Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Fourier Analysis and Change Detection

Prof. L. Richard Carley

carley@ece.cmu.edu

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Dynamic Network Analysis

- Key focus – Networks change over time
- Summary statistics – typically average all data
  - Useless for seeing changes over time
- Longitudinal Networks and Change
  - Getting longitudinal networks from communications logs
  - Stability, Evolution, Shock, Mutation
- Statistical Models of Networks to Detect Change
  - Link Probability Model (LPM) for Stability
  - Actor-Oriented Models for Evolution
  - Multi-Agent Simulation for Evolution, Shock, and Mutation
- Network Change Detection Algorithms
- Fourier Analysis to remove periodic variations

 Copyright © 2016 Kathleen M. Carley – Director CASOS, ISR, CMU

Basic Issue

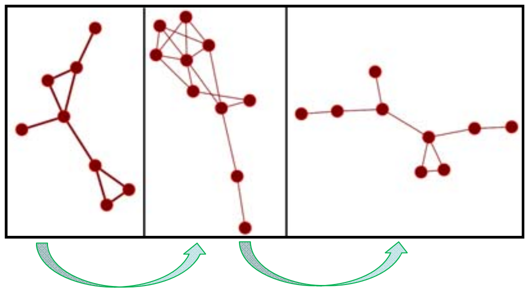
- Real Social Networks are not time independent
- Over time the set of nodes change
  - Agents die, agents are born
  - If data set has limited geographic focus,
    - Agents can enter region under study
    - Agents can leave region under study
- Network connections between agents can change
  - A network link between two agents can disappear
    - Two family members have a fight and refuse to talk to each other
  - A new network link can be created
    - People meet new people and form new relationships
    - Advertising campaigns can convince people to follow companies

Types of Changes in Network Data

- Stability: Relationships remain statistically the same over time
  - If you are a signal processing person, the Network is "Ergodic"
- Evolution: Interaction among agents cause the relationships to change over time.
  - All link weights / costs are evolving over time during obervations
- Shock: Change is exogenous to the social group.
  - E.g., like an earthquake hits Southern California
- Mutation: A shock stimulates evolutionary behavior.
  - E.g., after earthquake, people form many new links trying to survive

Dynamic Metrics on Over-Time Data

- *Identifying central nodes in a network*



T_1 T_2

Dynamically Changing Network Structure!!!

Proxy for Network Data

- 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
- Or, tracking changes over time using communications data
  - Communication is "proxy" for a network tie
  - Taking large amounts of communication data gives an approximate picture of the underlying social network (with some concerns)
  - Can use it to find Key Agents and other Social Structure measures
- Communication log data available from many sources
  - Cell Phone Service Providers – call logs, txt logs
  - E-mail Data logs – available within organization
  - Twitter, Facebook, FourSquare, etc.
  - Building Sensors, Cell Phone Sensors, RFID Tags, etc.

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

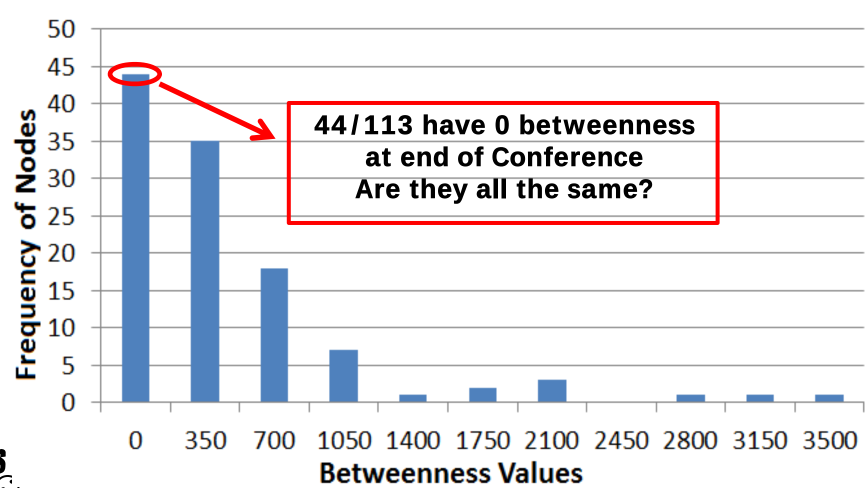
Example: Temporal Social Network

- 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

New Interactions → **Stronger Relations** → **Different Interactions**

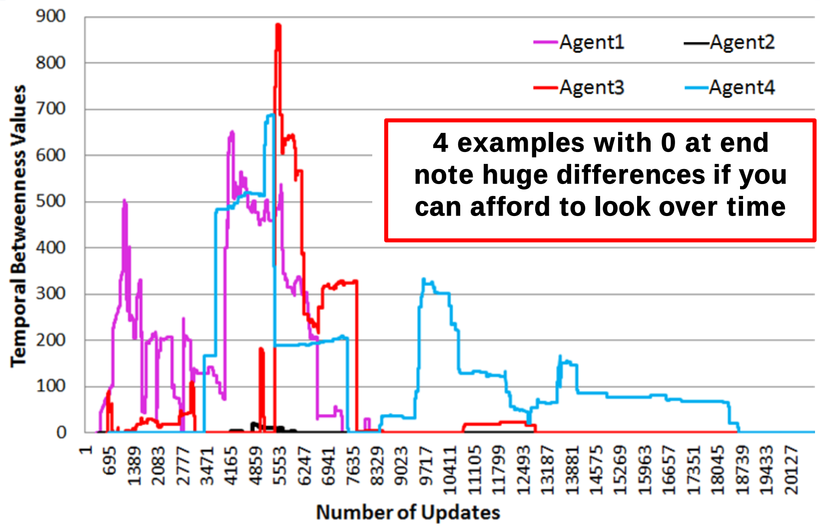
Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Socio-Patterns: Betweenness Centrality Distribution



