1 - p_{ij})/n_{ij}$$

   where $i \in$ [1,30] (i.e. range of initiation ages considered) and $j \in [12, \infty)$ (i.e. age range of respondents). Hence, the fit will give greater weight to those age groups that were over-sampled (i.e. teenagers); the s.d. of their probability estimates will be smaller than those of other age groups.

		Normal dist.			Studentized t dist.			
	mode	μ	σ	$\mathcal{L}$	μ	σ	df	$\mathcal{L}$
Tobacco	15	14.86	3.32	-327.97	14.82	3.01	7.03	-68.50
Alcohol	16	15.68	2.99	-298.90	15.62	2.68	4.48	-19.73
Marijuana	16	15.96	2.65	-177.52	15.94	2.47	6.51	-104.56
Cocaine	18	19.37	4.69	-1493.30	19.34	4.37	10.67	-1431.74

We fit both a normal distribution and a Studentized t distribution using the Newton-Raphson algorithm described in Chapter 2. The estimates from each distribution corroborate one another, and the sequence of initiation is clear: cigarette smoking precedes alcohol consumption, which precedes marijuana use, which precedes cocaine use. These findings are echoed in Elliott et al. (1989). Given that we employed all survey years to obtain these estimates, the differences in initiation ages between substances are all significant. Note, however, that the fit for cocaine is comparatively abysmal due to its skewness to the left, leaving a fat tail to the right. However, a peak initiation age does not necessarily imply the age at which risk of persistent use is highest.



(a) $p(y_{\text{CIGMON}} = 1 | y_{\text{AGE}}, y_{\text{CIGTRY}})$ (b) $p(y_{\text{CIGMON}} = 1 | y_{\text{AGE}} = 18, y_{\text{CIGTRY}})$

Figure A.1: Probability of Last Month Cigarette Use Given Age of Initiation. *On the left, each line represents the current age of the respondent, labeled by a single digit which is* $y_{\text{AGE}} \bmod 10$ *(e.g. 12 year-olds are represented by the '2' curve and 18 year-olds, by the '8' curve). In the right graph, we plot just one curve for 18 year-olds with error bars denoting standard deviations around each proportion estimate. Given the large sample sizes at each point (i.e.* > 1000*), the confidence intervals would not be visible. Also, data from respondents whose age equaled their initiation age are not displayed as last month use often reflects initiation; there would appear an upturn at the end of all the curves.*

In Figure A.1, we plot, for cigarette smoking, ages of initiation against the probability that the respondent used in the last month, a proxy for indicating a smoker. Here, the peaks of each curve lie between ages 9-12. We speculate that this association is not so much enforced by biological differences between younger and older teens but instead the manner in which their use behavior affects the evolution of their peer groups and vice versa. The following table lists the number of 18-year-old respondents who informed last month use at initiation age (right plot):

CIGTRY	6	7	8	9	10	11	12	13	14	15	16	17
n	65	137	226	280	486	559	1539	1676	1716	1940	1788	1163

	$y_{\text{CIGTRY}} = x, y_{\text{AGE}} = x + \ldots$					$y_{\text{AGE}} = x + \ldots$				
x	+1	+2	+3	+4	+5	+1	+2	+3	+4	+5
1	2	0	2	4	5	24903	26790	26581	25877	25805
2	8	6	7	4	5	24903	26790	26581	25877	25805
3	18	25	23	19	14	24903	26790	26581	25877	25805
4	39	31	38	36	25	24903	26790	26581	25877	25805
5	98	117	88	107	94	24903	26790	26581	25877	25805
6	118	149	125	137	116	24903	26790	26581	25877	25805
7	158	225	222	226	249	24903	26790	26581	25877	25805
8	205	288	337	347	319	24903	26790	26581	25877	25805
9	314	349	419	429	418	24903	26790	26581	25877	25805
10	535	719	793	744	796	24903	26790	26581	25877	25805
11	748	924	892	959	840	24903	26790	26581	25877	25805
12	1505	1837	1961	1983	1964	26790	26581	25877	25805	24558
13	2100	2238	2299	2092	1676	26581	25877	25805	24558	20310
14	2223	2231	2149	1716	1575	25877	25805	24558	20310	18451
15	2159	2236	1940	1747	1711	25805	24558	20310	18451	17736
16	1826	1788	1720	1738	1657	24558	20310	18451	17736	17499
17	467	435	391	387	382	7794	6987	6710	6609	6625
18	328	352	387	373	384	6987	6710	6609	6625	6929
19	173	200	194	177	188	6710	6609	6625	6929	7179
20	133	153	194	152	174	6609	6625	6929	7179	7546

Table A.1: Smoking Initiation Rates Pooled from All Survey Years. *The drop in samples after initiation age $x = 16$ denotes the under-sampling of adults, ages > 17. Remember, we do not count same year initiates. Also do due a lumping of adults into broader age categories after 1997, we use only data on adults prior to that year. We use 12-16 year-old respondents to determine initiation counts when age of first trying a cigarette is 11 or earlier.*

Appendix B

Distribution of Friends According to Data

We evaluate the appropriateness of using the Poisson by fitting it to several empirical distributions of peer group size. The data appears in Table B.1. The first distribution comes from data collected in 1987 by Deirdre Kirke, a sociologist, as part of her study on adolescent substance use in Ireland (Kirke, 1996). The second distribution comes from an NSDUH response item, asked in the 1979 and 1982 surveys, in which respondents were asked 'How many close friends do you have, who live in households?' (formal name: CLOSFRNS)

While the distributions are roughly similar, we should not expect a close match given that the Kirke friendship data was collected with greater care and attention than the NSDUH friendship data. For example, the spike in the NSDUH data where respondents expressed having ten close friends is clearly due to a rounding effect; the Kirke data displays no such aberrations. We use a multinomial likelihood to fit the fourteen categories of tabulated responses, n_{Kirke}, on the first fourteen densities of a Poisson, normalized, with a hypothetical parameter λ. That is, we seek the λ that produces a distribution that best fits that data:

$$\mathcal{L}(\lambda = 3.671|(n_{\text{Kirke}})) = -50.75$$

where (n_{Kirke}) denotes a single vector containing the fourteen data points. A Poisson with mean (and variance) 3.671 friends, best fits the Kirke data; this mean is almost identical the the mean number of friends of the Kirke distribution: 3.670. For comparison, a binomial is fit to the data:

$$\mathcal{L}(\theta = 0.2823|n = 13, (n_{\text{Kirke}})) = -84.98; \; n\theta = 3.67$$

where $n = 13$ refers to the maximum possible number of friends in the distribution; the mean is confirmed to be the same. Comparing $\mathcal{L}$ give evidence that friends count for the Kirke data, and perhaps friendship count data in general, arises more likely from a Poisson than a binomial. Finally, the negative-binomial distribution is employed to

Number of Friends	Respondents n_{Kirke}	%	n_{NSDUH}	%
None	2	0.7	179	7.0
One	24	9.0	132	5.2
Two	53	19.9	270	10.6
Three	77	28.8	342	13.4
Four	49	18.4	348	13.7
Five	21	7.9	418	16.4
Six	16	6.0	193	7.6
Seven	6	2.2	135	5.3
Eight	6	2.2	76	3.0
Nine	5	1.9	39	1.5
Ten	5	1.9	329	12.9
Eleven	1	0.4	20	0.8
Twelve	0	0.0	48	1.9
Thirteen	2	0.7	15	0.6
Total	267		2544	

Table B.1: Distribution of Sizes of Peer Group. *The* n_{Kirke} *columns show the tabulation from the respondents in Kirke's study. The* n_{NSDUH} *columns show the distributions of the "close friends" response item administered in the 1979 and 1982 survey years of the NSDUH.*

fit an over-dispersed Poisson and provides an even better fit; the mean $\frac{\alpha}{\beta}$ is identical to the means of the other two distributions.

$$\mathcal{L}(\alpha = 16.64, \beta = 4.53 | (n_{Kirke})) = -47.23; \ \frac{\alpha}{\beta} = 3.67$$

In Figure B.1, the three distributions, binomial, Poisson and negative binomial, are compared to the Kirke data; it is not obvious that there is that great a difference in the fits between the tested distributions.

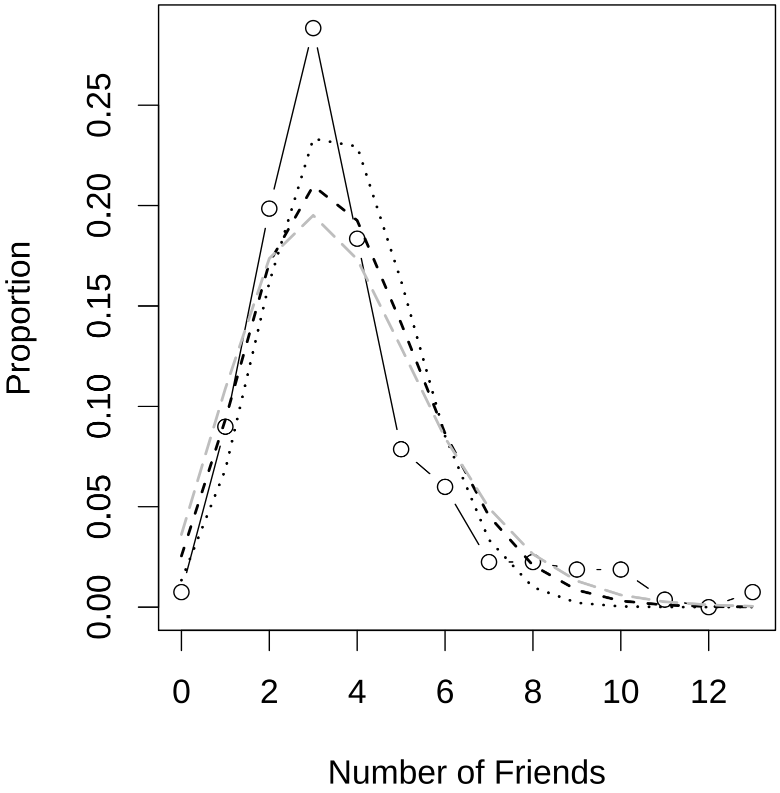
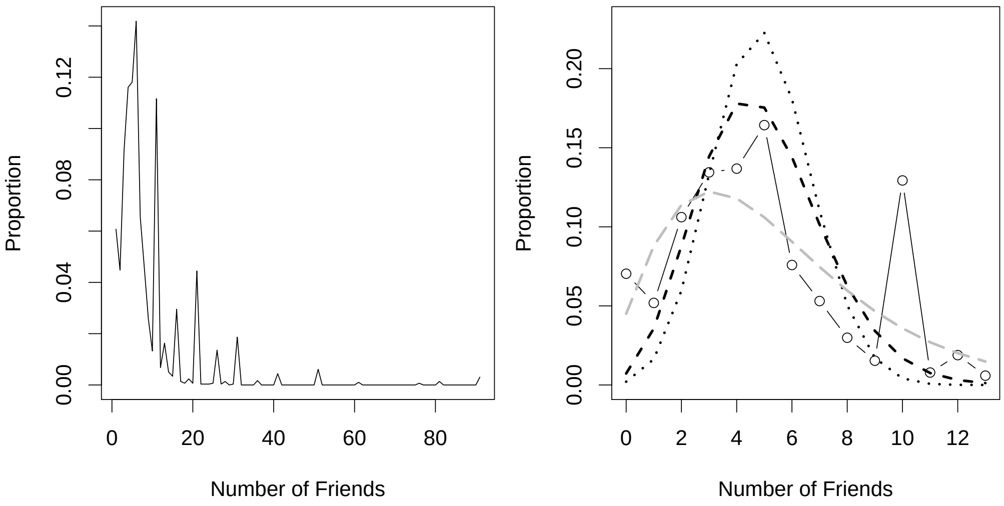


