Dynamics: Reasoning About Networks Over Time

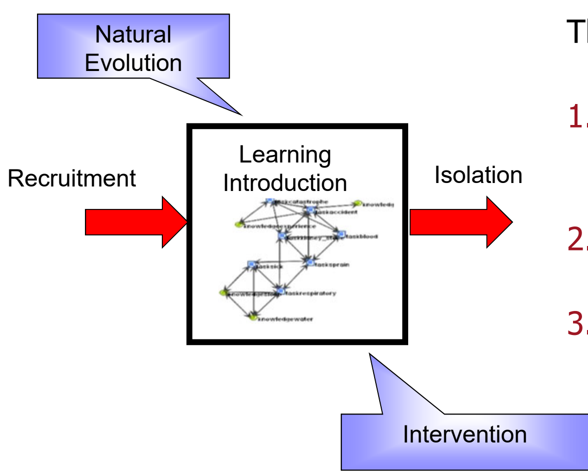
Prof. L. Richard Carley

carley@ece.cmu.edu

Carnegie Mellon

Center for Computational Analysis of Social and Organizational Systems
http://www.casos.cs.cmu.edu/

Changes in Networks Over Time



Three Ways to Study Dynamics

1. Comparison over time
   - Look at real data
2. Immediate Impact
   - Comparative statics
3. Near Term Impact
   - Utilizes simulation - DyNet

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Basic Issue

- Over time the set of nodes change
- What should you do?
  - Compare just nodes present in all time periods
    - For core group – how has it changed
  - Create a master network of all nodes
    - How has the flux altered the groups
  - Use whatever nodes are available
    - What are the natural dynamics
- No single right answer
  - It depends on what you want to know
  - It depends on how your underlying network changes with time
  - Often try two different approaches and see how much they differ

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Types of Over-Time Change in Networks

- Stability
  - Relationships remain the same over time but there are statistical fluctuations in who talks to who when
- Evolution
  - Interaction among agents cause relationships to change over time
- Shock
  - Change is exogenous to the social group.
- Mutation
  - A shock stimulates evolutionary of the social structure in respons.

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Models Used to Study Change

- Different models used to study different types of Change
  - Stability: LPM , ERGM, repeated measures
    focus is on modeling statistical fluctuations in interactions
  - Evolution: SIENA, multi-agent simulation, or both
    focus is on modeling systematic shift in networks structure
  - Shock: Change detection in real-world applications
    & Multi-agent simulation for experimentation (what if studies)
  - Mutation: Change detection coupled with evolutionary model
    for real world applications

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Social Networks are Continuously Emerging Structures

- Networks emerge from intersecting trails
  - Constrained and enabled
- Networks reinforce some trails
  - Secondary emphasis to some constraints
- Slices across trails are the "measured" or "observed" social network
- The level of aggregation determines the "width" of the slice
  - The greater the width – the higher the density of connections

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Aspects of Trails of Interest

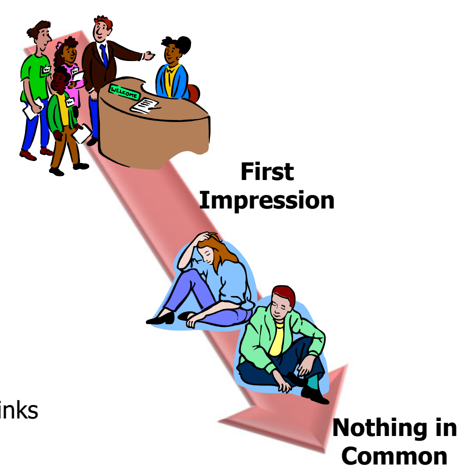
- PLACE – Physical
  - Who was where when
  - doing what (how (to/with whom (why)))
- EXPERTISE/KNOWLEDGE – Cognitive
  - Who was providing what information when
  - how (to whom (from where (why)))
- ACTIVITY – Occupation
  - Who was doing what when
  - how (with whom (where (why)))

Trails Provide Meta-Network Information

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Social Dynamics due to Learning

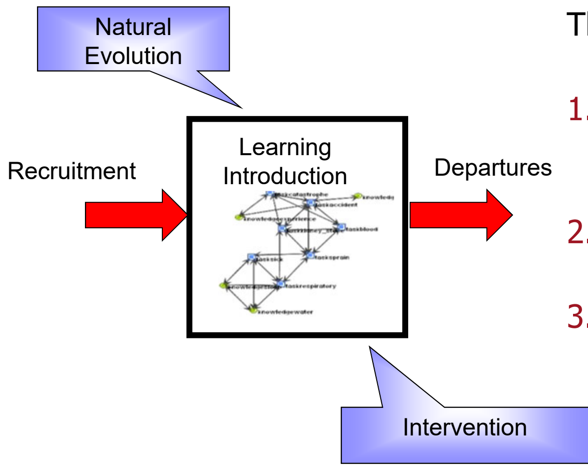
- Implicit link
  - seen together
  - common sources
  - seniority
- Explicit link
  - information exchange
  - learned from each other
  - mentoring



- When meeting a new person
  - Infer expertise based on implicit links
  - Baseline for trust
  - Social shakeout occurs as you move from implicit to explicit links

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Changes in Networks over Time



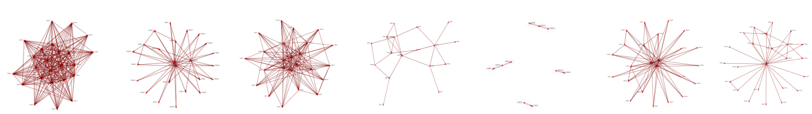
Three Ways to Study Dynamics

1. Comparison over past times
   – Look at real data
2. Immediate Impact
   – Comparative statics
3. Near Term Impact
   – Utilizes simulation - DyNet

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Longitudinal (Over Time) Networks

- Consider watching communications on a network, such as email. Mark a link between agents if communicated.



- Has this organization changed significantly?
- Has it evolved?
- Have people changed their position in the network?

One Issue: the node set

- Over time the set of nodes change
- What should you do?
  - Compare just nodes present in all time periods
    - For core group – how has it changed
  - Create a master network of all nodes
    - How has the flux altered the groups
  - Use whatever nodes are available
    - What are the natural dynamics
  - Note – choice changes many measures that are scaled by size
- No single right answer
  - Right answer depends on what you want to know
  - Often try two different approaches and see how much they differ

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Taxonomy of Change in Network Data

- Stability: Relationships remain the same over time.
  - But will still have significant "random" variations with time
- Evolution: Interaction among agents cause the relationships to change over time.
  - Normal state of affairs with humans beings as agents
  - Still has "random" variations as well
- Shock: Change is exogenous to the social group.
  - This is crucial for many real world applications
- Mutation: A shock stimulates evolutionary behavior.
  - This is longer term response of organization to changing environment

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Models Used to Classify "Change"

- Stability: LPM , ERGM, repeated measures
  - LPM is Link Probability Model
  - ERGMs are Exponential Random Graph Models
- Evolution: SIENA, multi-agent simulation (CONSTRUCT), or both
- Shock: Change detection in real-world applications
  Multi-agent simulation for experimentation
- Mutation: Change detection coupled with SIENA for real world applications
  Multi-agent simulation for experimentation

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Dynamic Analysis Techniques

- Visualization
- Comparative Statics – Immediate Impact
- Longitudinal Networks and Change
  - Stability, Evolution, Shock, Mutation
- QAP (Quadratic Assignment Procedure)
  and MRQAP (Multiple Regression QAP), Longitudinal QAP
- Statistical Models of Networks
  - Link Probability Model (LPM) for Stability
  - Actor-Oriented Models for Evolution
  - Multi-Agent Simulation for Evolution, Shock, and Mutation
  - Exponential Random Graph Models
- SIENA
- Statistical Process Control
- Network Change Detection
- Fourier Analysis
- Simulation (Agent-Based Dynamic Network)

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Communications as a Proxy

- "Ideal approach" – directly sample network each time period
  - E.g., have every member of society fill out survey every time period
  - Limited to very small societies and really motivated subjects
- Or, tracking changes over time using communications data
  - Communication is "proxy" for a network tie
  - Tracking large amounts of communication data gives approximate picture of the underlying social network structure
  - Can use it to find Key Entities and other Network measures
- Communication log data available from many sources
  - Cell Phone Service Providers – call logs, txt msg logs
  - E-mail Data logs – available within organizations
  - Software: Twitter, Facebook, FourSquare, etc.
  - Hardware: building sensors, cell phone sensors, RFID Tags, GPS, etc.

CASOS

June 2020

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Communications Log Data

- Data on who you talk to over monitored means, but NOT what you say (decreased privacy concerns relative to full text monitoring)
- Researchers often only have access to logs from 1 or 2 communications channels – not all possible channels
  - Missing data is substantial
- Communication event is taken as a proxy for a link
  - But this may not always be the case; e.g., calling a wrong #

CASOS

June 2020

Carnegie Mellon

isr institute for SOFTWARE RESEARCH

Is Com Log Data a good Proxy ?