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Socio-Patterns: Betweenness Over-Time Trends

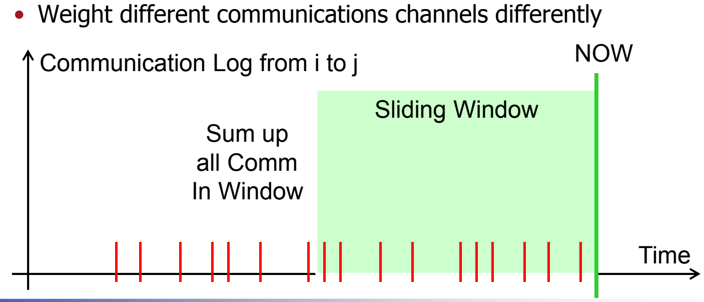


CASOS

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Sliding Window for Over-Time Links

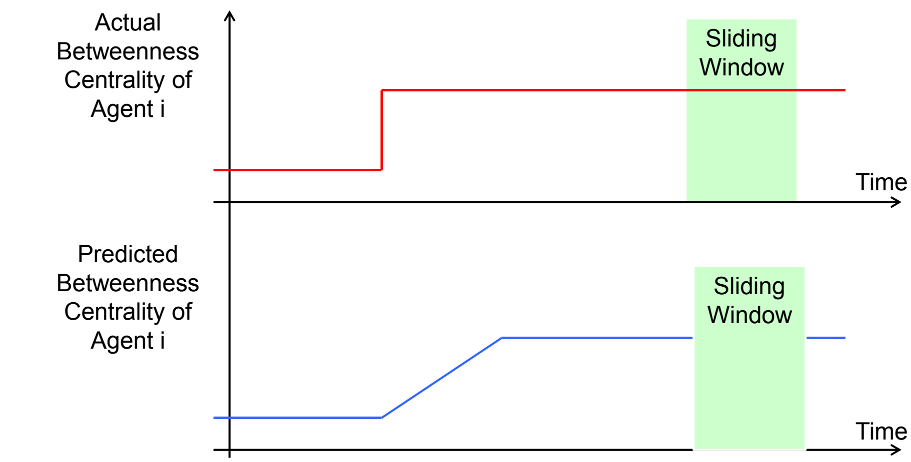
- Estimator for Link Weight (aka Link Cost)
  - Add up # of Communication Events between x-y in window
  - Take reciprocal. If # is 0, there is no Link between that pair
  - Then move window forward by a time step and repeat
  - Alternatives possible:
    - Incorporate duration of communication
    -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