Figure B.1: Kirke Friends Distribution. *Solid line is the Kirke 1996 data. Dashed line is the Poisson fit; the dotted line is the binomial fit; and the gray dashed line is the negative-binomial fit.*

In Figure B.2, we see just how much rounding issues are endemic in the NSDUH close friends data, with clear spiking occurring at 5, 10, 15, 25, 30, etc.[1] Furthermore,

[1] The abnormally higher frequency of responses at multiples of five suggest respondents were either uncertain in their count of friends or unwilling to determine the exact number and settled for a normative response. While no other research that addressed this issue was found, Bruine de Bruin et al. (2000, 2002) examined respondents who settle on 50/50 in evaluating various probabilities of events. We suspect similar mechanisms are at play when respondents are asked to express or recall these quantities.



(a) Entire Sample

(b) Thirteen Friends or Fewer

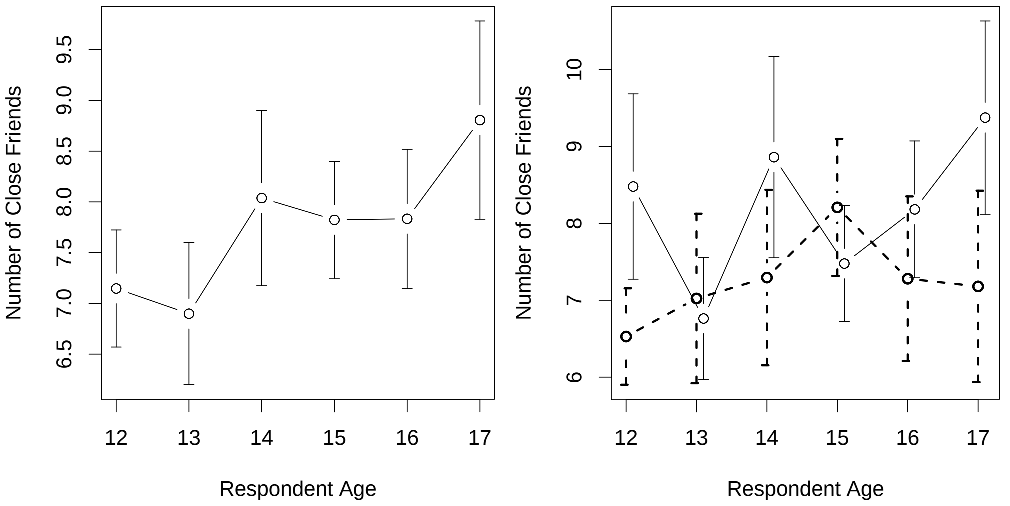
Figure B.2: 1979/1982 NSDUH Close Friends Distribution. *In the left plot, we show the weighted distribution of close friends count for all available adolescents of ages 12-17 in the 1979 and 1982 NSDUH survey years; sample size is* $n = 2949$*. On the right, we truncate the distribution and consider only responses of 13 friends or less; this sub-sample has a size of* $n = 2544$*. The dashed line represents the best Poisson fit while the dotted line, the best binomial.*

a sizable proportion of adolescent respondents (∼400 out of 3000) list having more than 20 close friends. Even when we consider the lack of specificity inherent in the phrase "close friends", we suspect that a lack of demand for accuracy is responsible for the aberration rather than it being the case that this many adolescents have so many close friends. So, we perform our tests on the first fourteen density points, allowing us to make some claims of comparison with the Kirke data. For the purposes of these confirmatory tests, we will retain the spikes as is; we suspect they reflect values surrounding them, and any fit will reflect that. At least visually, the Poisson appears to be the better fit.

$$\begin{array}{llcr}
\text{Poisson:} & \mathcal{L}(\lambda = 4.93|(n_0, \ldots, n_{13})) & = & -887.46 \\
\text{binomial:} & \mathcal{L}(\theta = 0.379|(n_0, \ldots, n_{13})) & = & -1971.85 \\
\text{negative-binomial:} & \mathcal{L}(\alpha = 3.07, \beta = 0.574|(n_0, \ldots, n_{13})) & = & -407.20
\end{array}$$

And, in fact, the Poisson substantially outperforms the binomial. The mean of the binomial is again similar to $\lambda = 4.93$: $n\theta = 4.92$. Even with the truncation, the mean friends here is higher than it was for the Kirke data. We suspect the aforementioned sources of inaccuracy (esp. the spiking) in the NSDUH data accounts for this difference.

We can also use the NSDUH close friends data to confirm the pattern of peer



(a) Peer Group Size Grows With Age (b) Cigarette Users Have More Friends

Figure B.3: Number of Friends, as a Function of Age and Smoking. *In the left plot, we demonstrate the number of close friends gradually increasing in older cohorts. In the right plot, we split the data to those who smoked at some point (solid line) vs. those who never have (dashed line).*

groups growing in size with increasing age and substance use.[2] In Figure B.3, we show the mean number of friends per age group, and their respective confidence intervals. While the mean number of close friends for at-least-once smokers are higher than those who never smoked, the overlap in confidence intervals renders the plot inconclusive. Instead, the following regression analysis reveals that both covariates are significant in predicting the increasing number of close friends; we consider the logarithm transformed data to be more accurate as its errors are normally distributed:

Dep. Var	$n_{friends}$	$\log(n_{friends}+1)$
Intercept	5.079***	1.539***
Age	0.168ˆ	0.020*
Ever Used	0.591ˆ	0.053ˆ
p	0.021*	0.004**
adj-R^2	0.002	0.003

ˆ= $p < 0.10$, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0001$

[2] While a proper analysis would take age, period, and cohort effects into account, there is little reason to believe the number of friends would be subject to variation within the roughly ten years of data required to account of the additional effects.

Appendix C

Additional Decompositions

Indicator y_{CIGFLAG}	λ	θ	σ_λ	σ_θ	$\mathcal{L}$	n	μ_{smoke}
0	3.48	0.198	0.039	0.0025	-314.51	15915	0.70
1	4.43	0.488	0.044	0.0031	-55.43	9135	2.15
				$\Sigma\mathcal{L} =$	-369.94		

y_{CIGRC3}	λ	θ	σ_λ	σ_θ	$\mathcal{L}$	n	μ_{smoke}
0	3.50	0.210	0.037	0.0024	-371.63	17284	0.73
1	4.58	0.513	0.049	0.0033	-29.24	7768	2.35
				$\Sigma\mathcal{L} =$	-400.87		

y_{CIGYR}	λ	θ	σ_λ	σ_θ	$\mathcal{L}$	n	μ_{smoke}
0	3.54	0.229	0.033	0.0023	-460.39	19150	0.79
1	4.66	0.551	0.056	0.0037	-17.97	5899	2.56
				$\Sigma\mathcal{L} =$	-478.36		

y_{CIGMON}	λ	θ	σ_λ	σ_θ	$\mathcal{L}$	n	μ_{smoke}
0	3.58	0.247	0.031	0.0023	-559.49	20911	0.88
1	4.76	0.607	0.067	0.0043	-12.92	4139	2.90
				$\Sigma\mathcal{L} =$	-572.41		

Table C.1: Decomposition by Cigarette Use Indicators

	Smoking in ...			
	Lifetime	Past Three Years	Past Year	Past Month
Age	p_{CIGFLAG}	p_{CIGRC3}	p_{CIGYR}	p_{CIGFLAG}
12	0.121	0.099	0.068	0.042
13	0.213	0.181	0.139	0.078
14	0.322	0.286	0.199	0.126
15	0.440	0.374	0.287	0.198
16	0.512	0.431	0.331	0.247
17	0.582	0.492	0.394	0.307

Table C.2: Age-Specific Proportions for Smoking.

y_{CIGRC3}	y_{AGE}	$\hat{\lambda}$	$\hat{\theta}$	$\sigma_{\hat{\lambda}}$	$\sigma_{\hat{\theta}}$	$\mathcal{L}$	n	μ_{smoke}
0	12	2.91	0.116	0.101	0.0048	-86.72	3607	0.34
0	13	3.38	0.156	0.095	0.0051	-68.47	3486	0.53
0	14	3.70	0.208	0.091	0.0056	-72.59	3105	0.77
0	15	3.92	0.250	0.092	0.0060	-58.26	2720	0.98
0	16	4.08	0.284	0.096	0.0065	-39.61	2321	1.16
0	17	4.31	0.295	0.106	0.0068	-21.31	2043	1.27
1	12	3.63	0.381	0.185	0.0170	-13.89	394	1.38
1	13	4.79	0.426	0.171	0.0105	-13.46	770	2.04
1	14	4.84	0.473	0.133	0.0081	-10.23	1246	2.29
1	15	4.43	0.520	0.102	0.0074	-10.71	1611	2.30
1	16	4.78	0.532	0.107	0.0068	-13.05	1761	2.54
1	17	4.79	0.566	0.099	0.0063	-17.73	1987	2.71
$\Sigma\mathcal{L}$						-426.03		