- Example: 2011-2013 NetSense Data Set from Notre Dame
  - Aaron Striegel, Shu Liu, Lei Meng, Christian Poellabauer, David Hachen, Omar Lizardo, "Lessons Learned from the NetSense Smartphone Study," Proceedings of HotPlanet'13, August 16, 2013, Hong Kong, China.
- They recruited 180+ incoming freshmen/freshwomen in 4 dorms to join study
  - Students received free cell phone (including phone plan)
  - Students had to agree to use provided Android cell phone as their primary cell phone
  - Students agreed to having calls and txt msgs logged
  - Students agreed to filling out monthly surveys
- Data collected from study for 4 academic semesters
  - Data from Summer survey too unreliable to use because many students were away from campus for summer

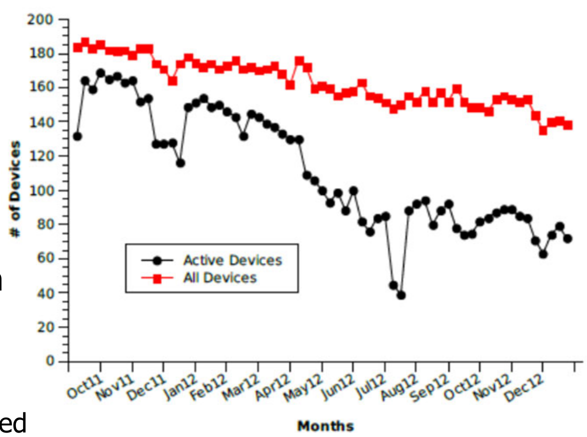
CASOS

Carnegie Mellon

isr institute for SOFTWARE RESEARCH

NetSense Details

- NetSense study surveyed all participants monthly + an extra long survey at end of each semester
  - Survey return rate nearly 100%
  - This work focused on the long survey at end each sem.
  - Long Survey Question asked top 10 people you interact with
- NetSense population changes over time
  - Students either quit or violate terms of study and are removed



CASOS

Carnegie Mellon isr institute for SOFTWARE RESEARCH

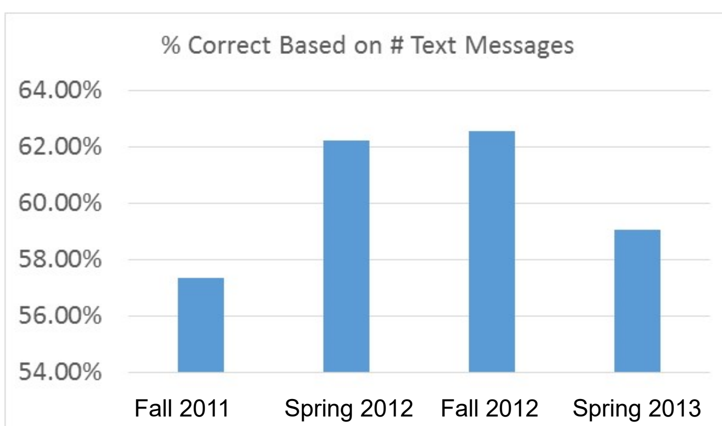
Methodology

- Question to be studied:
  - Accuracy of phone logs relative to survey for predicting network
- Survey
  - Asked students to list top 10 people they interact with regularly
  - Students didn't have to fill in all 10 slots
  - May of those listed were people outside of study (e.g., parents)
  - Keeping only those in study gave list of 0-10 others in the study that the surveyed individual considered strong interaction targets
- Cell Phone Data
  - Looked at # txt msgs, # txt chars, # phone calls, # secs on calls
  - Ranked in-study interactors based on these metrics
- Predictor Quality
  - Probability individual listed as one of N in-study individuals in survey is in the top N based on cell phone data

Carnegie Mellon isr institute for SOFTWARE RESEARCH

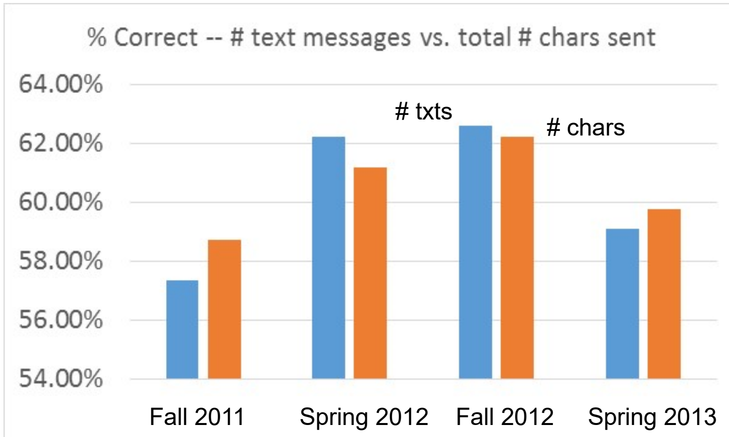
Text Messages

- This generation lives on text messages
  - Overall, # txt msgs accurage about 60% of the time



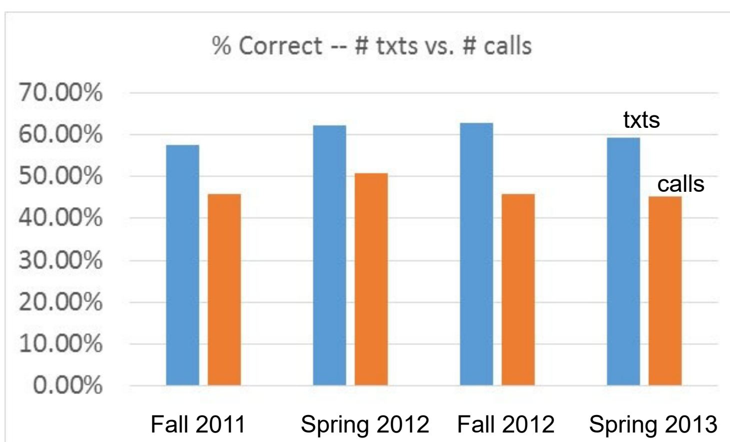
Chars in Text Messages

- Length of text message does not seem significant
  - Using # chars instead of # txt msgs is actually slightly worse



Phone Calls

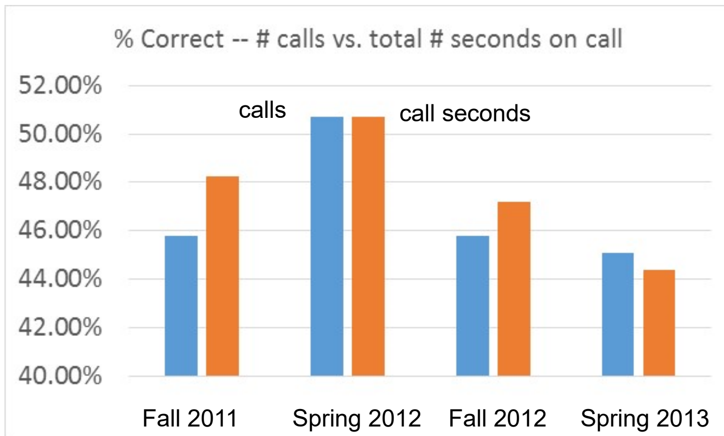
- How about using # phone calls instead of # txt msgs
  - Not good at all – only about 45% on average for phone calls



Carnegie Mellon

Call Seconds vs. # Calls

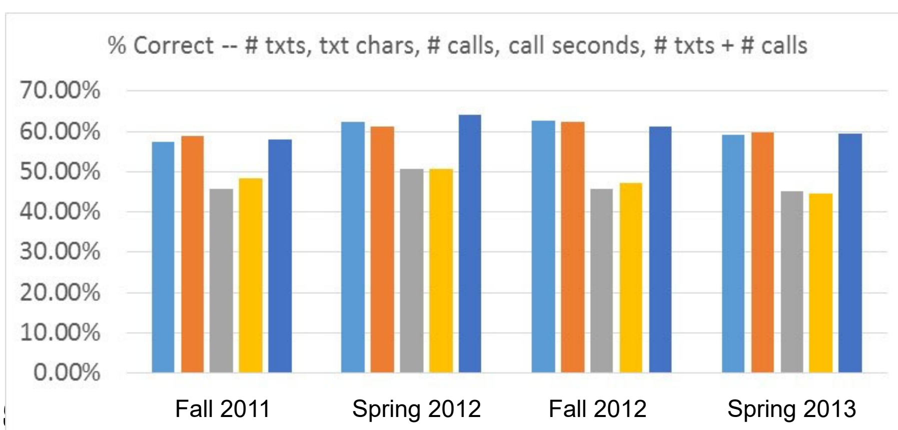
- How about using # call seconds instead of # txt msgs ?
  - A little better – about 48% on average for calls seconds



Carnegie Mellon

Composite Measures

- Comparing all to (#txts + #calls) as measure ?
  - This combination achieves highest score, above 60% accurate