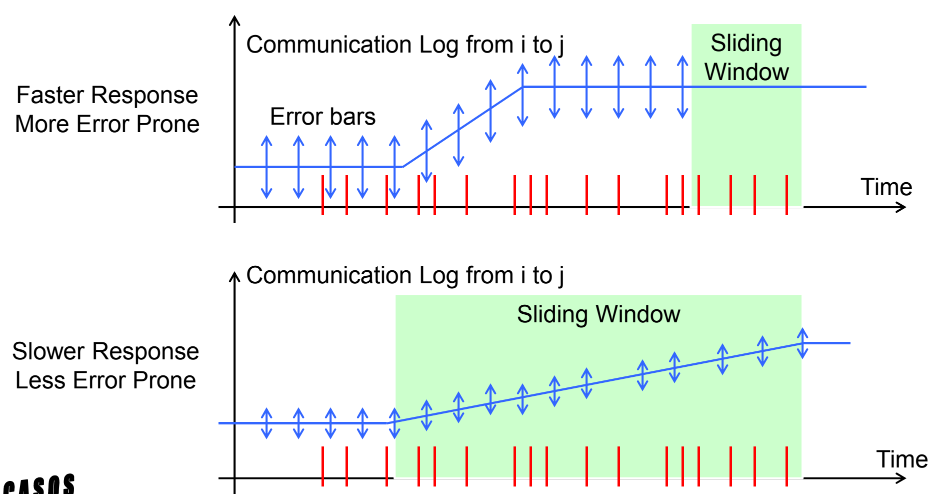
CASOS

CASOS

Smoothing Effect of Sliding Window

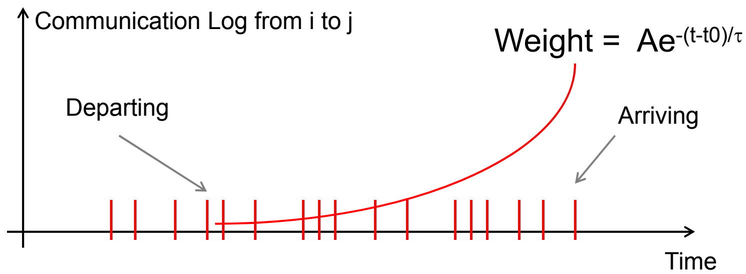


Adjusting Window Size



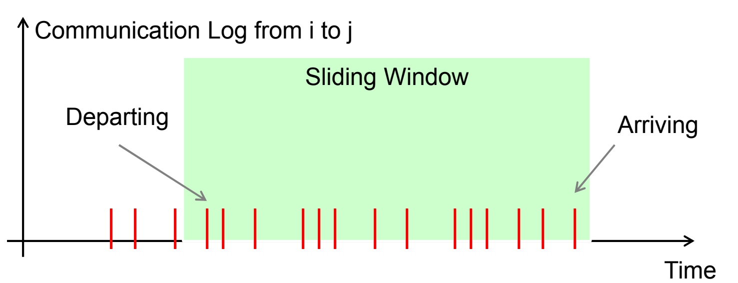
Mathematically Better Window

- 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



Incremental Sliding Window

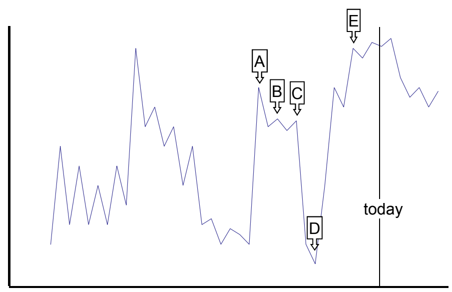
- Sliding Window is Synergistic with Incremental Analysis
  - As window moves forward in time
    - New events “arrive” and must be processed
    - Old events “fall out” of trailing edge of window and must be processed
    - BUT – all of the data in middle of window remains unchanged
    - Incremental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



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Changes in Network Data

- Various measures of a network are taken for a window at each time point.
- 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

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

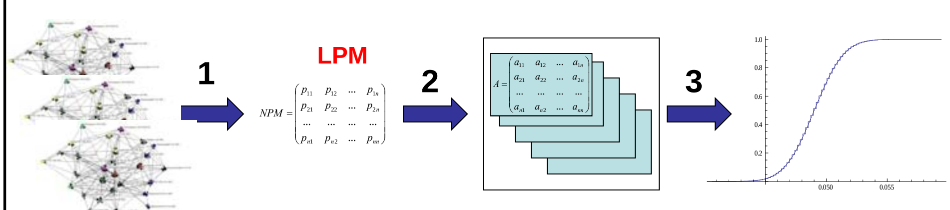
Statistical Models of Networks

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



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Statistical Models of Networks

Link Probability Model (LPM) for Stability

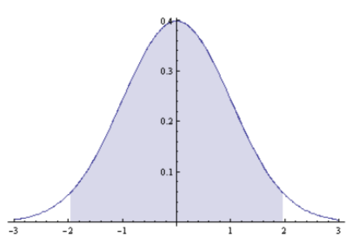
LPM simulated networks are compared to empirical networks and are shown to represent the network well.

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

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

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

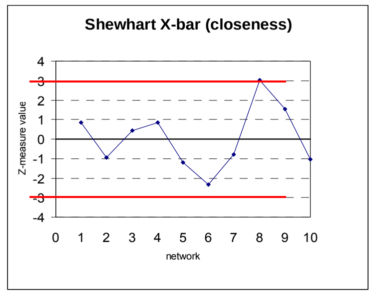
- Manufacturing processes are: stochastic, dependent, non-ergotic, complex, and involve human interaction.
- Shewhart (1927) X-bar Control Chart proposed to monitor change
- Calculate Z_t transform value for each time-period, t.

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

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

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

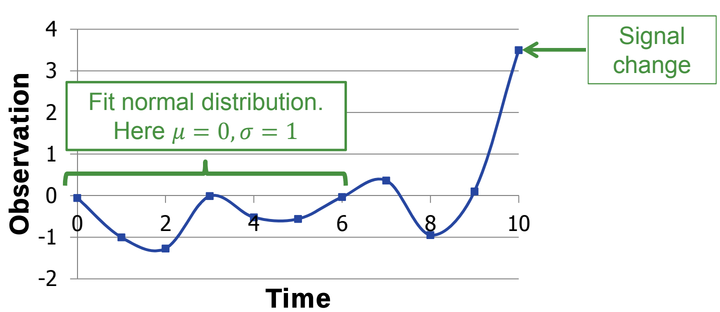
- Chart Signals when Z exceeds control limit, L.



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

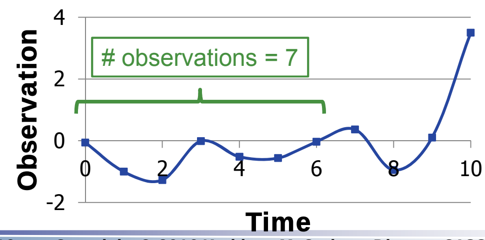
The Shewhart X-Bar Chart

- Overview
  - Fit normal distribution on early observations
  - Signal change if a subsequent observation is outside confidence interval
- Simple Example of technique



The Shewhart X-Bar Chart

- Parameters
  - \# observations used to fit distribution
  - False positive risk or decision interval
    - Trade-off between False positive risk & detection speed
- Assumption
  - Observations are normally distributed
    - Shewhart X-Bar chart used even when assumption is violated. However, false positive risk is inaccurate



Statistical Process Control

- New approaches detect change in less 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 data

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

- 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)