Table C.3: Estimates for Age and Past Three Years Cigarette Use

y_{CIGYR}	y_{AGE}	$\hat{\lambda}$	$\hat{\theta}$	$\sigma_{\hat{\lambda}}$	$\sigma_{\hat{\theta}}$	$\mathcal{L}$	n	μ_{smoke}
0	12	2.89	0.122	0.095	0.0049	-87.38	3730	0.36
0	13	3.40	0.166	0.090	0.0051	-78.84	3667	0.55
0	14	3.77	0.230	0.082	0.0053	-97.43	3487	0.86
0	15	3.93	0.269	0.083	0.0057	-64.70	3098	1.05
0	16	4.16	0.300	0.088	0.0060	-43.07	2732	1.26
0	17	4.38	0.317	0.095	0.0062	-32.95	2439	1.38
1	12	3.83	0.414	0.228	0.0199	-13.35	270	1.60
1	13	4.96	0.454	0.201	0.0117	-10.32	589	2.25
1	14	4.74	0.514	0.152	0.0098	-9.94	864	2.43
1	15	4.49	0.557	0.117	0.0083	-10.41	1235	2.51
1	16	4.90	0.573	0.123	0.0076	-11.58	1351	2.82
1	17	4.87	0.601	0.111	0.0069	-14.09	1592	2.92
$\Sigma\mathcal{L}$						-474.06		

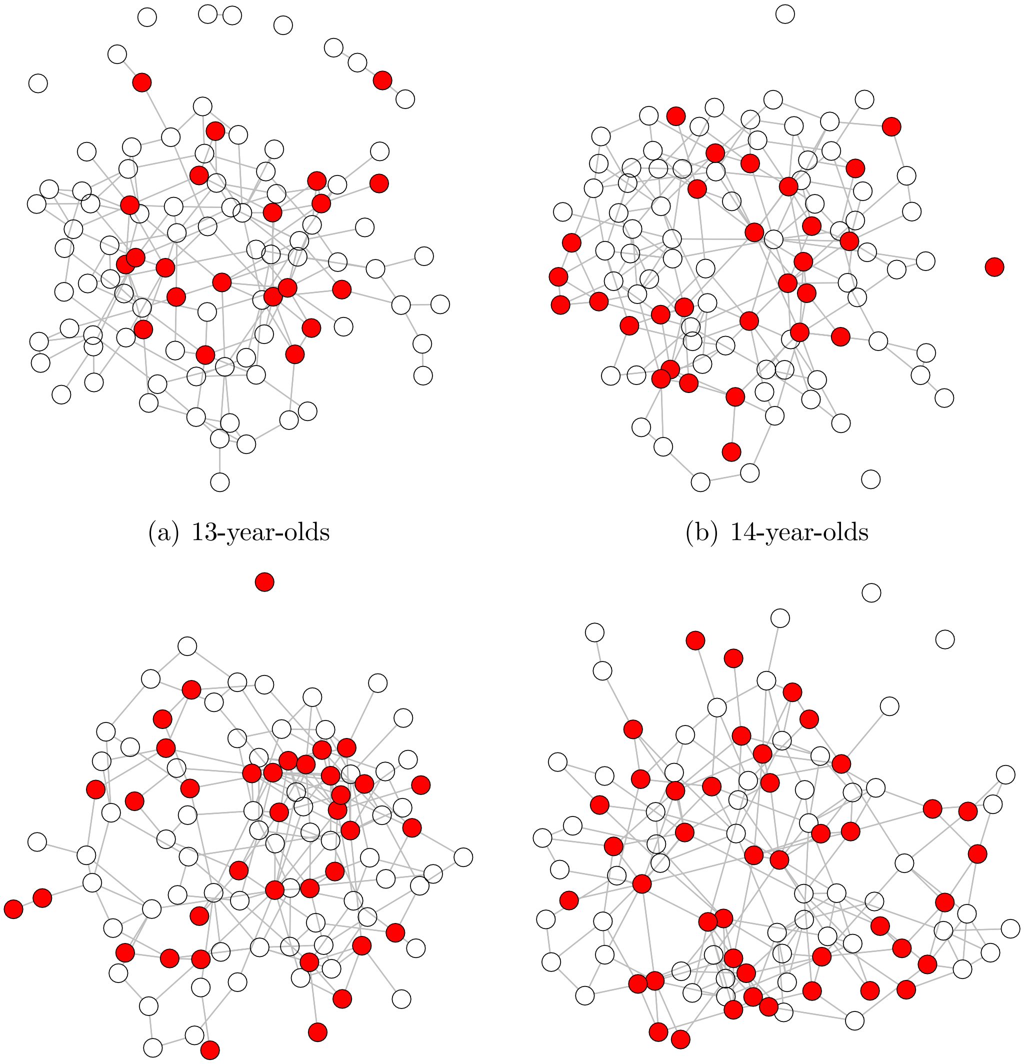
Table C.4: Estimates for Age and Past Year Cigarette Use

y_{CIGMON}	y_{AGE}	$\hat{\lambda}$	$\hat{\theta}$	$\sigma_{\hat{\lambda}}$	$\sigma_{\hat{\theta}}$	$\mathcal{L}$	n	μ_{smoke}
0	12	2.88	0.128	0.091	0.0049	-91.35	3833	0.37
0	13	3.41	0.182	0.082	0.0050	-91.24	3922	0.63
0	14	3.77	0.250	0.075	0.0052	-117.23	3803	0.94
0	15	4.00	0.283	0.077	0.0053	-67.97	3487	1.15
0	16	4.22	0.320	0.082	0.0056	-60.76	3077	1.33
0	17	4.45	0.332	0.089	0.0057	-40.91	2788	1.49
1	12	4.51	0.479	0.331	0.0229	-8.33	168	2.16
1	13	4.89	0.514	0.255	0.0155	-9.89	333	2.52
1	14	4.58	0.560	0.180	0.0123	-10.41	548	2.57
1	15	4.71	0.630	0.146	0.0095	-10.43	843	3.00
1	16	4.87	0.613	0.139	0.0087	-9.97	1004	3.00
1	17	4.99	0.648	0.127	0.0076	-13.40	1242	3.23
$\Sigma\mathcal{L}$						-531.89		

Table C.5: Estimates for Age and Past Month Cigarette Use

Appendix D

Additional Results for Ego-Network Matching



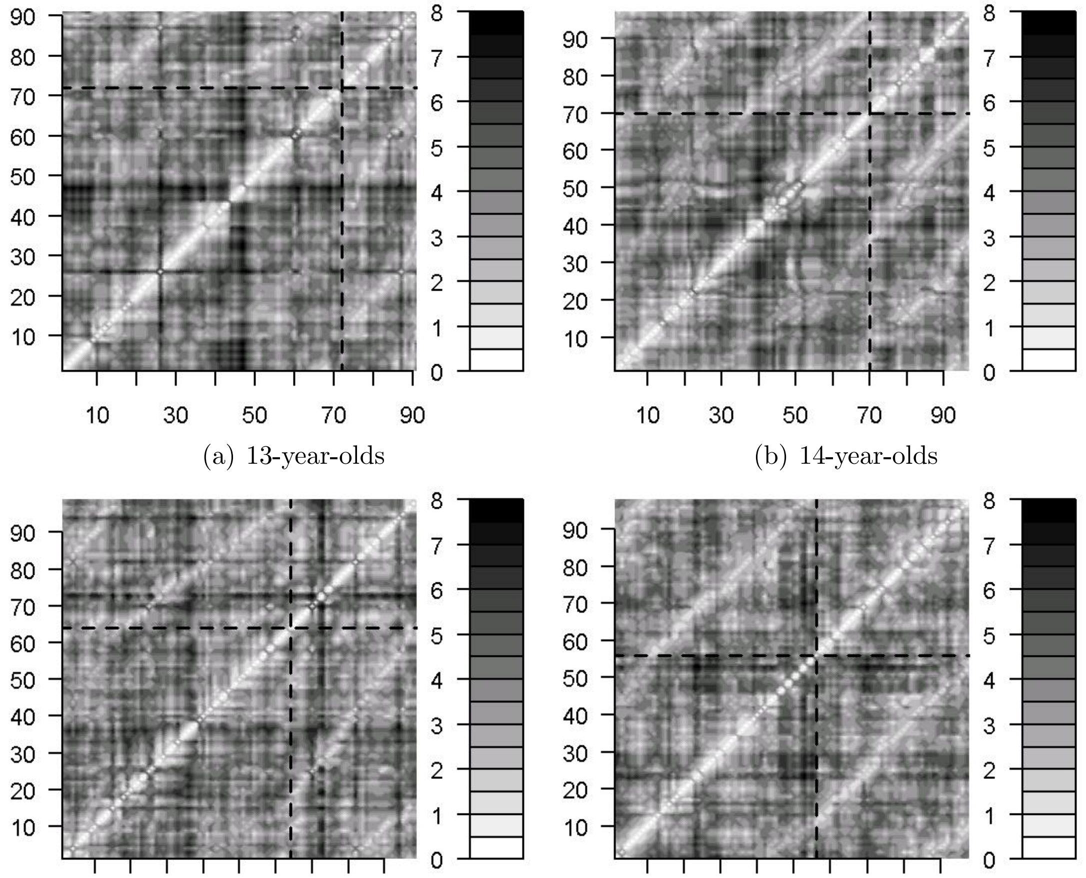
(a) 13-year-olds

(b) 14-year-olds

(c) 15-year-olds

(d) 16-year-olds

Figure D.1: Sample Networks for 13, 14, 15, and 16 year-olds. *Simulated, ego-network matched populations of 100 12-16 year-olds are displayed. The following recency of use indicator variables distinguish perceived smokers from non-smokers, respectively for each age-group:* CIGFLAG (12), CIGFLAG (13), CIGRC3 (14), CIGRC3 (15), CIGRC3 (16), *and* CIGYR (17). *The respective proportions of use are 0.21, 0.29, 0.37, 0.43, and 0.39. Dark colored circles represent cigarette smokers.*



(a) 13-year-olds

(b) 14-year-olds

(c) 15-year-olds

(d) 16-year-olds

Figure D.2: Closeness Between Nodes for 13-16 year-olds. *The geodesics (shortest paths) between all non-isolated nodes (in the main component) to all others appear in shades of gray; lighter color indicates higher closeness. The dashed lines denote the sub-population partition between non-smokers, left and below the partitions, and smokers, to the right and above the partitions. The axes differ slightly due to there being differing number of isolates per age sub-population. The graph is undirected hence the distances and plot are symmetric.*