Conclusions: Com Logs can be OK Proxy for Network Ties

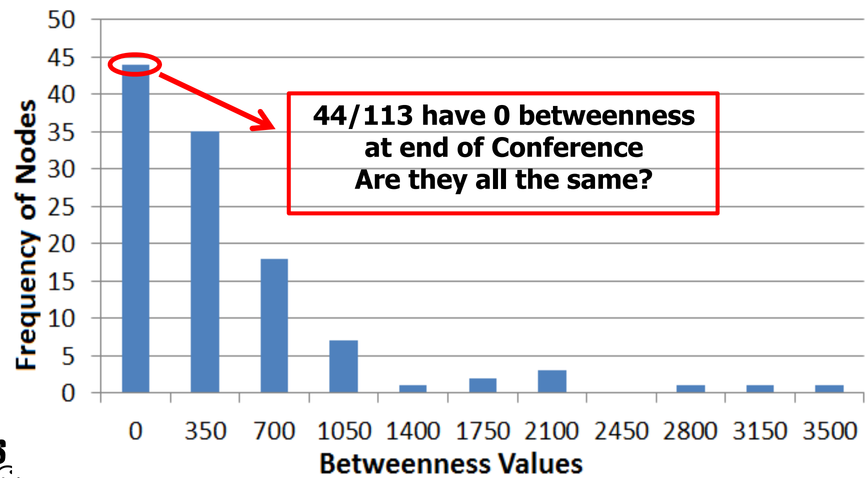
- # txt msg is good proxy for interaction propensity for this cohort
- Combinations of comm data metrics can slightly increase accuracy, but only a little
- Accuracy level of about 60% indicates that many interactions are mediated by other communications channels (e.g., face-to-face).
- Results of this analysis may vary widely for different communities – in 2011, freshmen/freshwomen are highly attached to txt msgs for communication
- Note, self-reporting errors may influence these results – e.g., participants took final survey less seriously

The Challenge of Temporal Evolving Social Networks

- Consider ACM Hypertext 2009 Conference
  - Badges with RFIDs
- Close Range Face-to-Face Contact
  - 1 - 1.5 meters of one another
  - Human body acts as an RF shield
- Collect sensor data every 20 seconds for 2.5 days
  - 20,818 real time data updates
  - 113 participants, 2196 undirected, weighted links

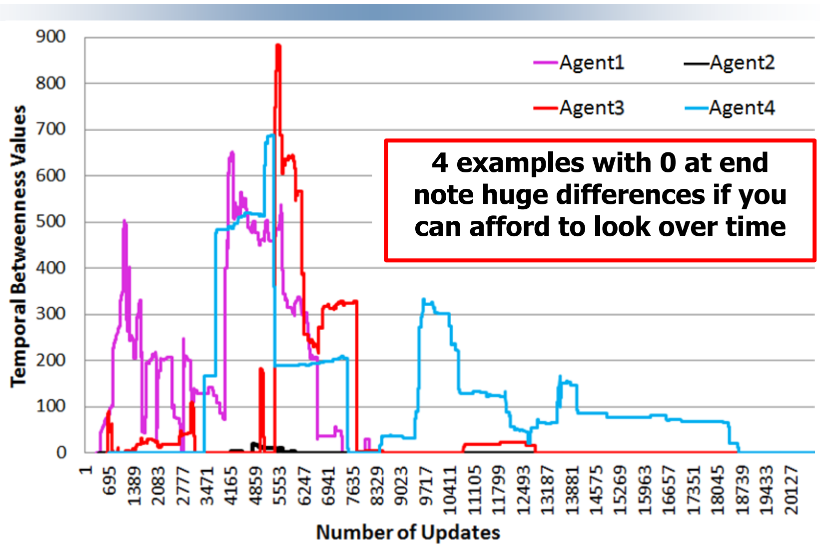
Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Socio-Patterns: Betweenness Centrality Distribution



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Socio-Patterns: Betweenness Over-Time Trends



Carnegie Mellon
institute for SOFTWARE RESEARCH

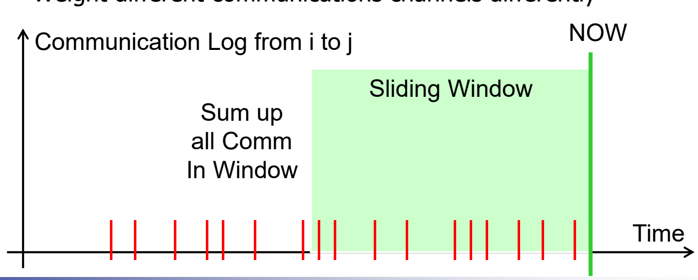
Critical Issue: Slicing and Dicing

- Approach 1: Cumulative network
  - Each time period is all prior links plus new
  - Good for data where links don't go away – e.g., citation networks
- Approach 2: Divide based on external shock
  - Number of time windows depends on external events e.g., before and after a referendum
  - Good for data where there is a major known change
- Approach 3: Divide into uniform periods
  - Number of time windows depends on collection and time slice
  - Good for large data and for doing periodicity studies
- Approach 4: Streaming
  - Only show most recent data using some moving average
  - Good when data too large to be stored – least developed

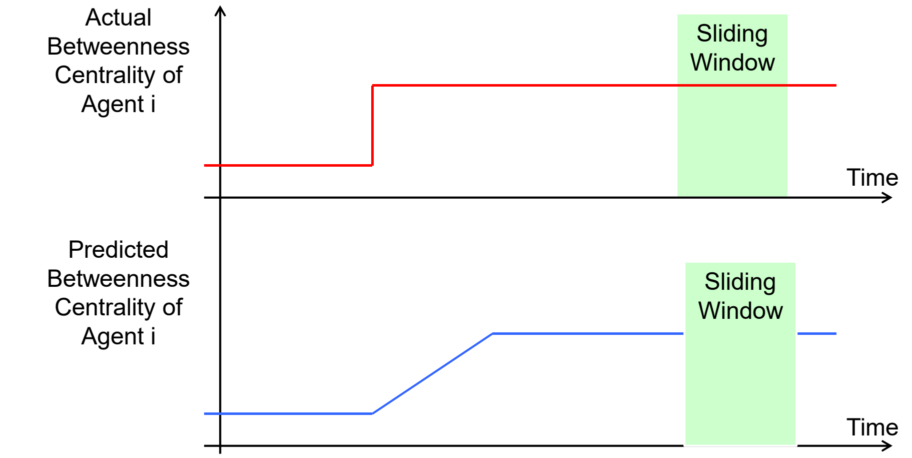
Carnegie Mellon
institute for SOFTWARE RESEARCH

Sliding Window for Over-Time Links

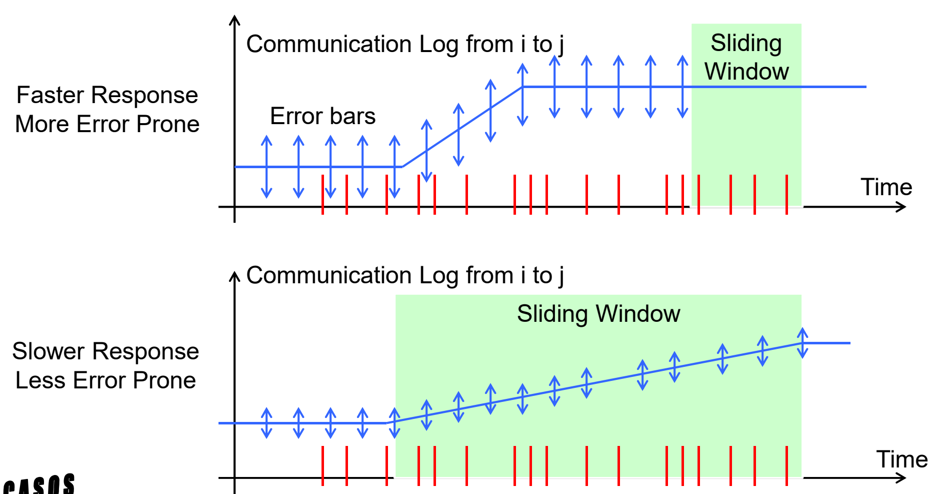
- Estimator for Link Weight (a.k.a. Link Cost)
  - Add up # of Communication Events between x & y in window
  - Take reciprocal. If # is 0, there is no Link between that pair
  - Then move window forward by a time step and repeat
  - Alternatives possible:
    - Incorporate duration of communication
    - Weight different communications channels differently



Smoothing Effect of Sliding Window

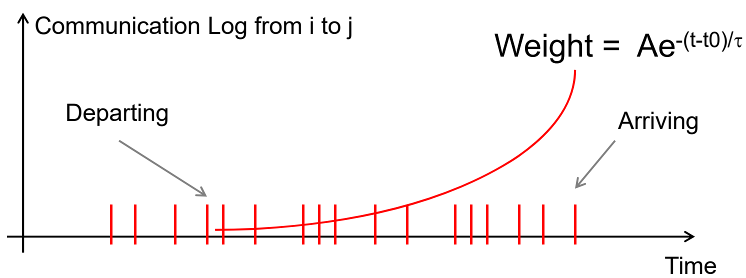


Adjusting Window Size



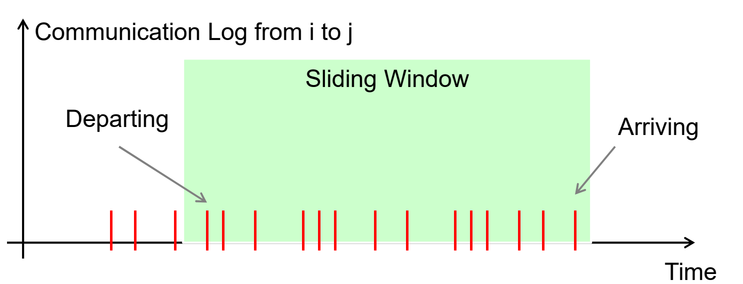
Mathematically Better Window

- Improved tradeoff between smoothing and averaging
  - Mathematically, Exponentially Weighted Moving Average (EWMA)
    - Considers all past known events in estimating current network
    - Old events receive smaller and smaller weighting
    - New events receive highest weighting
    - Exponential time constant – τ – sets how quickly past attenuates vs. how much averaging reduces variance of network