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

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

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

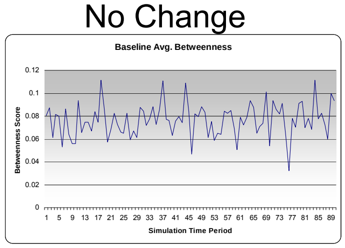
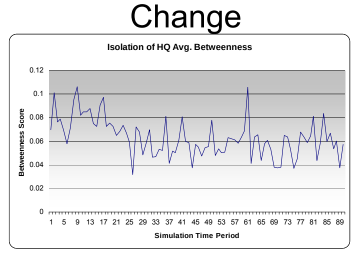
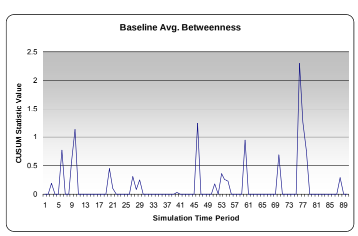
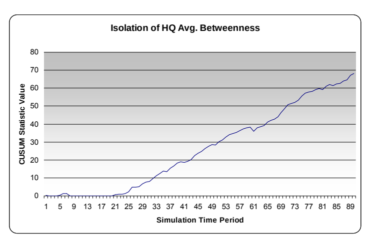
- Two Charts (To Detect Increase and Decrease)

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

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

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

Comparison of Change Detection Approaches

	No Change	Change
Over-Time Meas		
CUSUM Statistic		

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

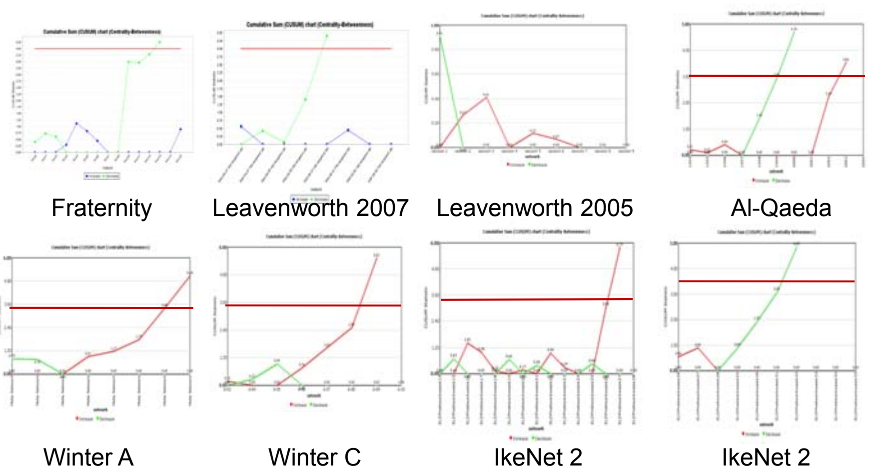
Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Network Change Detection: Analysis of Real World Data

	No 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

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

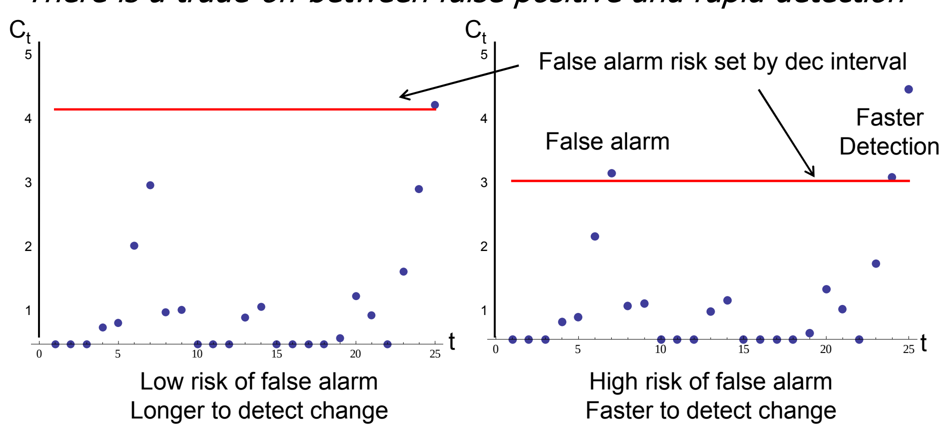
Network Change Detection: Analysis of Real World Data



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Summary of Change Detection Across Data Sets

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



Carnegie Mellon

Summary of Change Detection Across Data Sets

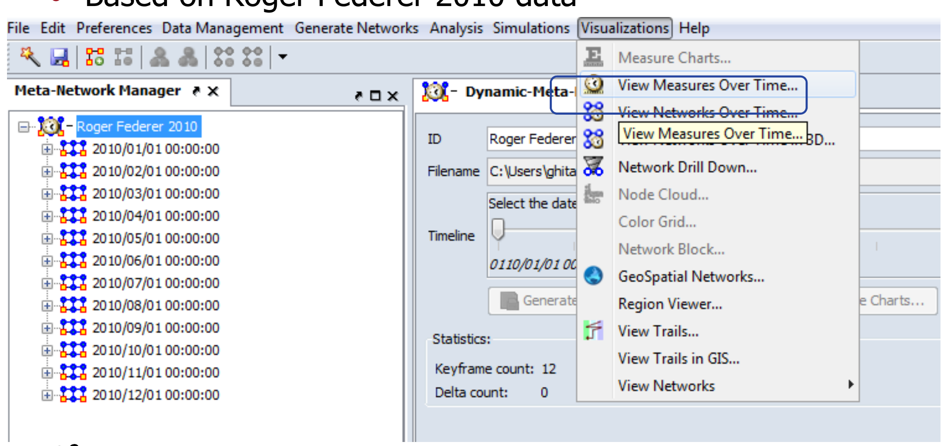
Too little risk may prevent change detection