Appendix E

Auxiliary Results for Linear Model

SEX	CIGRC3	AGE	λ	θ	μ_{smoke}	$\mathcal{L}$	n
0	0	12	3.08	0.109	0.34	-59.00	1765
0	0	13	3.37	0.151	0.51	-30.51	1688
0	0	14	3.76	0.202	0.76	-32.78	1477
0	0	15	3.90	0.260	1.01	-33.82	1316
0	0	16	4.04	0.297	1.20	-25.85	1186
0	0	17	4.02	0.306	1.23	-14.54	1069
0	1	12	3.21	0.432	1.38	-12.46	165
0	1	13	4.57	0.433	1.98	-12.26	376
0	1	14	4.83	0.513	2.48	-12.55	637
0	1	15	4.89	0.532	2.60	-9.49	798
0	1	16	4.79	0.544	2.61	-11.23	850
0	1	17	4.53	0.595	2.69	-11.17	948
1	0	12	2.77	0.121	0.33	-39.82	1841
1	0	13	3.38	0.162	0.55	-45.61	1799
1	0	14	3.66	0.213	0.78	-48.50	1628
1	0	15	3.93	0.241	0.95	-32.96	1404
1	0	16	4.12	0.271	1.11	-21.32	1136
1	0	17	4.69	0.284	1.33	-16.41	974
1	1	12	4.04	0.342	1.38	-9.25	230
1	1	13	5.04	0.420	2.11	-8.97	394
1	1	14	4.93	0.428	2.11	-10.84	609
1	1	15	4.05	0.509	2.06	-9.73	814
1	1	16	4.79	0.520	2.49	-10.40	909
1	1	17	5.11	0.539	2.75	-17.99	1039

Table E.1: Parameter Estimates for Sex, Age, and Past Three Years Cigarette Use

SEX	CIGRC3	AGE	σ_λ	σ_θ	$\sigma_{\mu_{smoke}}$
0	0	12	0.155	0.00663	0.0148
0	0	13	0.140	0.00730	0.0192
0	0	14	0.136	0.00803	0.0266
0	0	15	0.128	0.00871	0.0342
0	0	16	0.130	0.00916	0.0409
0	0	17	0.135	0.00970	0.0442
0	1	12	0.244	0.02805	0.1155
0	1	13	0.230	0.01535	0.1076
0	1	14	0.182	0.01127	0.1002
0	1	15	0.164	0.00995	0.0947
0	1	16	0.153	0.00972	0.0894
0	1	17	0.132	0.00932	0.0861
1	0	12	0.132	0.00701	0.0146
1	0	13	0.130	0.00718	0.0196
1	0	14	0.122	0.00780	0.0261
1	0	15	0.131	0.00833	0.0317
1	0	16	0.142	0.00922	0.0404
1	0	17	0.168	0.00949	0.0498
1	1	12	0.278	0.02097	0.1025
1	1	13	0.254	0.01430	0.1145
1	1	14	0.198	0.01161	0.0923
1	1	15	0.131	0.01095	0.0738
1	1	16	0.150	0.00946	0.0849
1	1	17	0.151	0.00853	0.0862

Table E.2: σ's for Sex, Age, and Past Three Years Cigarette Use

Appendix F

Additional Results from Joint Analysis

	Have You Ever Consumed Alcohol in ...				
		Past	Past	Past	
	Lifetime	3 Yrs	Year	Month	Recency
Intercept	-8.758*** (0.159)	-8.993*** (0.160)	-8.978*** (0.168)	-9.326*** (0.215)	-8.882*** (0.145)
Is Male	0.099* (0.031)	0.102* (0.031)	0.036 (0.032)	0.178* (0.039)	0.123** (0.028)
Age	0.361*** (0.010)	0.373*** (0.010)	0.355*** (0.011)	0.323*** (0.013)	0.344*** (0.009)
Friends' Use	1.116*** (0.022)	1.123*** (0.022)	1.128*** (0.022)	1.093*** (0.025)	1.106*** (0.019)
Adults' Use	0.378*** (0.020)	0.371*** (0.020)	0.328*** (0.021)	0.214*** (0.025)	0.321*** (0.018)
n	24733	24733	24733	24733	24733
Pseudo-R^2	0.397	0.401	0.386	0.315	0.361
p-value	0.000	0.000	0.000	0.000	0.000
AIC	24791	24692	23377	17473	46068
BIC	24832	24732	23418	17513	46133

Table F.1: Alcohol Consumption: Logistic and Ordinal Regressions

	Have You Ever Used Marijuana in ...				
		Past	Past	Past	
	Lifetime	3 Yrs	Year	Month	Recency
Intercept	-11.571***	-11.399***	-10.735***	-10.557***	-11.469***
	(0.226)	(0.227)	(0.244)	(0.318)	(0.213)
Is Male	0.265***	0.251***	0.236**	0.283**	0.268***
	(0.041)	(0.041)	(0.044)	(0.056)	(0.038)
Age	0.418***	0.403***	0.333***	0.266***	0.384***
	(0.014)	(0.014)	(0.015)	(0.020)	(0.013)
Friends' Use	1.378***	1.387***	1.439***	1.417***	1.398***
	(0.028)	(0.028)	(0.030)	(0.035)	(0.026)
Adults' Use	0.743***	0.743***	0.703***	0.657***	0.698***
	(0.032)	(0.032)	(0.034)	(0.040)	(0.029)
n	24526	24526	24526	24526	24526
Pseudo-R^2	0.443	0.441	0.425	0.379	0.383
p-value	0.000	0.000	0.000	0.000	0.000
AIC	15887	15756	13698	9036	25983
BIC	15927	15796	13738	9077	26048

Table F.2: Marijuana Use: Logistic and Ordinal Regressions.

	Alcohol Consumption in ...			
	Lifetime	Past Three Years	Past Year	Past Month
Age	p_{ALCFLAG}	p_{ALCRC3}	p_{ALCYR}	p_{ALCFLAG}
12	0.107	0.097	0.064	0.025
13	0.222	0.212	0.163	0.066
14	0.346	0.332	0.277	0.131
15	0.489	0.479	0.417	0.233
16	0.588	0.576	0.496	0.275
17	0.655	0.649	0.564	0.341

Table F.3: Age-Specific Proportion for Alcohol Consumption

	Marijuana Use in ...			
	Lifetime	Past Three Years	Past Year	Past Month
Age	p_{MRJFLAG}	p_{MRJRC3}	p_{MRJYR}	p_{MRJFLAG}
12	0.022	0.021	0.018	0.006
13	0.052	0.051	0.043	0.024
14	0.124	0.122	0.098	0.056
15	0.223	0.218	0.184	0.102
16	0.297	0.290	0.235	0.121
17	0.362	0.350	0.274	0.157

Table F.4: Age-Specific Proportion for Marijuana Use

	y_{FDMJ} = None				y_{FDMJ} = Few			
	y_{FDALC} =				y_{FDALC} =			
y_{FDCIG}	None	Few	Most	All	None	Few	Most	All
None	10.58	2.98	0.90	0.17	0.11	0.25	0.08	0.02
Few	4.75	23.41	14.11	2.65	0.30	5.15	4.80	0.84
Most	0.93	4.97	4.41	1.39	0.08	3.09	6.41	1.69
All	0.17	0.33	0.28	0.19	0.01	0.21	0.42	0.20

	y_{FDMJ} = Most				y_{FDMJ} = All			
	y_{FDALC} =				y_{FDALC} =			
y_{FDCIG}	None	Few	Most	All	None	Few	Most	All
None	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.02
Few	0.02	0.17	0.32	0.11	0.01	0.13	0.08	0.13
Most	0.01	0.15	1.42	0.29	0.01	0.06	0.07	0.14
All	0.00	0.05	0.23	0.25	0.02	0.02	0.02	0.34

Table F.5: Joint Substance Adults' Use.

Joint Results for Tobacco and Alcohol:

y_{AGE}	λ	θ_{00}	θ_{10}	θ_{01}	θ_{11}	$\mathcal{L}$	n	μ_{CIG}	μ_{ALC}
12	2.60	0.820	0.080	0.024	0.076	-269.94	3972	0.41	0.26
13	3.08	0.730	0.085	0.041	0.145	-319.34	4225	0.71	0.57
14	3.37	0.603	0.098	0.083	0.216	-354.59	4317	1.06	1.01
15	3.55	0.517	0.082	0.107	0.294	-238.68	4296	1.33	1.42
16	3.84	0.447	0.106	0.143	0.304	-258.60	4059	1.57	1.72
17	4.03	0.404	0.101	0.145	0.350	-205.10	4015	1.82	2.00

Joint Results for Tobacco and Marijuana:

y_{AGE}	λ	θ_{00}	θ_{10}	θ_{01}	θ_{11}	$\mathcal{L}$	n	μ_{CIG}	μ_{MRJ}
12	2.58	0.833	0.111	0.008	0.048	-336.26	3975	0.41	0.14
13	3.09	0.761	0.144	0.013	0.082	-334.28	4219	0.70	0.29
14	3.29	0.660	0.164	0.029	0.146	-439.28	4317	1.02	0.58
15	3.44	0.582	0.164	0.041	0.213	-377.02	4300	1.30	0.87
16	3.61	0.518	0.188	0.064	0.230	-414.04	4051	1.51	1.06
17	3.74	0.484	0.184	0.060	0.272	-316.56	4016	1.70	1.24

Joint Results for Alcohol and Marijuana:

y_{AGE}	λ	θ_{00}	θ_{10}	θ_{01}	θ_{11}	$\mathcal{L}$	n	μ_{ALC}	μ_{MRJ}
12	2.29	0.879	0.062	0.013	0.047	-232.85	3979	0.25	0.14
13	2.83	0.786	0.111	0.019	0.084	-337.80	4226	0.55	0.29
14	3.01	0.663	0.151	0.025	0.161	-355.74	4305	0.94	0.56
15	3.26	0.558	0.179	0.029	0.234	-345.36	4293	1.35	0.86
16	3.44	0.506	0.196	0.036	0.262	-341.68	4048	1.57	1.02
17	3.65	0.461	0.206	0.039	0.294	-307.38	4020	1.83	1.22

Table F.6: Age-Specific Results for Friends' Joint Two Substance Use.