Incremental Sliding Window

- Sliding Window is Synergistic with Incremental Analysis
  - As window moves forward in time
    - New events "arrive" and must be processed
    - Old events "fall out" of trailing edge of window and must be processed
    - BUT – all of the data in middle of window remains unchanged
    - Incremental algorithms fast because only small part of data changes



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Change Detection

- Goal: Rapidly detect that a change has occurred
- Detect *shocks*, not evolutionary changes
  - Evolutionary change: change due to interaction among actors in a network
    - Example: change of interaction patterns over time among new students as they get to know each other
  - Shock: change reason is exogenous to the network
    - Example: change of interaction patterns among students after they graduate
  - Another way to say it: detect "fast" change not "slow" change
- Another goal is to identify *change point*
  - Likely time when change occurred
  - Limits the scope of explanation for network change

CASOS

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Theory of Change and Network Evolution – is it "Change"?

- When is observation statistically different from normal fluctuations ?
- Need a theory for how links fluctuate over time – Null Hypothesis
- Random – assume Network links appear at random
- Heiderian balance
- Blau exchange
- Socio-Cognitive needs
  - Homophilly
  - Expertise
  - work
- The Rich get Richer
  - Popularity
    - Most likely link is to nodes that others link to
  - Preferential attachment
    - Variation on the theme
  - Limits to growth/interest
    - Link to those not over-committed

CASOS

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Common Attachment Biases

- Interaction Logics/Biases – going beyond random
  - Homophilly
    - Relative similarity
  - Relative expertise
  - Need to work
  - Need to coordinate
  - Activity
  - Node intelligence
  - Preferential attachment
  - Distance decay
- Often referred to as generative Grammars

CASOS

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Random

- Network ties are random
- Each time period just generate a random network of a particular size and density
  - Size and density may grow or shrink via other models
- The “naïve” baseline

CASOS

CASOS

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Classic Random Graph Models

- In the **G(n,p)** random graph model:
  1. There are ***n*** nodes.
  2. There is an edge between any two nodes with probability ***p***.

Proposed by Erdös and Renyi in 1960s.

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Properties of Online G(n,p)

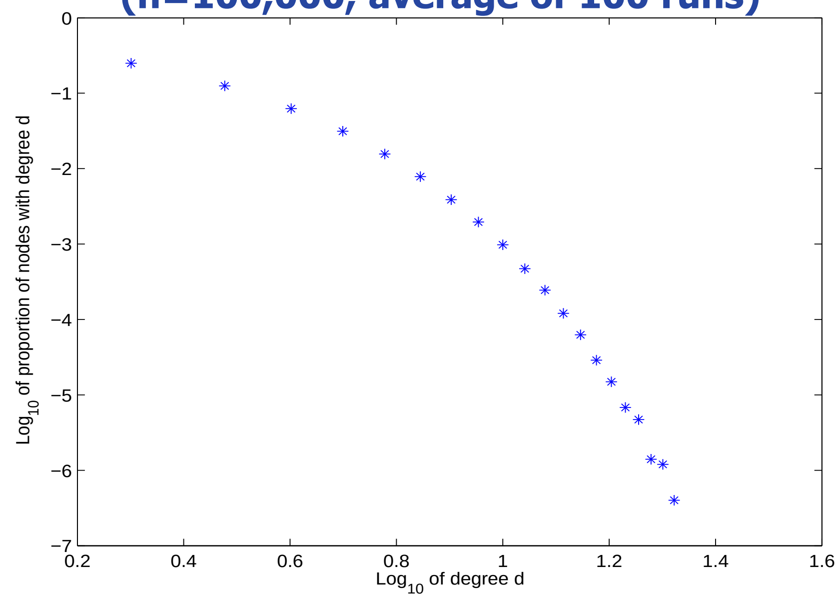
- E[degree of first node] = $1 + 1/2 + 1/3 + 1/4 + \ldots 1/n$ = $\Theta(\log n)$

- E[max degree] = $\Theta(\log n)$

- X_k = Proportion of nodes with degree k
  $E[X_k] = \Theta(½^k)$

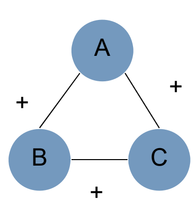
This does NOT generate a POWER LAW

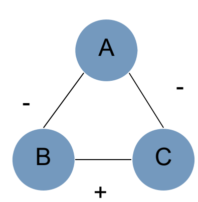
Online G(n,p)
(n=100,000, average of 100 runs)



Heiderian Balance

- Instead of 0/1 Links, let us allow -1 / 0 / 1 links
- Actors are only comfortable in balanced relations
- Balance is achieved when there are three positive links or two negatives with one positive.
- Two positive links and one negative creates imbalance.

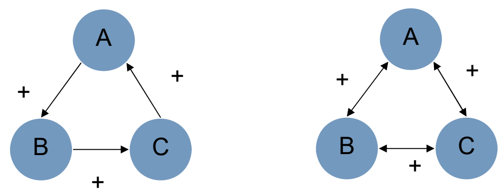




Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Blau Exchange

- Exchange strengthen ties
- Tendency to reciprocity
- Reciprocity is strongest when in triadic relations



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Theory Based Inference: Meta Network

- Homophilly
  - Knowledge
  - Resources
  - Attributes
  - Etc
- Two mode networks needed:
  - Such as People by expertise or People by resources
- Operationalized as
  - Similarity
  - Relative similarity
  - Similarity on shared and non shared characteristics
  - Relative similarity on shared and non shared characteristics

Carnegie Mellon isri institute for SOFTWARE RESEARCH

Interaction Style: Need for Communicative Ease - Homophily

- Relative similarity = degree of shared knowledge i & j relative to i's shared knowledge with all others
- AKik is knowledge network
  - Knowledge network is agent by knowledge ("facts")
- Homophily → interaction proportional to relative similarity

I = max number of agents
K = max number of ideas, facts, pieces of knowledge

$$RS_{ij} = \frac{\sum_{k=0}^{K} (AK_{ik} * AK_{jk})}{\sum_{j=0}^{I} \sum_{k=0}^{K} (AK_{ik} * AK_{jk})}$$

$$\text{Cutoff} = \sum_{i=0}^{I} RS_{ij} \; / \; (I * (I - 1))$$

If $RS_{ij} \geq$ Cutoff the Expected interaction = 1
else 0

Carnegie Mellon isri institute for SOFTWARE RESEARCH

Relative Similarity – Why ?

- Similarity: individuals tend to interact with those whom they perceive to be more similar to themselves
  - Comfort
  - Ease of interaction
  - Ease of access
  - Common language
  - More effective for getting information
  - Shared expectations about reciprocity
- Relative: individuals judge similarity relative to others
  - There is a comparison group
  - There is a generalized other

Carnegie Mellon

Interaction Style: Need to Know Relative Expertise

- Relative expertise = how much i thinks j knows that i does not know divided by how much i thinks all others know that i does not know
- AKik is knowledge network
- Expected interaction based on relative expertise

If $AK_{ik} = 0$ then $X_{jk} = AK_{jk}$
Else $X_{jk} = 0$

$$RE_{ij} = \frac{\sum_{k=0}^{K} (X_{jk})}{\sum_{j=0}^{I} \sum_{k=0}^{K} (X_{jk})}$$

$$\text{Cutoff} = \sum_{j=0}^{I} RE_{ij} \; / \; (I * (I - 1))$$

If $RE_i \geq$ Cutoff the Expected interaction = 1
else 0

I = max number of people
K = max number of ideas

CASOS

Carnegie Mellon

Relative Expertise - Why

- Expertise: individuals tend to interact with those whom they believe to have information that they need
  - Information gathering
  - Desire to achieve
  - Desire for increase in power
  - Information as power
- Relative: individuals judge expertise relative to others
  - There is a comparison group
  - There is a generalized other

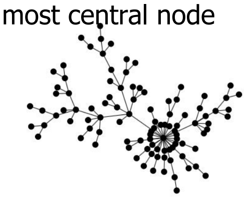
CASOS

CASOS

Carnegie Mellon
ISRI institute for SOFTWARE RESEARCH

The Rich Get Richer Centrality Increases Models

- Popularity – rich get richer
  - As size goes up new nodes link to most central node
  - On average
- Preferential Attachment (Yule or Matthew effect)
  - New nodes are connected to old according to the number of others already connected
  - Can generate power laws
- Limits to Growth
  - As size goes up new nodes are added to the most central node that has not hit its limit
  - On average



Carnegie Mellon
ISRI institute for SOFTWARE RESEARCH

Preferential Attachment

- In the Preferential Attachment model, each new node connects to the existing nodes with a probability proportional to their degree.