Data	Change	α = 0.05	α = 0.02	α = 0.01	α = 0.005	α = 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

Carnegie Mellon

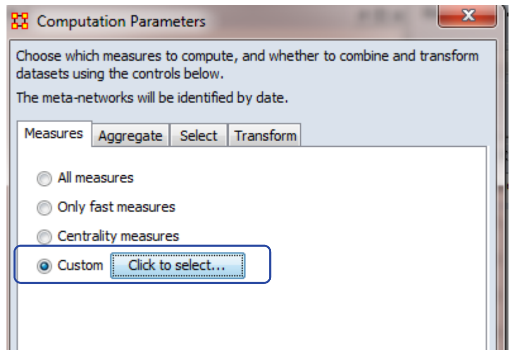
Change Detection Hands-On

- Based on Roger Federer 2010 data



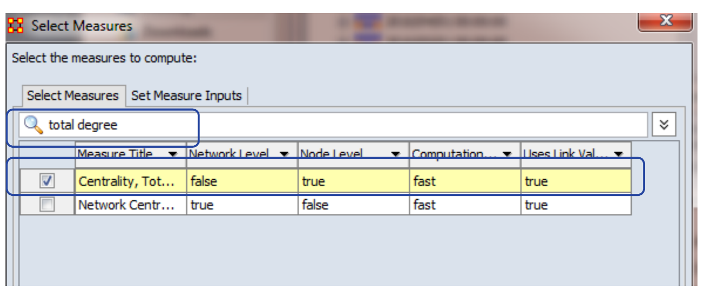
Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Change Detection Hands-On



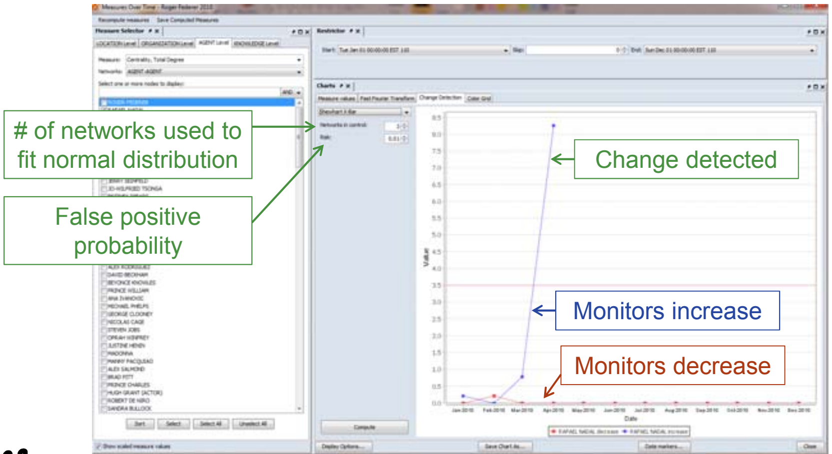
Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Change Detection Hands-On



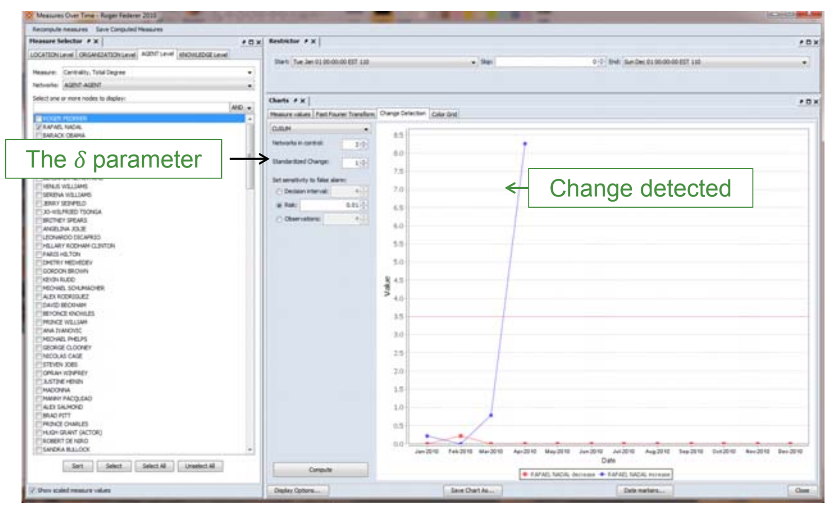
Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Change Detection Hands-On
The Shewhart X-Bar Chart



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Change Detection Hands-On
CUMSUM Method



Carnegie Mellon

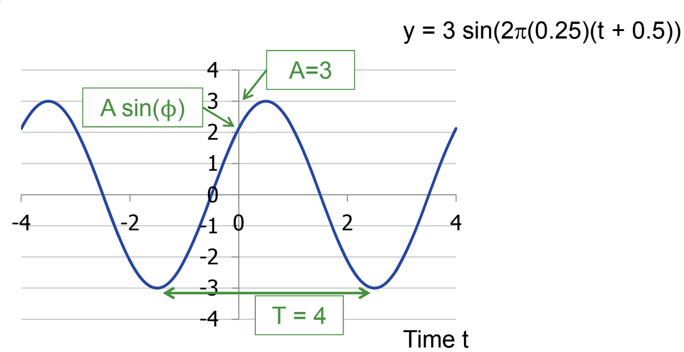
Fast Fourier Transform (FFT)