Appendix G

Additional Results for Risk of Initiation

Age	λ_i	θ_i	σ_{λ_i}	σ_{θ_i}	$\mathcal{L}$	n_i	μ_i	σ_{μ_i}
12	4.28	0.373	0.503	0.0360	-5.81	74	1.60	0.204
13	4.34	0.334	0.347	0.0233	-6.74	173	1.44	0.125
14	3.97	0.443	0.229	0.0193	-7.98	276	1.76	0.111
15	4.65	0.519	0.244	0.0164	-8.10	315	2.41	0.137
16	4.12	0.428	0.249	0.0196	-7.82	257	1.76	0.117
17	4.40	0.461	0.307	0.0220	-7.36	188	2.03	0.154

Table G.1: Alcohol Initiation by Age

Age	λ_i	θ_i	σ_{λ_i}	σ_{θ_i}	$\mathcal{L}$	n_i	μ_i	σ_{μ_i}
12	4.99	0.358	1.192	0.0659	-3.00	19	1.76	0.468
13	5.07	0.300	0.638	0.0333	-5.32	75	1.51	0.203
14	5.18	0.510	0.461	0.0249	-6.27	121	2.64	0.255
15	4.27	0.406	0.342	0.0252	-6.97	151	1.73	0.154
16	6.04	0.512	0.560	0.0226	-6.18	126	3.09	0.300
17	5.80	0.441	0.596	0.0260	-5.75	103	2.56	0.277

Table G.2: Marijuana Initiation by Age

	Number of Alcohol Drinking Friends, $m = n_{drink}$						
Age	0	1	2	3	4	5	6
12	0.031	0.218	0.711	0.956	0.995	0.999	1.000
13	0.047	0.154	0.403	0.714	0.903	0.972	0.992
14	0.051	0.130	0.290	0.528	0.754	0.894	0.959
15	0.022	0.062	0.162	0.363	0.627	0.832	0.936
16	0.031	0.056	0.101	0.173	0.282	0.424	0.580

Table G.3: Risk of Alcohol Initiation by Age

	Number of Marijuana Smoking Friends, $m = n_{mrj}$						
Age	0	1	2	3	4	5	6
12	0.006	0.085	0.598	0.960	0.997	1.000	1.000
13	0.021	0.122	0.471	0.851	0.974	0.996	0.999
14	0.012	0.072	0.336	0.769	0.956	0.993	0.999
15	0.024	0.071	0.191	0.420	0.690	0.873	0.955
16	0.006	0.028	0.126	0.416	0.778	0.945	0.988

Table G.4: Risk of Marijuana Initiation by Age

	0	1	2	3	4	5	6
0	0.022	0.000	0.000	0.000	0.000	0.000	0.000
1	0.021	0.110	0.000	0.000	0.000	0.000	0.000
2	0.020	0.105	0.400	0.000	0.000	0.000	0.000
3	0.019	0.100	0.388	0.783	0.000	0.000	0.000
4	0.018	0.096	0.376	0.774	0.951	0.000	0.000
5	0.017	0.092	0.365	0.765	0.949	0.991	0.000
6	0.017	0.088	0.353	0.756	0.946	0.990	0.998
7	0.016	0.084	0.342	0.747	0.944	0.990	0.998
8	0.015	0.080	0.331	0.738	0.941	0.989	0.998
9	0.014	0.077	0.320	0.728	0.938	0.989	0.998
10	0.014	0.073	0.310	0.718	0.935	0.988	0.998
11	0.013	0.070	0.299	0.708	0.932	0.987	0.998
12	0.012	0.067	0.289	0.698	0.929	0.987	0.998
13	0.012	0.064	0.279	0.687	0.926	0.986	0.998
14	0.011	0.061	0.269	0.676	0.922	0.985	0.997
15	0.011	0.058	0.259	0.666	0.919	0.985	0.997

Table G.5: Joint Risk of Cigarette Initiation for 12 Year-Olds Given n and m

Appendix H

Additional Results for Sequencing

	Have You Ever ... ?		
	Smoked	Drunk Alcohol	Used Marijuana
	(CIGFLAG)	(ALCFLAG)	(MRJFLAG)
Intercept	-5.465*** (0.180)	-6.773*** (0.178)	-7.677*** (0.263)
Is Male	0.116** (0.032)	0.040 (0.033)	0.215*** (0.045)
Age	0.307*** (0.010)	0.364*** (0.011)	0.393*** (0.016)
Talking to Parents	-0.344*** (0.020)	-0.320*** (0.020)	-0.304*** (0.028)
Friends' Use	0.878*** (0.024)	0.944*** (0.023)	1.017*** (0.031)
Adults' Use	0.293*** (0.024)	0.351*** (0.020)	0.584*** (0.035)
Attitide Towards Friends' Use	-0.201*** (0.024)	-0.068* (0.026)	-0.420*** (0.032)
Attitude Towards Own Age Use	-0.489*** (0.023)	-0.444*** (0.026)	-0.691*** (0.032)
n	24402	24267	24083
Pseudo-R^2	0.392	0.432	0.524
p	0.000	0.000	0.000
AIC	23770	23316	13176
BIC	23810	23357	13216

Table H.1: Predicting Use with Parental Attachment and Attitudes. *The attitude towards use variables vary across the substances. For cigarette smoking, it refers to smoking at least one pack a day; for alcohol consumption, it refers to others' drinking alcohol daily; and for marijuana use, it refers to others' ever having used marijuana. In just the 1998 survey year, respondents are also asked about their attitudes towards others' marijuana use once a month; however, that model underperforms the one presented here.*

Appendix I

Age, Period, and Cohorts

Survey	Age of Respondent, y_{AGE}					
Year	12	13	14	15	16	17
1979	0.339	0.499	0.481	0.571	0.606	0.741
1982	0.260	0.394	0.428	0.514	0.627	0.692
1985	0.253	0.312	0.455	0.504	0.579	0.599
1988	0.186	0.263	0.426	0.452	0.524	0.601
1990	0.169	0.247	0.394	0.434	0.516	0.609
1991	0.165	0.268	0.334	0.426	0.493	0.573
1992	0.141	0.229	0.305	0.411	0.468	0.490
1993	0.152	0.218	0.305	0.420	0.467	0.522
1994	0.154	0.269	0.318	0.419	0.485	0.489
1995	0.142	0.253	0.371	0.482	0.470	0.581
1996	0.154	0.207	0.303	0.443	0.501	0.551
1997	0.139	0.257	0.332	0.457	0.528	0.597
1998	0.128	0.209	0.300	0.442	0.503	0.584
1999	0.114	0.216	0.344	0.439	0.519	0.580
2000	0.103	0.187	0.292	0.435	0.503	0.558
2001	0.098	0.175	0.288	0.404	0.485	0.544
2002	0.087	0.182	0.298	0.399	0.481	0.570
2003	0.081	0.153	0.260	0.379	0.461	0.542

Table I.1: Lifetime Cigarette Use for All Survey Years

Survey	Ages of Transition				
Year	12-13	13-14	14-15	15-16	16-17
1990	0.078	0.146	0.041	0.081	0.093
1991	0.103	0.066	0.092	0.067	0.080
1992	0.088	0.075	0.107	0.057	0.022
1993	0.065	0.087	0.115	0.047	0.055
1994	0.115	0.049	0.100	0.066	0.004
1995	0.111	0.117	0.111	-0.012	0.111
1996	0.053	0.096	0.140	0.058	0.050
1997	0.119	0.074	0.125	0.071	0.069
1998	0.081	0.092	0.141	0.062	0.081
1999	0.103	0.127	0.095	0.080	0.062
2000	0.084	0.105	0.143	0.068	0.055
2001	0.077	0.113	0.116	0.082	0.058
2002	0.095	0.116	0.101	0.082	0.089
2003	0.072	0.107	0.120	0.082	0.081
μ_i	0.089	0.098	0.110	0.064	0.065
σ_i	0.020	0.026	0.026	0.024	0.028

Table I.2: Initiation Rates for Cigarette Use. *Initiation rates are calculated for cohorts across pairs of years. There are no adjacent survey years prior to 1990.*

Bibliography

Adler, Patricia A. 1993. *Wheeling and Dealing: An Ethnography of an Upper-Level Drug Dealing and Smuggling Community*. New York, NY: Columbia University Press.

Agnew, Robert. 1991. "The interactive effects of peer variables on delinquency." *Criminology* 29:47–72.

Akers, Ronald L., Martin D. Krohn, Lonn Lanza-Kaduce, and Marcia Radosevich. 1979. "Social learning and deviant behavior: a specific test of a general theory." *American Sociological Review* 44:635–655.

Alexander, Cheryl, Marina Piazza, Debra Mekos, and Thomas W. Valente. 2001. "Peer networks and adolescent cigarette smoking: An analysis of the national longitudinal study of adolescent health." *Journal of Adolescent Health* 29:22–30.

Almack, John C. 1922. "The influence of intelligence on the selection of associates." *School and Society* 16:529–530.

Almeder, Christian, Jonathan P. Caulkins, Gustav Feichtinger, and Gernot Tragler. 2000. "Age-specific multi-stage drug initiation models: insights from considering heterogeneity." *U.N. Bulletin on Narcotics* July 2000. RR246.

Almeder, Christian, Jonathan P. Caulkins, Gustav Feichtinger, and Gernot Tragler. 2004. "An Age-Structured Single-State Initiaton Model – Cycles of Drug Epidemics and Optimal Prevention Programs." *Socio-Economic Planning Sciences* 38:91–109.

Anderson, Brigham S., Carter T. Butts, and Kathleen M. Carley. 1999. "The interaction of size and density with graph level indices." *Social Networks* 21:239–267.

Ardelt, Monika and Laurie Day. 2002. "Parents, Siblings, and Peers: Close Social Relationships and Adolescent Deviance." *Journal of Early Adolescence* 22:310–349.

Arrow, Kenneth J. 1951. *Social Choice and Individual Values*. New York, NY: Wiley.

Bachman, Jerald G., Lloyd D. Johnston, and Patrick M. O'Malley. 1990. "Explaining the recent decline in cocaine use among young adults: further evidence that perceived risks and disapproval lead to reduced drug use." *Journal of Health and Social Behavior* 31:173–184.

Bachman, Jerald G., Lloyd D. Johnston, Patrick M. O'Malley, and Ronald H. Humphrey. 1988. "Explaining the recent decline in marijuana use among young adults: differentiating the effects of perceived risks, disapproval, and general lifestyle factors." *Journal of Health and Social Behavior* 29:92–112.

Baker, Wayne E. and Robert R. Faulkner. 1993. "The social organization of conspiracy: illegal networks in the heavy electrical equipment industry." *American Sociological Review* 58:837–860.

Banks, David L. and Kathleen M. Carley. 1997. "Models for network evolution." *Journal of Mathematical Sociology* 21:173–196.

Bauman, Karl E. and Susan T. Ennett. 1996. "On the importance of peer influence for adolescent drug use: commonly neglected considerations." *Addiction* 91:185–198.

Bearman, Peter S., James Moody, and Katherine Stovel. 2004. "Chains of affection: the structure of adolescent romantic and sexual networks." *American Journal of Sociology* 110:44–91.

Behrens, Doris A., Jonathan P. Caulkins, Gernot Tragler, and Gustav Feichtinger. 2000. "Optimal control of drug epidemics: prevent and treat – but not at the same time?" *Management Science* 46:333–347.