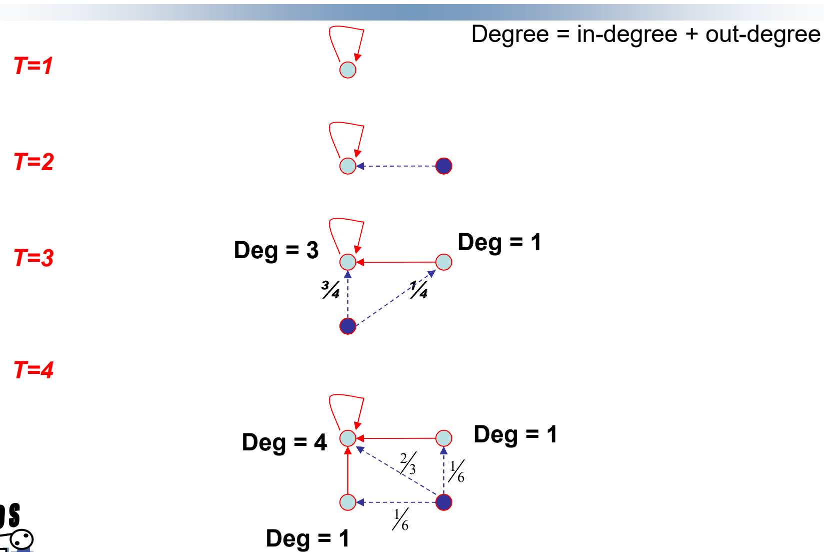
(1) *Growth*: Starting with a small number (m_0) of nodes, at every timestep we add a new node with $m(\leq m_0)$ edges that link the new node to m different nodes already present in the system.

(2) *Preferential attachment*: When choosing the nodes to which the new node connects, we assume that the probability Π that a new node will be connected to node i depends on the degree k_i of node i, such that

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}$$

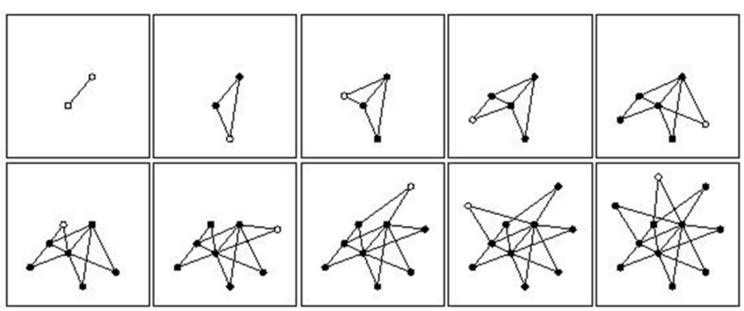
CASOS

Preferential Attachment



Preferential Attachment

- Linear preferential attachment processes in which the number of nodes increases are known to produce a distribution of node centralities following the so-called Yule distribution.
- The fraction of nodes having k links in the limit $P(k) = \frac{\mathrm{B}(k+a,\gamma)}{\mathrm{B}(k_0+a,\gamma-1)},$



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Measuring preferential attachment

- Is it the case that the rich get richer?
- Look at the network for an interval [t,t+dt]
- For node i, present at time t, we compute

$$D_i = \frac{dk_i}{dk}$$

  - dk_i = increase in the degree
  - dk = number of edges added
- Fraction of edges added to nodes of degree k

$$f(k) = \sum_{i:k_i=k} D_i$$

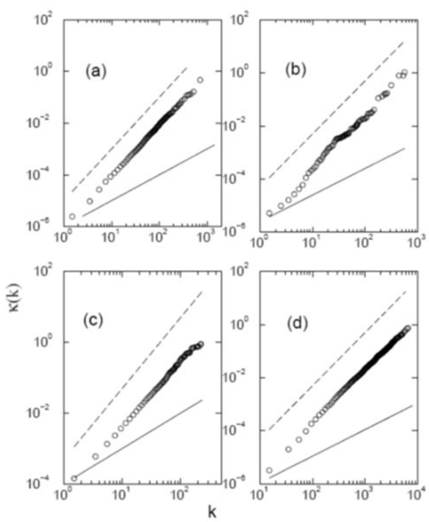
- Cumulative: fraction of edges added to nodes of degree at most k

$$F(k) = \sum_{j=1}^{k} f(j)$$

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Measuring preferential attachment

- plot F(k) as a function of k



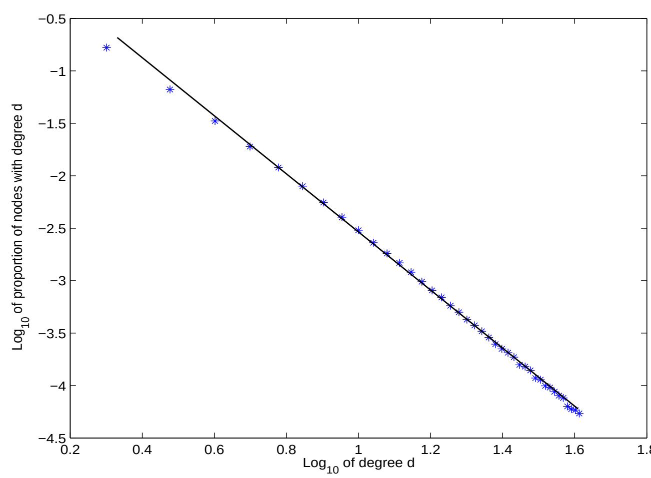
(a) citation network
(b) Internet
(c) scientific collaboration network
(d) actor collaboration network

Preferential Attachment

E[degree of 1st node] = √n

Preferential Attachment gives a **power-law** degree distribution. [**Mitzenmacher, Cooper & Frieze 03, KRRSTU00**]

Preferential Attachment



Carnegie Mellon

Network models and temporal evolution

- For most of the existing models it is assumed that
  - number of edges grows linearly with the number of nodes
  - the diameter grows at rate logn, or loglogn
- What about real graphs?
  - Leskovec, Kleinberg, Faloutsos 2005

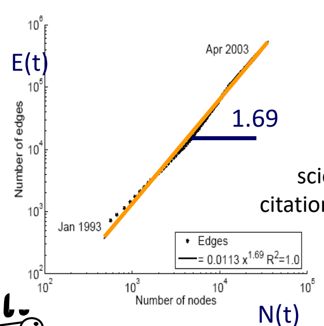
Carnegie Mellon

Densification laws

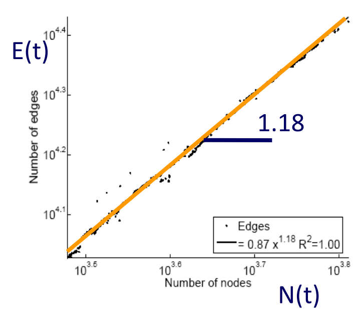
- In real-life networks the average degree increases! – networks become denser!

$$\boxed{E(t) \propto N(t)^{a}}$$

α = densification exponent



scientific citation network

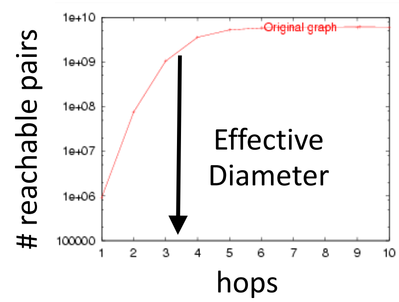


Internet

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

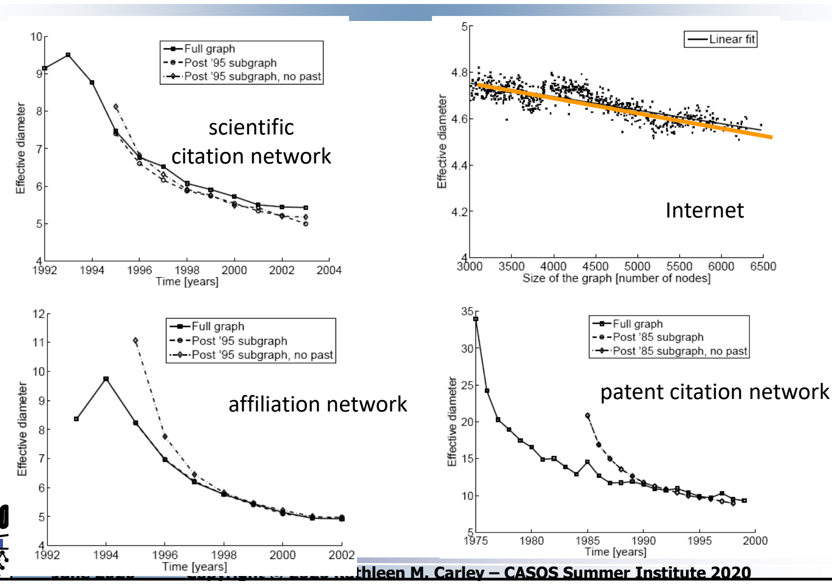
What about diameter?