- Goal: detect periodicity in over-time data
- Examples
  - Weekly periodicity in email data
  - Time of the day effects
- Fourier's theorem
  - Any time signal is composed of sinusoidal functions with different frequencies, amplitudes and phase shifts
- Fourier transform finds sinusoids that decompose a signal
  - Analogy: given a dish, find the ingredients
  - Sinusoids have the advantage that they are orthogonal

Carnegie Mellon

Sinusoidal Function

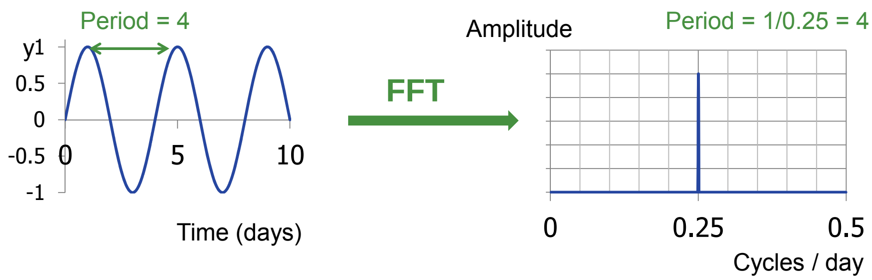
- A sinusoidal function $y = A\sin(2\pi f t + \phi)$ has
  - A amplitude
  - f frequency $(T = \frac{1}{f}$ is the period)
  - ϕ phase



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

FFT Example: Sinusoidal Function

- Fast Fourier transform of sinusoidal function is a spike at the sinusoidal frequency
- Example $y = sin(2\,\pi\,0.25\,t)$

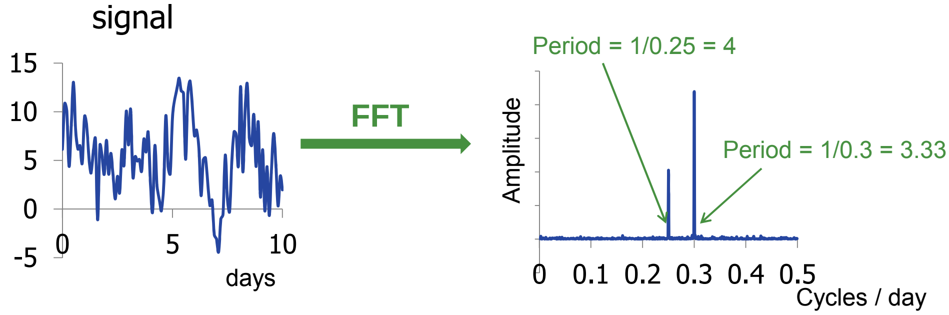


Over-time signal | FFT

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

FFT Example 2

- FFT finds periodicities that may be unclear in over-time signal



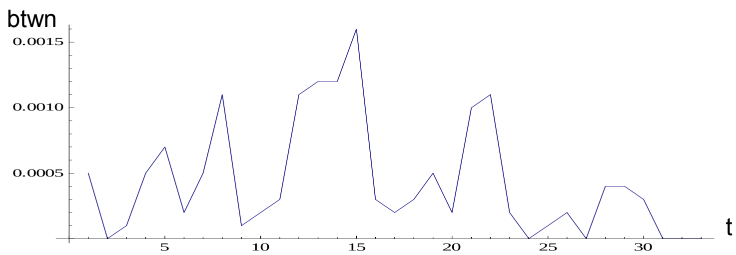
Over-time signal | FFT

Hidden "recipe": over – time signal computed as
y(t) = 2 sin(2 pi 0.25 t) + 3 sin(2 pi 0.3 t + 0.2) + noise

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Fourier Analysis Example 3

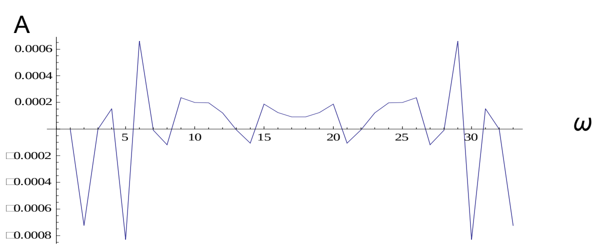
- 24 cadets in a regimental chain of command agreed to have their email monitored to form a social network data set known as IkeNet3.
- The betweenness over the first month in their duty positions.