Behrens, Doris A., Jonathon P. Caulkins, Gernot Tragler, and Gustav Feichtinger. 2002. "Why present-oriented societies undergo cycles of drug epidemics." *Journal of Economic Dynamics and Control* 26:919–936.

Behrens, Doris A., Jonathan P. Caulkins, Gernot Tragler, Josef L. Haunschmied, and Gustav Feichtinger. 1999. "A Dynamic Model of Drug Initiation: Implications for Treatment and Drug Control." *Mathematical Biosciences* 159:1–20. RAND Reprint RP-910.

Biemer, Paul P. and Michael Witt. 1996. "Estimation of measurement bias in self-reports of drug use with applications to the National Household Survey on Drug Abuse." *Journal of Official Statistics* 12:275–300.

Biemer, Paul P. and Michael Witt. 1997. "Repeated measures estimation of measurement bias for self-reported drug use with applications to the National Household Survey on Drug Abuse." In *NIDA Resesarch Monograph 167: The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates*, edited by Lana Harrison and Arthur Hughes. Rockville, MD: National Institute on Drug Abuse.

Blau, Peter M. 1977. *Inequality and Heterogeneity: A Primitive Theory of Social Structure*. New York, NY: Free Press.

Blau, Peter M. and Joseph E. Schwartz. 1984. *Crosscutting Social Circles: Testing a Macrostructural Theory of Intergroup Relations*. New York, NY: Academic Press.

Borgatti, Stephen P. and Martin G. Everett. 1999. "Models of core/periphery structures." *Social Networks* 21:375–395.

Bott, Helen. 1928. "Observation of play activities in a nursery school." *Genetic Psychology, Monographs* 4:44–88.

Boyd, John P., William J. Fitzgerald, and Robert J. Black. 2006. "Computing core/periphery structures and permutatoin tests for social relations data." *Social Networks* 28:165–178.

Brook, J.S., M. Whiteman, and A.S. Gordon. 1983. "Stages of drug use in adolescence: personality, peer and family correlates." *Developmental Psychology* 19:269–277.

Bruine de Bruin, Wandi, Paul S. Fischbeck, Neil A. Stiber, and Baruch Fischhoff. 2002. "What number is 'fifty-fifty'? Redistributing excess 50responses in risk perception studies." *Risk Analysis* 22:725–735.

Bruine de Bruin, Wandi, Baruch Fischhoff, Susan G. Millstein, and Bonnie L. Halpern-Felsher. 2000. "Verbal and numerical expressions of probability: 'It's a fifty-fifty chance.'." *Organizational Behavior and Human Decision Processes* 81:115–131.

Burt, Ronald S. 1984. "Network items and the General Social Survey." *Social Networks* 6:293–339.

Burt, Ronald S. 1985. "General social survey network items." *Connections* 8:119–123.

Carley, Kathleen M. 1990. "Group stability: a socio-cognitive approach." In *Advances in Group Processes*, edited by Edward Lawler, Barry Markovsky, Cecilia Ridgeway, and Henry Walker, volume 7, pp. 1–44. Greenwich, CT: JAI Press.

Carley, Kathleen M. 1991. "A Theory of Group Stability." *American Sociological Review* 56:331–354.

Carley, Kathleen M. 2003. "Dynamic Network Analysis." In *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, edited by Ronald Breiger, Kathleen M. Carley, and Philippa Pattison. Washington, D.C.: National Academies Press.

Carley, Kathleen M., Ju-Sung Lee, and David Krackhardt. 2001. "Destabilizing Networks." *Connections* 24:31–34.

Caulkins, Jonathan P. 1993. "Local drug markets' response to focused police enforcement." *Operations Research* 41:848–863.

Caulkins, Jonathan P. 2000a. "The evolution of drug initiation: from social networks to public markets." In *Optimization, Dynamics, and Economic Analysis – Essays in Honor of Gustav Feichtinger*, edited by Engelbert J. Dockner, Richard F. Hartl, Mikulas Luptacik, and Gerhard Sorger, pp. 353–367. Heidelberg, DE: Physica-Verlag.

Caulkins, Jonathan P. 2000b. "Measurement and analysis of drug problems and drug control efforts." In *Criminal Justice 2000, Volume 4: Measurement and Analysis of Crime and Justice*, edited by David Duffee, David McDowall, Lorraine Green Mazerolle, and Stephen D. Mastrofski. Washington, DC: National Institute of Justice. 2000:391449.

Caulkins, Jonathan P. 2000c. "Measurement and analysis of drug problems and drug control efforts." *Measurement and Analysis of Crime and Justice* 4:391–449.

Caulkins, Jonathan P., Doris A. Behrens, Claudia Knoll, Gernot Tragler, and Doris Zuba. 2004. "Markov Chain Modeling of Initiation and Demand: The Case of the US Cocaine Epidemic." *Health Care Management Science* 7:319–329.

Caulkins, Jonathan P., Susan S. Everingham, C. Peter Rydell, James Chiesa, and Shawn Bushway. 1999. *An Ounce of Prevention, A Pound of Uncertainty*. Santa Monica, CA: RAND: Drug Policy Research Center.

Chassin, Laurie, Clark C. Presson, and Steven J. Sherman. 1984. "Cigarette smoking and adolescent psychosocial development." *Basic and Applied Social Psychology* 5:295315.

Cohen, Jere M. 1977. "Sources of peer group homogeneity." *Sociology of Education* 50:227–241.

Coleman, James, Elihu Katz, and Herbert Menzel. 1957. "The diffusion of an innovation among physicians." *Sociometry* 20:253–270.

Curran, Geoffrey M., Helene R. White, and Stephen Hansell. 2000. "Personality, environment, and problem drug use." *Journal of Drug Issues* 30:375–406.

Curtis, Richard, Samuel R. Friedman, Alan Neaigus, Benny Jose, Marjorie Goldstein, and Gilbert Ildefonso. 1995. "Street-level drug markets: Network structure and HIV risk." *Social Networks* 17:229–249.

D'Amico, Elizabeth J. and Denis M. McCarthy. 2006. "Escalation and Initiation of Younger Adolescents' Substance Use: The Impact of Perceived Peer Use." *Journal of Adolescent Health* 39:481–487.

Dangalchev, Chavdar. 2006. “Residual closeness in networks.” *Physica A: Statistical Mechanics and Its Applications* 365:556–564.

Doreian, Patrick, Roman Kapuscinski, David Krackhardt, and Janusz Szczypula. 1996. “A brief history of balance through time.” *Journal of Mathematical Sociology* 21:113–131.

Doreian, Patrick and Frans N. Stokman. 1997. “The Dynamics and Evolution of Social Networks.” In *Evolution of Social Networks*, edited by Patrick Doreian and Frans N. Stokman, pp. 1–17. New York, NY: Gordon and Breach.

Edelen, Maria Orlando, Joan S. Tucker, and Phyllis L. Ellickson. 2006. “A discrete time hazard model of smoking initiation among West Coast youth from age 5 to 23.” *Preventive Medicine* 44.

Ellickson, Phyllis L. and Robert M. Bell. 1990. “Drug prevention in junior high: A multi-site longitudinal test.” *Science, New Series* 247:1299–1305.

Elliott, Delbert S., David Huizinga, and Scott Menard. 1989. *Multiple Problem Youth: Delinquency, Substance Use, and Mental Health Problems*. New York, NY: Springer-Verlag.

Ennett, Susan T. and Karl E. Bauman. 1993. “Peer group structure and adolescent cigarette smoking: a social network analysis.” *Journal of Health and Social Behavior* 34:226–236.

Ennett, Susan T. and Karl E. Bauman. 1994. “The contribution of influence and selection to adolescent peer group homogeneity: the case of adolescent cigarette smoking.” *Journal of Personality and Social Psychology* 64:653–663.

Ennett, Susan T., Karl E. Bauman, Vangie A. Foshee, Michael Pemberton, and Kathern A. Hicks. 2001. “Parent-Child communication about adolescent tobacco and alcohol use: what do parents say and does it affect youth behavior?” *Journal of Marriage and Family* 63:48–62.

Erickson, Bonnie E. 1981. “Secret societies and social structure.” *Social Forces* 60:188–210.

Everett, S.A., C.W. Warren, D. Sharp, L. Kann, C.G. Husten, and L.S. Crossett. 1999. “Initiation of cigarette smoking and subsequent smoking behavior among U.S. high school students.” *Preventive Medecine* 29:327–333.

Everingham, Susan S. and C. Peter Rydell. 1994. “Modeling the Demand of Cocaine.” Research Report MR-332-ONDCP/A/DPRC, RAND, Santa Monica, CA.

Fararo, Thomas J. and John Skvoretz. 1984. “Biased networks and social structure theorems: part II.” *Social Networks* 6:223–258.

Fararo, Thomas J. and Morris H. Sunshine. 1964. *A Study of a Biased Friendship Net*. Syracuse, NY: Syracuse University Youth Development Center and Syracuse University Press.

Festinger, Leon. 1950. "Informal social communication." *Psychological Review* 57:217–282.

Festinger, Leon. 1953. "An Analysis of Compliant Behavior." In *Group Relations at the Crossroads*, edited by M. Sherif and M. O. Wilson, pp. 232–256. New York, NY: Harper.

Festinger, Leon. 1954. "A theory of social comparison processes." *Human Relations* 7:117–140.

Fischer, Claude S. 1977. *Networks and Places: Social Relations in the Urban Setting*. New York, NY: Free Press.

Fischer, Claude S. 1982. *To Dwell among Friends*. Chicago, IL: University of Chicago Press.

Flom, Peter L., Samuel R. Friedman, Alan Neaigus, and Milagros Sandoval. 2003. "Boundary-crossing and drug use among young adults in a low-income, minority, urban neighborhood." *Connections* 25:77–87.

Frank, Ove. 1977. "Survey sampling in graphs." *Journal of Statistical Planning and Inference* 1:235–264.

Frank, Ove. 1979. "Estimation of population totals by use of snowball samples." In *Perspectives on Social Network Research*, edited by P.W. Holland and S. Leinhardt. New York, NY: Academic Press.

Frank, Ove and Tom A. B. Snijders. 1994. "Estimating the size of hidden populations using snowball sampling." *Journal of Official Statistics* 10:53–67.

Fraser, Mark and J. David Hawkins. 1984. "Social network analysis and drug misuse." *Social Science Review* 58:81–97.

Freeman, Linton C. 1977. "A set of measures of centrality based on betweenness." *Sociometry* 40:35–41.