- Effective diameter: the interpolated value where 90% of node pairs are reachable



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

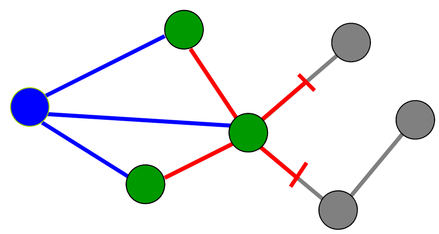
Diameter shrinks



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

"Forest Fire" Model

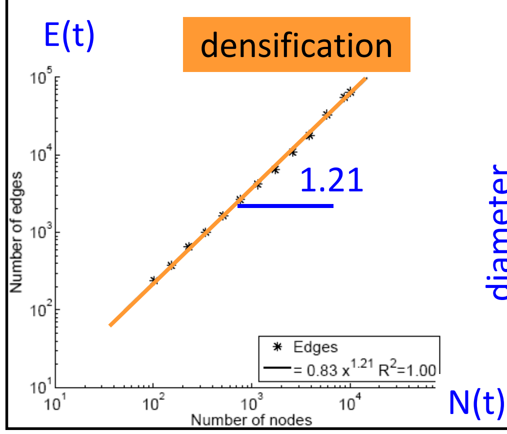
- A node arrives
- Randomly chooses an "ambassador"
- Starts burning nodes (with probability p) and adds links to burned nodes
- "Fire" spreads recursively

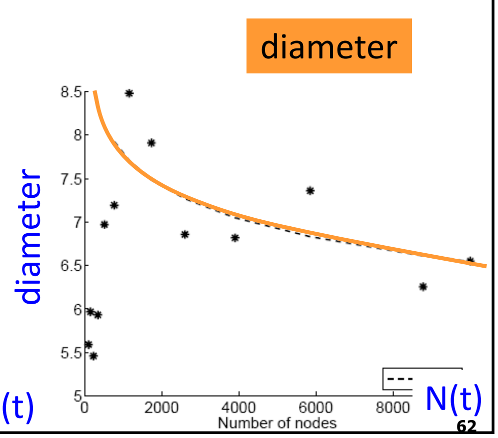


Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Forest Fire in Action (1)

- Forest Fire generates graphs that Densify and have Shrinking Diameter

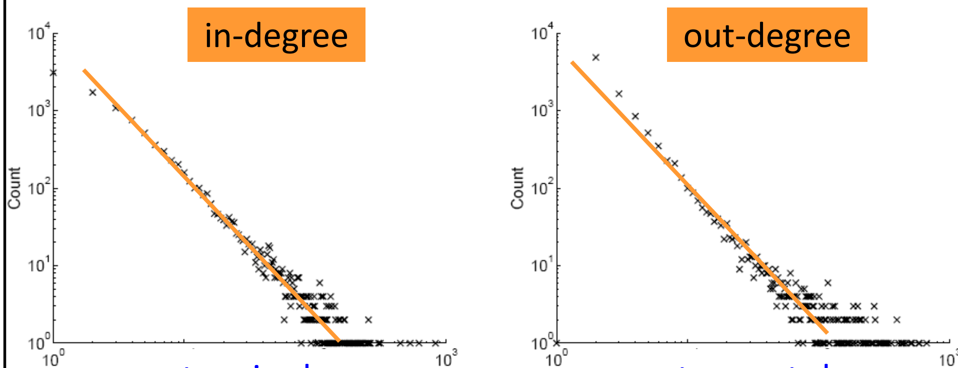




Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Forest Fire in Action (2)

- Forest Fire also generates graphs with heavy-tailed degree distribution



count vs. in-degree count vs. out-degree

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Forest Fire model – Justification

- Densification Power Law:
  - Similar to Community Guided Attachment
  - The probability of linking decays exponentially with the distance – Densification Power Law
- Power law out-degrees:
  - From time to time we get large fires
- Power law in-degrees:
  - The fire is more likely to reach hubs
- Communities:
  - Newcomer copies neighbors’ links
- Shrinking diameter

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Networks Heal Themselves

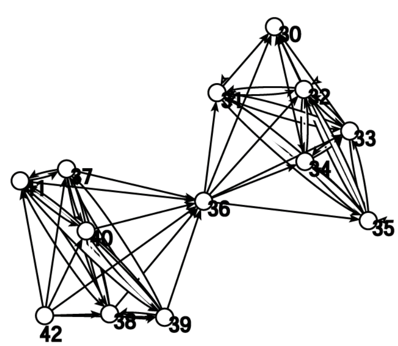
- The rules for where networks will add ties are actually even more complex than any of the above
- Networks can add ties intentially
- Networks, particularly cellular networks, can withstand high levels of turnover
- Agents the are in structurally "equivalent positions" are replaceable by others that are "equivalent"
  - Connected to same others
- Agents in specialized positions, e.g., those with high cognitive load, are harder to replace
- Newcomers typically enter as neither structurally equivalent with a key actor nor high in cognitive load
  - Low transactive memory
  - Few pre-existing ties
  - "start off on simple tasks"

CASOS

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Network Evolution

- Causes
  - Learning
  - Physical movement
  - New resources
  - Attrition
  - Removal
- Interaction Logics
  - Homophily
  - Information seeking
  - Co-work
  - Co-location



1. Original Cell Configuration

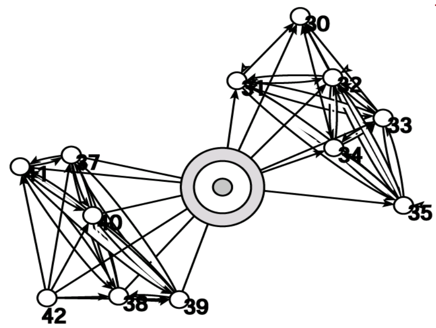
CASOS

CASOS

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Network Evolution

- Causes
  - Learning
  - Physical movement
  - New resources
  - Attrition
  - Removal
- Interaction Logics
  - Homophily
  - Information seeking
  - Co-work
  - Co-location

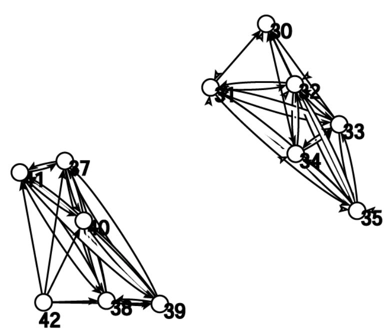


2. Node 36 Targeted for Isolation

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Network Evolution

- Causes
  - Learning
  - Physical movement
  - New resources
  - Attrition
  - Removal
- Interaction Logics
  - Homophily
  - Information seeking
  - Co-work
  - Co-location

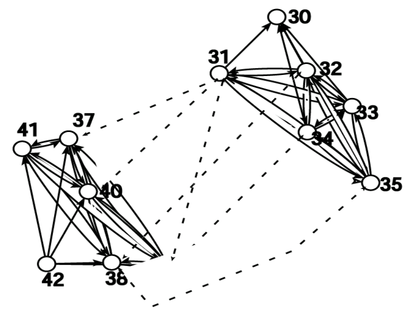


3.Two Cells are Disconnected

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Network Evolution

- Causes
  - Learning
  - Physical movement
  - New resources
  - Attrition
  - Removal
- Interaction Logics
  - Homophily
  - Information seeking
  - Co-work
  - Co-location

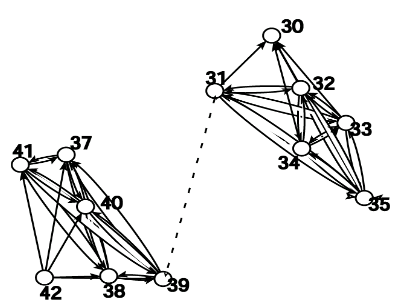


4. Agents use referential data to attempt to reconnect cells

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Network Evolution

- Causes
  - Learning
  - Physical movement
  - New resources
  - Attrition
  - Removal
- Interaction Logics
  - Homophily
  - Information seeking
  - Co-work
  - Co-location

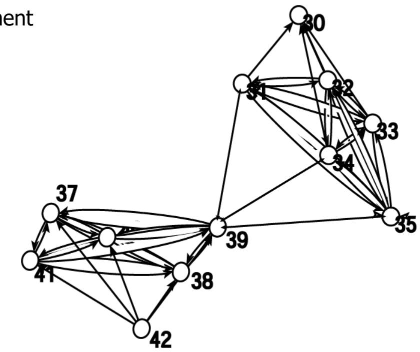


5. One of the connections succeeds
Others are refused

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Network Evolution

- Causes
  - Learning
  - Physical movement
  - New resources
  - Attrition
  - Removal
- Interaction Logics
  - Homophily
  - Information seeking
  - Co-work
  - Co-location



6. New Cell Leader Has Emerged

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Be Careful What Network Models

- Many Network analyses applies to “flow” through edges
- Things that can flow
  - Data
  - Ideas or Beliefs
  - Money or Resources
  - New Technology
  - Disease
  - Current / Power / water
- Each has different flow properties because
  - Retention
  - Acceptance
  - Resistance

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Propagation models

- Epidemics
  - How do epidemic diseases propagate through society?
  - One of the major reasons that people started studying social networks in the community
- Consumer's society
  - How do products propagate and innovations get accepted ?
  - Early reason for studying online social networks
- Fads
  - How do ideas and beliefs diffuse?
  - One of the major reasons that people started studying social networks in the workplace