CASOS

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Fourier Analysis – Example 3



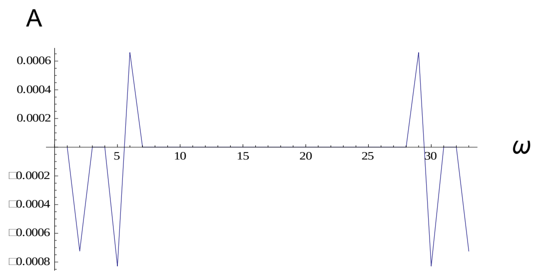
Fourier transform
Symmetric around the midpoint
3 main components (in terms of magnitude)

That is why we typically only display from origin up to midpoint

CASOS

CASOS

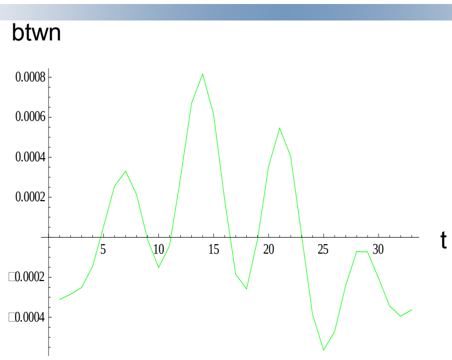
Filtering



3 main (high magnitude) components picked out
The others have been clipped out

ω Frequency in radians per second

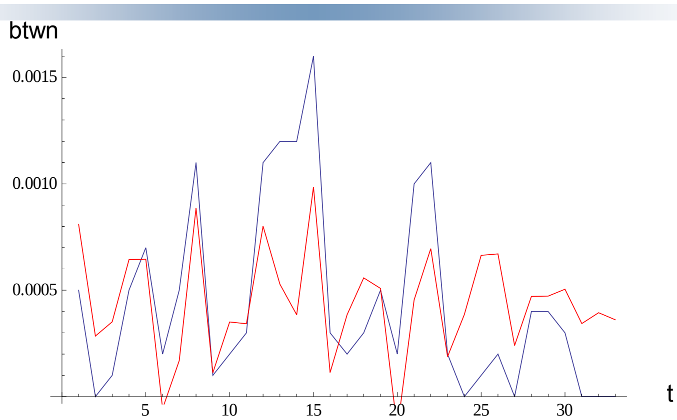
Inverse Fast Fourier Transform



This is the inverse Fourier transform
The filtered 3 components have been reconverted to time
There is a weekly, two week and three week cycle

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Anomaly Detection

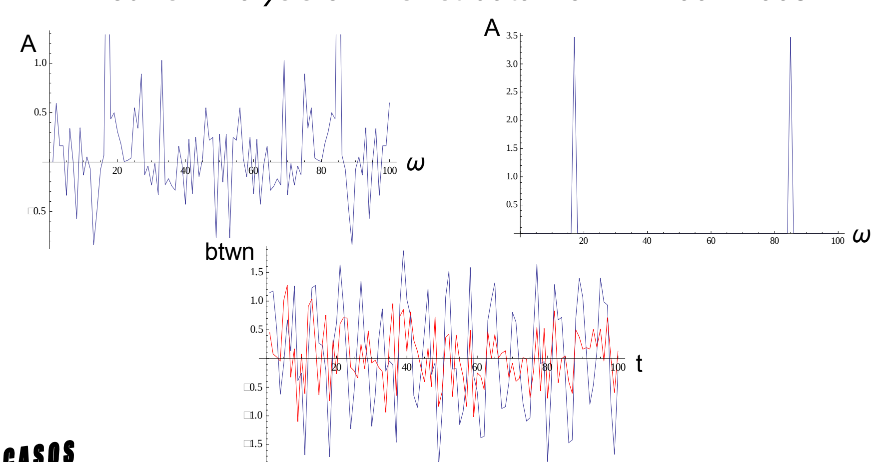


The filtered pattern has been subtracted from the original
The red is what is left – the anomalies

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Fourier Analysis to Handle Periodicity

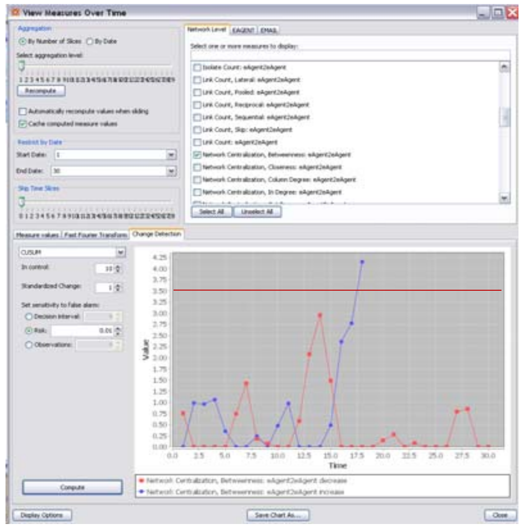
Fourier Analysis on IkeNet data from AY2007-2008



Carnegie Mellon
institute for SOFTWARE RESEARCH

Supporting Technology

- Classic view measures over time
- Fourier analysis
- Network change detection
- Aggregate networks
- Graph & Agent meas.
- Reports
  - Change detection
  - Stat distribution fitter



Carnegie Mellon
institute for SOFTWARE RESEARCH

FFT Hands On Session

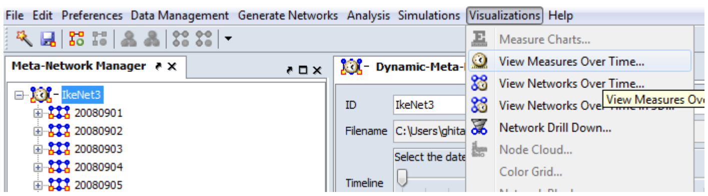
File – Fourier-Example-3.xml

Walk through analysis on screen and on your laptops

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

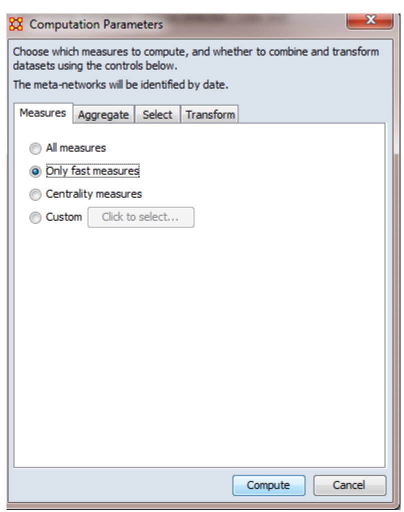
FFT Example Hands-On (1/4)