Freeman, Linton C. 1979. "Centrality in social networks: a conceptual clarification." *Social Networks* 1:215–239.

Freeman, Linton C., A. Kimball Romney, and Sue C. Freeman. 1987. "Cognitive Structures and Informant Accuracy." *American Anthropologist* 89:310–325.

Frey, Frederick W., Elias Abrutyn, David S. Metzger, George E. Woody, Charles P. O'Brien, and Paul Trusiani. 1995. "Focal networks and HIV risk among African-American male intravenous drug users." In *NIDA Monograph 151: Social Networks, Drug Abuse, and HIV Transmission*, edited by Richard H. Needle, Susan L. Coyle, Sander G. Genser, and Robert T. Trotter II, pp. 89–108. Rockville, MD: U.S. Dept. of Health and Human Services.

Friedkin, Noah and Eugene Johnsen. 1999. "Social influence networks and opinion change." In *Advances in Group Processes*, edited by Shane R. Thye, Michael W. Macy, Henry Walker, and Edward Lawler, volume 16, pp. 1–29. Stamford, CT: JAI Press.

Friedkin, Noah E. 1990. "Social networks in structural equation models." *Social Psychology Quarterly* 53:316–328.

Friedkin, Noah E. 1993. "Structural bases of interpersonal influence in groups." *American Sociological Review* 58:861–872.

Friedkin, Noah E. 1998. *A Structural Theory of Social Influence*. New York, NY: Cambridge University Press.

Friedkin, Noah E. and Eugene C. Johnsen. 1990. "Social Influence and Opinions." *Journal of Mathematical Sociology* 15:193–205.

Friedman, Samuel R. 1995. "Promising social network research results and suggests for a research agenda." In *NIDA Monograph 151: Social Networks, Drug Abuse, and HIV Transmission*, edited by Richard H. Needle, Susan L. Coyle, Sander G. Genser, and Robert T. Trotter II, pp. 196–215. Rockville, MD: U.S. Dept. of Health and Human Services.

Friedman, Samuel R., Richard Curtis, Alan Neaigus, Benny Jose, and Don C. Des Jarlais. 1999. *Social Networks, Drug Injectors' Lives, and HIV/AIDS*. New York: Kluwer Academic.

Friedman, Samuel R., Alan Neaguis, Benny Jose, Richard Curtis, Marjorie Goldstein, Gilbert Ildefonso, Richard Rothenberg, and Don C. Des Jarlais. 1997. "Sociometric risk networks and risk for HIV infection." *American Journal of Public Health* 87:1289–1296.

Gainey, Randy R., Peggy L. Peterson, Elizabeth A. Wells, J. David Hawkins, and Richard F. Catalano. 1995. "The social networks of cocaine users seeking treatment." *Addiction Research* 3:17–32.

Gaughan, Monica. 2003. "Predisposition and pressure: mutual influence and adolescent drunkenness." *Connections* 25:17–31.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2003. *Bayesian Data Analysis*. Boca Radon, FL: Chapman and Hall/CRC.

Granovetter, Mark. 1973. “Strength of weak ties.” *American Jounral of Sociology* 78:1360–1380.

Granovetter, Mark S. 1976. “Network sampling: some first steps.” *American Journal of Sociology* 81:1287–1303.

Hagman, Elizabeth P. 1933. “The companionships of preschool children.” *University of Iowa Studies in Child Welfare* 7:10–69.

Hahn, Ginger, Ventura L. Charlin, Steve Sussman, Clyde W. Dent, Jorge Manzi, Alan W. Stacy, Brian Flay, William B. Hansen, and Dee Burton. 1990. “Adolescent’s first and most recent use situations of smokeless tobacco and cigarettes: Similarities and differences.” *Addictive Behaviors* 15:414–430.

Hall, Jeffrey A. and Thomas W. Valente. 2007. “Adolescent smoking networks: the effects of influence and selection on future smoking.” *Addictive Behaviors* 32:3054–3059.

Hallinan, Maureen T. 1978. “The process of friendship formation.” *Social Networks* 1:193–210.

Hallinan, Maureen T. and Warran N. Kubitschek. 1988. “The effects of individual and structural characteristics on intransitivity in social networks.” *Social Psychology Quarterly* 51:81–92.

Hallinan, Maureen T. and R. A. Williams. 1987. “The stability of students interracial friendships.” *American Sociological Review* 52:653–664.

Hallinan, Maureen T. and Richard A. Williams. 1990. “Students’ characteristics and the peer-influence process.” *Sociology of Education* 63:122–132.

Harary, Frank. 1969. *Graph Theory*. Reading, MA: Addison-Wesley.

Harrison, Lana. 1997. “The validity of self-reported drug use in survey research: an overview and critique of research methods.” In *NIDA Resesearch Monograph 167: The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates*, edited by Lana Harrison and Arthur Hughes. Rockville, MD: National Institute on Drug Abuse.

Harrison, Lana D., Steven S. Martin, Tihomir Enev, and Deborah Harrington. 2007. “Comparing Drug Testing and Self Report of Drug Use Among Youths and Young Adults in the General Population.” Technical Report DHHS Publication No. (SMA)07-4249, Substance Abuse and Mental Health Services Administration, Office of Applied Studies, Rockville, MD.

Hawkins, J. David and Mark W. Fraser. 1985. “The social networks of street drug users: a comparison of two theories.” *Social Work Research and Abstracts* 4:3–12.

Heckathorn, Douglas D. 1997. “Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations.” *Social Problems* 44:174–199.

Heckathorn, Douglas D. 2002. “Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations.” *Social Problems* 49:11–34.

Heckathorn, Douglas D. 2007. “Extensions of Respondent-Driven Sampling: Analyzing Continuous Variables and Controlling for Differential Recruitment.” *Sociological Methodology* pp. 151–208. forthcoming.

Heckathorn, Douglas D., Robert S. Broadhead, Denise L. Anthony, and David L. Weakliem. 1999. “AIDS and social networks: HIV prevention through network mobilization.” *Sociological Focus* 32:159–179.

Heckathorn, Douglas D., S. Semaan, R.S. Broadhead, and J.J. Hughes. 2002. “Extensions of Respondent-Driven Sampling: A New Approach to the Study of Injection Drug Users Aged 18-25.” *AIDS and Behavior* 6:55–67.

Hubbard, Ruth M. 1929. “A method of studying spontaneous group formation.” In *In Some New Techniques for Studying Social Behavior*, edited by Dorothy Swaine Thomas, pp. 76–85. New York, NY: Teachers College, Columbia University, Child Development Monographs.

Iannotti, Ronald J. and Patricia J. Bush. 1992. “Perceived vs. actual friends’ use of alcohol, cigarettes, marijuana, and cocaine: Which has the most influence?” *Journal of Youth and Adolescence* 21.

Iannotti, Ronald J., Patricia J. Bush, and Kevin P. Weinfurt. 1996. “Perception of friends’ use of alcohol, cigarettes, and marijuana among urban schoolchildren: a longitudinal analysis.” *Addictive Behaviors* 21:615–632.

Jessor, Richard and Shirley L. Jessor. 1978. “Theory testing in longitudinal research on Marihuana use.” In *Longitudinal Research on Drug Use: Empirical Findings and Methodological Issues*, edited by Denise B. Kandel, pp. 41–71. New York, NY: Hemisphere-Halsted Press.

Kandel, Denise and Mark Davies. 1991. “Friendship networks, intimacy, and illicit drug use in young adulthood: a comparison of two competing theories.” *Criminology* 29:441–469.

Kandel, Denise B. 1978a. "Convergences in prospective longitudinal surveys of drug use in normal populations." In *Longitudinal Research on Drug UseEmpirical Findings and Methodological Issues*, edited by Denise B. Kandel, pp. 3–38. New York, NY: Hemisphere-Halsted Press.

Kandel, Denise B. 1978b. "Homophily, selection, and socialization in adolescent friendships." *American Journal of Sociology* 84:427–436.

Kandel, Denise B., G.E. G.E. Kiros, C. Schaffran, and MC. Hu. 2004. "Racial/ethnic differences in cigarette smoking initiation and progression to daily smoking: a multilevel analysis." *American Journal of Public Health* 94:128–135.

Kandel, Denise B. and John A. Logan. 1984. "Patterns of drug use from adolescence to young adulthood: I. Periods of risk for initiation, continued use, and discontinuation." *American Journal of Public Health* 74:660–666.

Kaplan, Howard B., Steven S. Martin, and Cynthia Robbins. 1984. "Pathways to adolescent drug use: self-derogation, peer influence, weakening of social controls, and early substance use." *Journal of Health and Social Behavior* 25:270–289.

Kawaguchi, Daiji. 2004. "Peer effects on substance use among American teenagers." *Journal of Population Economics* 17:351–367.

Killworth, Peter D., Eugene C. Johnsen, H. Russell Bernard, Gene Ann Shelley, and Christopher McCarty. 1990. "Estimating the size of personal networks." *Social Networks* 12:289–312.

Kirke, Deirdre M. 1996. "Collecting Peer Data and Delineating Peer Networks in a Complete Network." *Social Networks* 18:333–346.

Kirke, Deirdre M. 2004a. "Chain reactions in adolescents' cigarette, alcohol and drug use: similarity through peer influence or the patterning of ties in peer networks?" *Social Networks* 26:3–28.

Kirke, Deirdre M. 2004b. "Gender and chain reactions in teenagers' social networks." *Connections* 15:19–27.

Kretzschmar, Mirjam and Lucas G. Wiessing. 1998. "Modeling the spread of HIV in social networks of injecting drug users." *AIDS* 12:801–811.

Krohn, Marvin D, , and Terence P. Thornberry. 1993. "Network theory: A model for understanding drug abuse among African-American and Hispanic youth." In *NIDA Resesarch Monograph 130: Drug Abuse Among Minority Youth: Advances in Research Methodology*, edited by Mario R. De La Rosa and Juan-Luis Recio Adrados. Rockville, MD: U.S. Dept. of Health and Human Services.

Krohn, Marvin D., James L. Massey, and Mary Zielinski. 1988. "Role overlap, network multiplexity, and adolescent deviant behavior." *Social Psychology Quarterly* 51:346–356.

Latkin, Carl A. 1995. "A personal network approach to AIDS prevention: an experimental peer group intervention for street-injecting drug users: the SAFE study." In *NIDA Monograph 151: Social Networks, Drug Abuse, and HIV Transmission*, edited by Richard H. Needle, Susan L. Coyle, Sander G. Genser, and Robert T. Trotter II, pp. 181–195. Rockville, MD: U.S. Dept. of Health and Human Services.