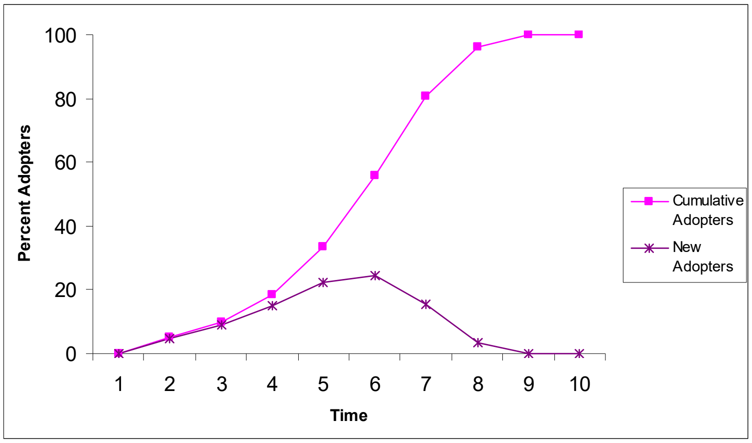
Carnegie Mellon
isr institute for SOFTWARE RESEARCH

ELEMENTS OF THE DIFFUSION OF INNOVATIONS

1. The rate of diffusion is influenced by the perceived characteristics of the innovation such as relative advantage, compatibility, observability, triability and complexity, radicalness, and cost.
2. Diffusion occurs over time such that the rate of adoption often yields a cumulative adoption S-shaped pattern.
3. Individuals can be classified as early or late adopters.
4. Individuals pass through stages during the adoption process typically classified as (1) knowledge, (2) persuasion, (3) decision, (4) implementation or trial, and (5) confirmation.

Carnegie Mellon

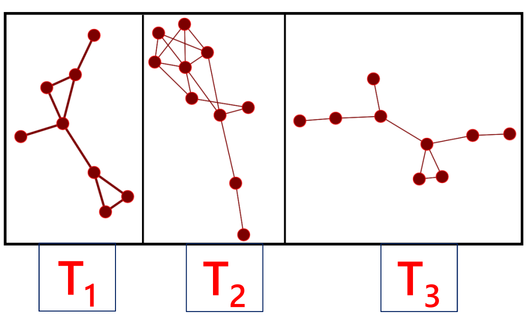
Hypothetical Cumulative and Incidence Adoption Curves for Diffusion Homogenous Mixing



Carnegie Mellon

Dynamic Metrics on Over-Time Data

- *Identifying central nodes in a network*



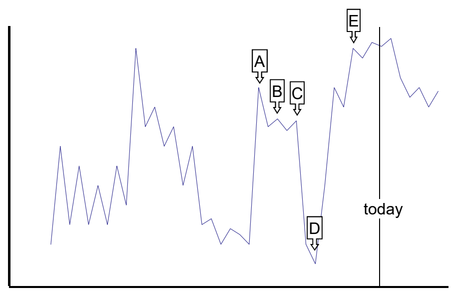
Dynamically Changing Network Structure!!!

Focus on calculating measures for each network
Then evaluate how measures change over time

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Changes in Network Data Measures

- Various measures of a network are calculated for a window of network data at a multiple points in time
- Change detection: quickly determine *that* a change occurs.
- Change point identification: *when* did the change occur.



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Change Detection

- Goal: Rapidly detect that a change has occurred
- Detect *shocks*, not evolutionary changes
  - Evolutionary change: change due to interaction among actors in a network
    - Example: change of interaction patterns over time among new students as they get to know each other
  - Shock: change reason is exogenous to the network
    - Example: change of interaction patterns among students after they graduate
  - Another way to say it: detect "fast" change not "slow" change
- Another goal is to identify *change point*
  - Likely time when change occurred
  - Limits the scope of explanation for network change

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Statistical Process Control (SPC)

- Change detection based on SPC
- Statistical Process Control
  - Used in manufacturing to maintain quality control
  - Monitors a process to detect potential changes
  - Calculates a statistic from observed measurements of a process and compares it to a decision interval
  - If the statistic exceeds the decision interval, it is said to "signal", that a potential change may have occurred
  - A quality engineer will then begin to search for the specific cause of change

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

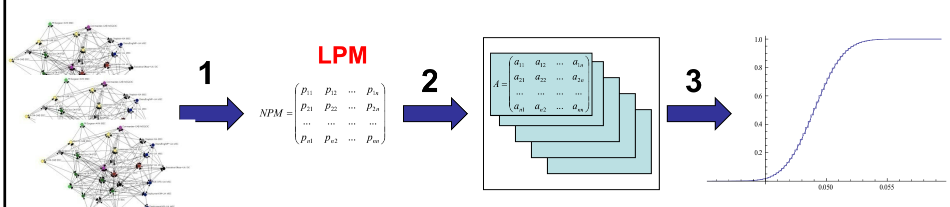
Statistical Models of Networks

Link Probability Model (LPM) for Stability

- LPM is a model for a network in *Stability*
- The probability that an email is sent from *i* to *j* within some period of time *t* is:

$$p = \int_0^t f_{ij}(x \mid \theta_{ij})\, dx$$

  - (p, as a function of t, is a CDF: f is the PDF that best fits cell ij in an NPM)
- LPM can be used to simulate stable longitudinal networks



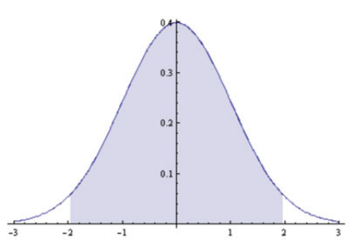
Statistical Models of Networks
Link Probability Model (LPM) for Stability

LPM simulated networks are compared to empirical networks and are shown to represent the network well.

M	*8*		*N*	*60000*			
e_mean	**e_stdev**		**s_mean**	**s_stdev**		**t-val**	**p**
409.2857	38.5604		358.0939	12.77466		3.754923	0.00
365.8571	18.2978		320.0974	12.7394		7.073195	0.00
365.8571	29.04266		320.1638	12.79331		4.449958	0.00
377.8571	38.24669		330.6744	12.77289		3.489244	0.00
375.2857	36.10039		328.3765	12.79551		3.675254	0.00
349.8571	38.15944		306.0783	12.7845		3.244918	0.00
373.8571	48.45076		327.0728	12.82622		2.731135	0.01
362.4286	55.63529		317.1509	12.77754		2.301849	0.02

Probability Background

- Consider a normal distribution with μ=0 and σ=1.
- 95% of the time, observations are between ±1.9597
- When an observation occurs in the tail, we don't believe it and think that something unusual might be going on.



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Statistical Process Control

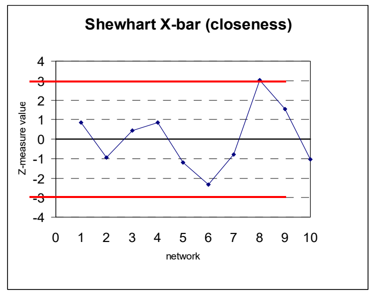
- Manufacturing processes are: stochastic, dependent, non-ergotic, complex, and involve human interaction.
- Shewhart (1927) X-bar Control Chart proposed to monitor change of any process
- Calculate Z_t transform value for each time-period, t.

$$Z_t = (x_t - \mu_0)/\sigma$$

- Calculate a control limit, L, based on risk for false alarm.

$$\int_L^{\infty} f(x)dx = \alpha$$

- Chart Signals when Z exceeds control limit, L.

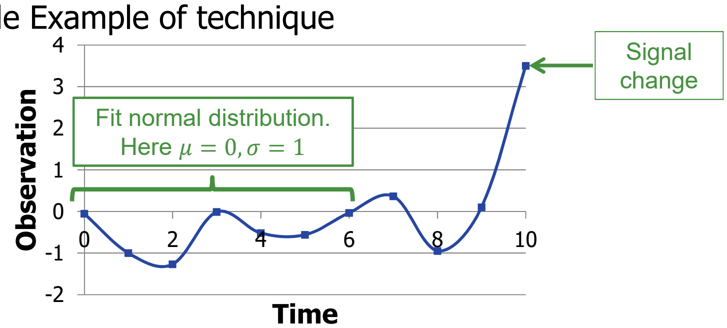


CASOS

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

The Shewhart X-Bar Chart

- Overview
  - Fit normal distribution on "control period" (early observations)
    > assumed to represent the "normal state"
  - Signal change if a subsequent observation is outside confidence interval
- Simple Example of technique



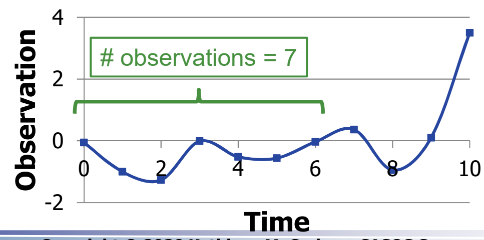
CASOS

CASOS

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

The Shewhart X-Bar Chart

- Parameters
  - # observations used to fit distribution (the "normal" period)
  - False positive risk or decision interval
    - Trade-off between False positive risk & detection speed
- Assumption
  - Observations are normally distributed as independent random vars
    - Shewhart X-Bar chart used even when assumption is violated. However, false positive risk probability may be inaccurate



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Statistical Process Control (cont.)