- IkeNet data (IkeNet3-dynamic.xml)
  - Email exchange data among mid-career officers in a one-year graduate program at Columbia University
  - Granularity: day
  - Duration: month



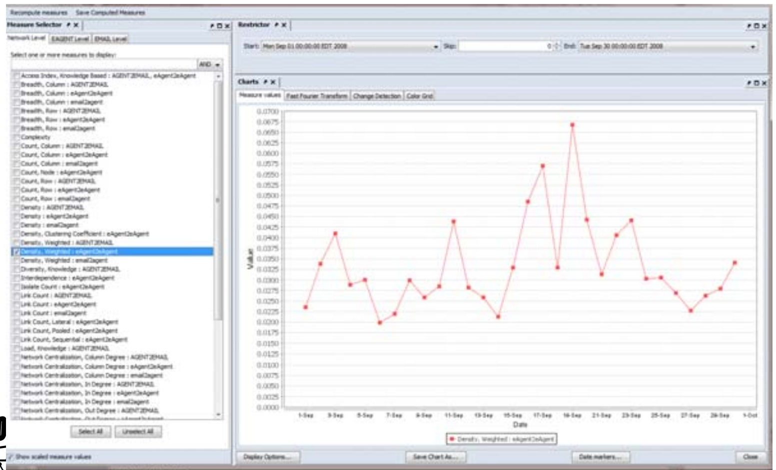
Carnegie Mellon
isr institute for SOFTWARE RESEARCH

FFT Example Hands On (2/4)



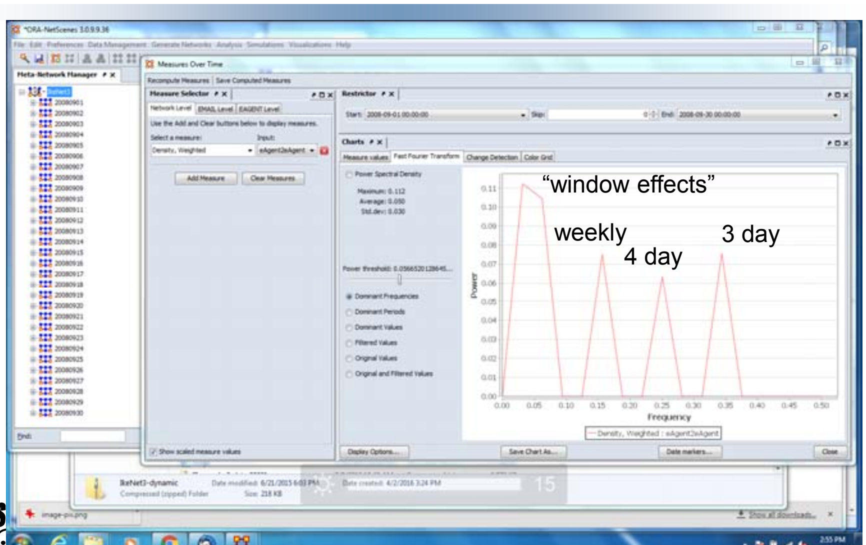
Carnegie Mellon
isr institute for SOFTWARE RESEARCH

FFT Example Hands On (3/4)



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

FFT Example Hands On (4/4)



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Fourier Analysis to Handle Periodicity

- Fourier analysis can effectively identify periodic trends in longitudinal network data.
- Identification of periodic trends can allow the analyst to aggregate relational data over the period to remove over-time dependence.
- The inverse Fourier transform of the significant period can be used to filter out periodicity from longitudinal network data.
- Further exploration of wavelets may produce greater insights in to network dynamics.

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Scalability

- The change detection algorithm is linear, thus the time consuming part is calculating network measures.
- Networks with less than 20 nodes tend to have a higher variance in over time measures. When a link is added or removed, it affects (n-1)(n-2) triads.
- Requires at least 3 time periods: 2 to determine typical behavior and 1 to compare. In practice, 10+ network time points are preferred.
- No difference in number of required networks for each technique: CUSUM, EWMA, Scan Statistic, x-bar, eyeball
- Wavelet/Fourier based approach needs many more time periods

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Limitations

- View findings on data with caution.
- Examine errors associated with technique through extensive simulations.
- Investigate more real world data sets.
- Investigate the degree to which network measures are correlated to understand the effects of compounding error.
- Investigate multi-dimensional network properties such as the cosine similarity between the triad census at different time periods.

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Summary Results

- Rapid change detection may allow an analyst to get inside a decision cycle and shape network evolution.
- Simulation is important for modeling longitudinal network behavior.
- Isolating when networks change enables more focused study on the causes of evolution, shock, and mutation, which may lead to future predictive analysis.
- Statistical process control is a useful tool for understanding social behavior.

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Conclusions

- Change detection
  - Detect occurrence of shocks i.e. change due to reasons exogenous to the network
- Fourier analysis
  - Detect periodicity in over-time data