Latkin, Carl A., Wallace Mandel, Maria Oziemkowska, David Celentano, David Vlahov, Margaret Ensminger, and Amy Knowlton. 1994. "Using social network analysis to study patterns of drug use among urban drug users at high risk for HIV/AIDS." *Drug and Alcohol Dependence* 38:1–9.

Lee, Ju-Sung. 2002. "Linking ego-networks using cross-ties." Working paper and conference paper for the 2002 Annual Meeting of the American Sociological Association, Chicago, IL.

Lee, Ju-Sung. 2004. "Generating Networks of Illegal Drug Users Using Large Samples of Partial Ego-Network Data." In *Proceedings from the Second Symposium on Intelligence and Security Informatics*, edited by Hsichun Chen, Reagan Moore, Daniel D. Zeng, and John Leavitt, pp. 390–403. Springer-Verlag.

Loomis, C. P. 1946. "Political and occupational cleavages in a Hanoverian village." *Sociometry* 9:316–333.

Mark, Noah. 1998a. "Beyond individual differences: social differentiation from first principles." *American Sociological Review* 63:309–330.

Mark, Noah. 1998b. "Birds of a feather sing together." *Social Forces* 77:453–485.

Marsden, Peter V. 1987. "Core discussion networks of Americans." *American Sociological Review* 52:122–131.

Marsden, Peter V. 1988. "Homogeneity in confiding relations." *Social Networks* 10:57–76.

McPherson, J. Miller, Pamela A. Popielarz, and Sonja Drobnic. 1992. "Social networks and organizational dynamics." *American Sociological Review* 57:153–170.

McPherson, J. Miller and Lynn Smith-Lovin. 1987. "Homophily in voluntary organizations: status distance and the composition of face-to-face groups." *American Sociological Review* 52:370–379.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. "Birds of a feather: homophily in social networks." *Annual Review of Sociology* 27:415–444.

Michell, Lynn and Amanda Amos. 1997. “Girls, pecking order and smoking.” *Social Science and Medicine* 44:1861–1869.

Mollison, Denis, Valerie Isham, and Bryan Grenfell. 1994. “Epidemics: models and data.” *Journal of the Royal Statistics Society* 157:115–149.

Morris, Martina. 1993. “Epidemiology and social networks: Modeling structured diffusion.” *Sociological Methods and Research* 22:99–126.

Musto, David F. 1987. *The American Disease*. New York, NY: Oxford University Press.

Newcomb, Michael D. and P.M. Bentler. 1988. “Impact of adolescent drug use and social support on problems of young adults: a longitudinal study.” *Journal of Abnormal Psychology* 97:64–75.

Newcomb, Theodore .M. 1961. *The Acquaintance Process*. New York, NY: Holt, Rhinehart, and Winston.

Olds, R. Scott, Dennis L. Thombs, and Jennifer Ray Tomasek. 2005. “Relations between normative beliefs and initiation intentions towards cigarette, alcohol, and marijuana.” *Journal of Adolescent Health* 37:75.e7–75e13.

Pearson, Michael and Lynn Michell. 2000. “Smoke Rings: Social network analysis of friendship groups, smoking and drug-taking.” *Drugs: Education, Prevention and Policy* 7:21–37.

Pearson, Michael, Christian E. Steglich, and Tom Snijders. 2006. “Homophily and assimilation among sport-active adolescent substance users.” *Connections* 27:47–63.

Pearson, Michael and Patrick West. 2003. “Drifting Smoke Rings: Social Network Analysis and Markov Processes in a Longitudinal Study of Friendship Groups and Risk-taking.” *Connections* 25:59–76.

Rapoport, Anatol. 1957. “A contribution to the theory of random and biased nets.” *Bulletin of Mathematical Biophysics* 19:257–271.

Rapoport, Anatol. 1958. “Nets with reciprocity bias.” *Bulletin of Mathematical Biophysics* 20:191–201.

Rapoport, Anatol. 1963. “Mathematical models of social interaction.” In *Handbook of Mathematical Psychology, vol. 2*, edited by Duncan R. Luce, R.R. Bush, and E. Galanter, pp. 493–579. New York, NY: Wiley.

Rapoport, Anatol. 1979. “A probabilistic approach to networks.” *Social Networks* 2:1–18.

Rice, Ronald E., Lewis Donohew, and Richard Clayton. 2003. “Peer Network, Senstation Seeking, and Drug Use among Junior and Senior High School Students.” *Connections* 25:32–58.

Rothenberg, Richard R., Donald E. Woodhouse, John J. Potterat, Stephen Q. Muth, William W. Darrow, and Alden S. Klovdahl. 1995. “Social networks in disease transmission: the Colorado Springs study.” In *NIDA Monograph 151: Social Networks, Drug Abuse, and HIV Transmission*, edited by Richard H. Needle, Susan L. Coyle, Sander G. Genser, and Robert T. Trotter II, pp. 144–180. Rockville, MD: U.S. Dept. of Health and Human Services.

Rubin, Donald B. 1987. *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley.

Rydell, C. Peter, Jonathan P. Caulkins, and Susan S. Everingham. 1996. “Enforcement or treatment? Modeling the relative efficacy of alternatives for controlling cocaine.” *Operations Resesarch* 44:687–695.

Salganik, Matthew J. and Douglas D. Heckathorn. 2004. “Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling.” *Sociological Methodology* 34:193–239.

Sanil, Ashish, David Banks, and Kathleen Carley. 1995. “Models for evolving fixed node networks: model fitting and model testing.” *Social Networks* 17:65–81.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. Boca Raton, FL: Chapman and Hall.

Simons-Morton, Bruce, Denise L. Haynie, Aria D. Crump, Patricia Eitel, and Keith E. Saylor. 2001. “Peer and Parent Influences on Smoking and Drinking among Early Adolescents.” *Health Education and Behavior* 28:95–107.

Simons-Morton, Bruce G. 2002. “Prospective Analysis of Peer and Parent Influences on Smoking Initiation Among Early Adolescents.” *Prevention Science* 3:275–283.

Simons-Morton, Bruce G. 2007. “Social Influences on Adolescent Substance Use.” *American Journal of Health Behavior* 31:672–684.

Simons-Morton, Bruce G. and Rusan S. Chen. 2006. “Over time relationships between early adolescent and peer substance use.” *Addictive Behavior* 31:1211–1223.

Simons-Morton, Bruce G., Rusan S. Chen, L. Abroms, and Denise L. Haynie. 2004. “Latent growth curve analyses of peer and parent influences on smoking progression among early adolescents.” *Health Psychology* 23:612–621.

Skinner, William F., James L. Massey, Marvin D. Krohn, and Ronald M. Lauer. 1985. "Social influences and constraints on the initiation and cessation of adolescent tobacco use." *Journal of Behavioral Medicine* 8:353–376.

Skvoretz, John, Thomas J. Fararo, and Filip Agneesens. 2004. "Advances in biased net theory: definitions, derivations, and estimations." *Social Networks* 26.

Snijders, Tom A.B., Christian E. Steglich, Michael Schweinberger, and Mark Huisman. 2005. *Manual for SIENA version 2.1*. ICS, Department of Sociology, Groningen, NL. http://stat.gamma.rug.nl/snijders/siena.html.

Sparrow, Malcolm K. 1991. "The application of network analysis to criminal intelligence: An assessment of the prospects." *Social Networks* 13:251–274.

Stacy, Alan W., Michael D. Newcomb, and Peter M. Bentler. 1992. "Interactive and high-order effects of social influences on drug use." *Journal of Health and Social Behavior* 33:226–241.

Steglich, Christian E., Tom A.B. Snijders, and Patrick West. 2006a. "Applying SIENA: An illustrative analysis of the co-evolution of adolescents' friendship networks, taste in music, and alcohol consumption." *Methodology* 2:48–56.

Steglich, Christian E., Tom A. B. Snijders, and Pearson West. 2006b. "Applying SIENA: An illustrative analysis of the co-evolution of adolescents friendship networks, taste in music and alcohol consumption." *Journal of Research Methods for the Behavioral and Social Sciences* 2:48–56.

Strogatz, Steven. 2001. "Exploring complex networks." *Nature* 410:268–276.

Tien, Allen. 2001. *Sociometrica LinkAlyzer*. MDLogix, Inc., Baltimore, MD. National Institute on Drug Abuse through a Small Business Innovation Research (SBIR) Phase I project (DA12306: "A Tool for Network Research on HIV Among Drug Users").

Tragler, Gernot, Jonathan P. Caulkins, and Gustav Feichtinger. 2001. "Optimal dynamic allocation of treatment and enforcement in illicit drug control." *Operations Research* 49:352–362.

Trotter II, Robert T., Julie A. Baldwin, and Anne M. Bowen. 1995a. "Network structure and proxy network measures of HIV, drug and incarceration risks for active drug users." *Connections* 18:89–104.

Trotter II, Robert T., Anne M. Bowen, and James M. Potter, Jr. 1995b. "Network models for HIV outreach and prevention programs for drug users." In *NIDA Monograph 151: Social Networks, Drug Abuse, and HIV Transmission*, edited by Richard H. Needle, Susan L. Coyle, Sander G. Genser, and Robert T. Trotter II, pp. 144–180. Rockville, MD: U.S. Dept. of Health and Human Services.

Trotter II, Robert T., Richard B. Rothenberg, and Susan Coyle. 1995c. “Drug abuse and HIV prevention research: expanding paradigms and network contributions to risk reduction.” *Connections* 18:29–45.

Urberg, Kathryn A., Serdar M. Degirmencioglu, and Colleen Pilgrim. 1997. “Close friend and group influence on adolescent cigarette smoking and alcohol use.” *Developmental Psychology* 33:834–844.

Verbrugge, Lois M. 1977. “The Structure of Adult Friendship Choices.” *Social Forces* 56:576–597.

Wasserman, Stanley and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge, UK: Cambridge University Press.

Watts, Duncan J. 1999a. “Networks, dynamics, and the small-world phenomenon.” *American Journal of Sociology* 105:493–527.

Watts, Duncan J. 1999b. *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton, NJ: Princeton University Press.

Watts, Duncan J. and Steven Strogatz. 1998. “Collective dynamics of ’small-world’ networks.” *Nature* 393:440–442.

Wellman, Beth. 1929. “The school childs choice of companions.” *Journal of Educucation Research* 14:126–132.

Zeggelink, Evelien. 1994. “Dynamics of structure: An individual-oriented approach.” *Social Networks* 16:295–333.

Zeggelink, Evelien. 1995. “Evolving friendship networks: An individual-oriented approach implementing similarity.” *Social Networks* 17:83–110.