- Newer approaches detect change in fewer observations subject to the same rate of false positives.
- Scan Statistic (Fisher, 1934)
- Exponentially Weighted Moving Average (EWMA) (Roberts, 1959)
  - Good at detecting small changes in mean over time
  - Performs well on time series with closely spaced data samples

$$w_t = \lambda \bar{x}_t + (1-\lambda) w_{t-1} \qquad \mu_0 \pm L\sigma_{\bar{x}} \left(\frac{\lambda}{2-\lambda} \left[1-(1-\lambda)^{2T} \right] \right)^{1/2},$$

- Cumulative-Sum (CUSUM) Control Chart (Page, 1961)
  - Good at detecting small changes in mean over time
  - Built-in change point detection
  - Two Charts (To Detect Increase and Decrease)

$$C_t^+ = \max\{0, Z_t - k + C_{t-1}^+\} \qquad C_t^- = \max\{0, -Z_t - k + C_{t-1}^-\}$$

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Cumulative Sum (CUMSUM)

- Cumulative-Sum Control Chart
  - Good at detecting small changes in mean over time
  - Built-in change point detection
- Calculate Z_t transform for each time-period, t

$$Z_t = \left(x_t - \mu_0\right) / \sigma$$

- Two Charts (To Detect Increase and Decrease)

$$C_t^+ = \max\{0,\ Z_t - \frac{\delta}{2} + C_{t-1}^+\}$$

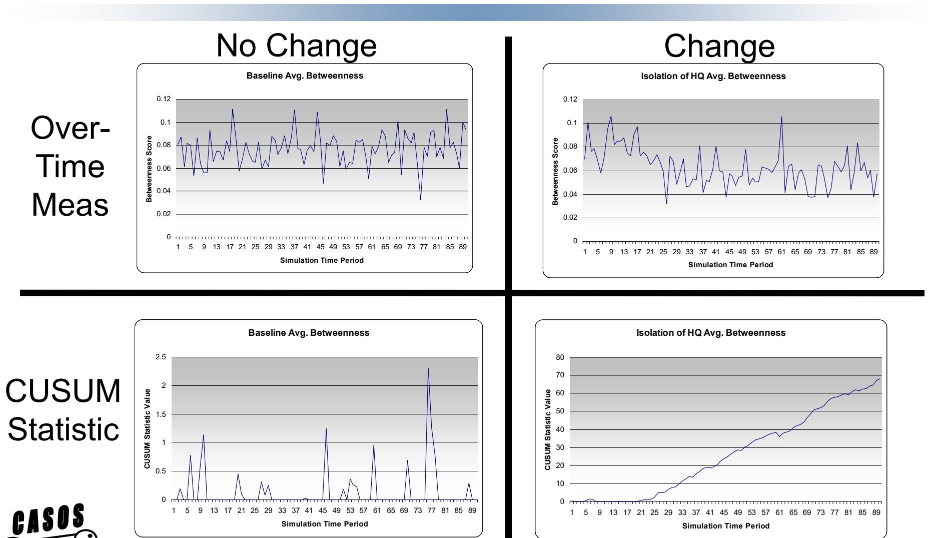
- Chart Signals when C^+ or C^- statistic exceeds decision interval

$$C_t^- = \max\{0,\ -Z_t - \frac{\delta}{2} + C_{t-1}^-\}$$

- Sensitivity in CUSUM due to discrete integration of error

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Comparison of Change Detection Approaches



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Comparison of Change Detection Approaches

	CUSUM $k = 0.5$	EWMA $r = 0.1$	EWMA $r = 0.2$	EWMA $r = 0.3$	Scan Statistic
Average Betweenness	9.32	8.24	10.16	11.52	6.76
Maximum Betweenness	14.36	14.72	15.72	17.08	13.24
Std Dev. Betweenness	16.44	16.24	16.92	18.52	15.24
Average Closeness	10.68	9.08	13.60	17.52	10.48
Maximum Closeness	8.76	6.00	10.60	37.96	8.64
Std Deviation Closeness	34.48	34.72	34.52	35.68	27.08
Average Eigenvector	31.28	31.28	31.28	31.28	24.00
Minimum Eigenvector	14.36	14.36	14.28	15.56	14.88
Maximum Eigenvector	5.24	5.40	5.80	7.52	4.00
Std. Dev Eigenvector	5.92	4.88	6.40	6.96	3.64

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

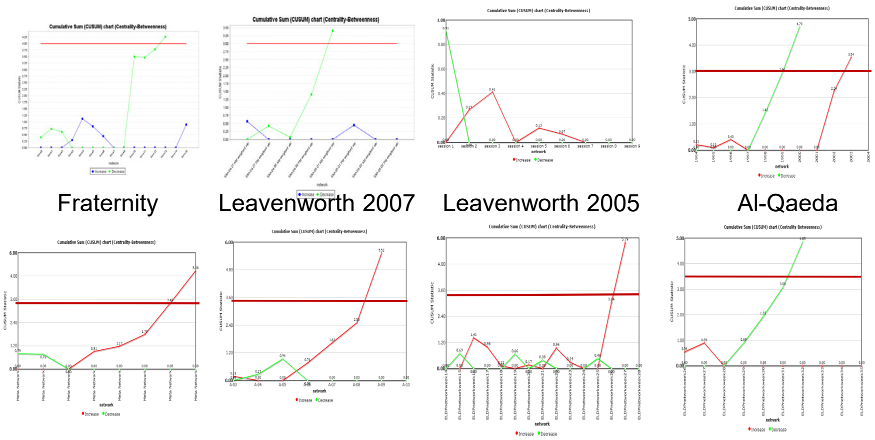
Network Change Detection: Analysis of Real World Data

	# Nodes	Time Periods	Method of Collection	Type of Relation	Design	Known Change
Fraternity	17	15	Survey	Ranking	Fixed	Yes
Leav 07	68	8	Survey	Rating	Free	Yes
Leav 05	158	9	Survey	Rating	Free	None
Al-Qaeda	62-260	17	Text	Rating	Free	Yes
Winter C	22	9	Observation & Survey	Rating	Fixed	Yes
Winter A	28	9	Observation & Survey	Rating	Fixed	Yes
IkeNet 2	22	46	Email	Count Msg	Free	Yes
IkeNet 3	68	121	Email	Count Msg	Free	Yes

CASOS June 2020

Carnegie Mellon
institute for SOFTWARE RESEARCH

Network Change Detection: Analysis of Real World Data



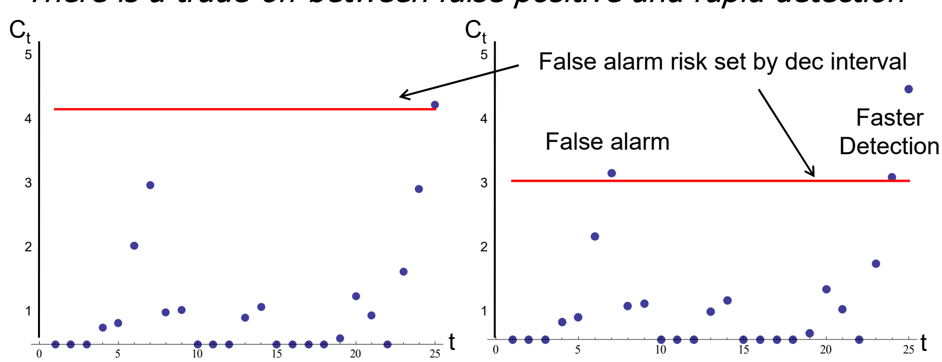
Fraternity | Leavenworth 2007 | Leavenworth 2005 | Al-Qaeda

Winter A | Winter C | IkeNet 2 | IkeNet 3

Carnegie Mellon
institute for SOFTWARE RESEARCH

Summary of Change Detection Across Data Sets

There is a trade-off between false positive and rapid detection



Low risk of false alarm
Longer to detect change

High risk of false alarm
Faster to detect change

Summary of Change Detection Across Data Sets

Too little risk may prevent change detection all together

Data	Change	$\alpha = 0.05$	$\alpha = 0.02$	$\alpha = 0.01$	$\alpha = 0.005$	$\alpha = 0.001$
Fraternity	8	10	10	10	13	Never
Leav 07	3	5	5	5	Never	Never
Leav 05	None	No F.A.	No F.A.	No F.A.	No F.A.	No F.A.
Al-Qaeda	1997	1999	1999	2000	2000	Never
Winter C	May	Sept	Sept	Oct	Oct	Never
Winter A	May	Aug	Sept	Sept	Sept	Oct
IkeNet 2	25	26	26	27	27	27
IkeNet 3	14	15	18	19	19	20

Conclusions

- Change detection
  - Detect occurrence of shocks i.e. change due to reasons exogenous to the network
- Practical Change Detection normally focuses on metrics
  - Calculate selected metrics at points in time
  - Characterize the statistics of the metric under normal conditions (note, typically this involves assuming that if is AWGN)
  - Detect Change as a statistically unlikely event for metric
  - Can do multivariable change detection on multiple metrics at the same time
- Hands on Practice and use of Fourier Transform next