<Your Name>

Dynamic Networks

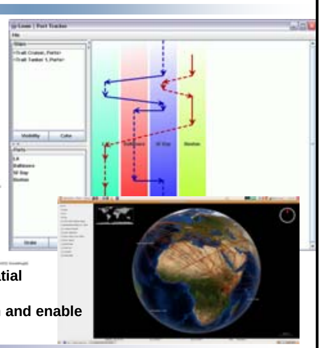
Kathleen M. Carley

kathleen.carley@cs.cmu.edu

Overview

- Networks
- Common metrics
- Dynamics
- Spatial
- Where does data come from

Network Science - Connecting the Dots and Trails to Predict and Explain Behavior



Identify Groups / Threats Using Network and Geo-Spatial Information
How does space – constrain and enable networks

<Your Name>

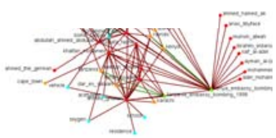
What is a Network?

Ties Between Nodes (links)

- Who do you like or respect?
- Transfer of resources
- Authority lines
- Association or affiliation
- Alliance
- Substitution
- Precedence
- Proximity

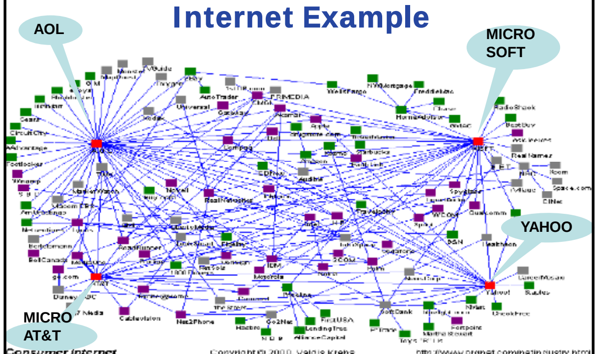
Nodes

- People
- Units of action
- Coalition partners
- Departments
- Resources
- Ideas or Skills
- Events
- Nation-states



Networks are ubiquitous

Internet Example



Some Nodes Stand Out

What is Dynamic Network Analysis?

- The study of how entities are constrained and enabled by the relations among them and the process that lead to change in these relations
- Combines social networks analysis, link analysis, multi-agent modeling, machine learning, graph theory, and non-parametric statistics
- Complex Meta-Networks: multiple networks, multiple types of nodes, multiple relations
- Key Issues: Scalability, Robustness, Flexibility, Error
  - Relations among nodes are flexible and vary in strength and certainty
  - Node membership may be questionable
  - Networks may be large 10^6 nodes
  - Classes of data may not be discoverable

Dynamic Network Analysis

- *The Network Perspective*
  - *It's not just the elements (composition) of a system, but how they are put together*
  - *non-reductionist, holistic*
- What are networks and how do you analyze them?
- Social Network Analysis, Link Analysis, Network Text Analysis, Dynamic Network Analysis
- Network Elites
- Groups and clustering
- Consensus and networks
- Network Topology
- Compare and contrast networks
- Network dynamics
- Network Visualization

Where are Dynamic Network Analysis Models Used

- Designing adaptive teams for Command and Control
- Evaluating organizational structures and evaluating changes such as downsizing
  - E.g., hospitals, health departments NY Dutchess County
- Estimating effectiveness and adaptability of new structures
  - E.g., SSG – Comcargru, Army Unit of Action, CPOF (IRAQ)
- Estimating size, shape and vulnerabilities in organizational designs and covert networks
  - E.g., NASA, Counter-terrorism, drug, terrorist, tax-avoiders
- Network management and IT intervention/effectiveness analysis
  - E.g., NASA, Knowledge Wall in JTF, supply chains, various companies
- Impact analysis of actions in asymmetric warfare situations
- Impact on cities of weaponized biological or chemical attacks
- Identifying key actors and emergent groups
  - E.g., Counter terrorism, Health Units, Merchant Marine
- Prevention and intervention
  - E.g., IRS tax avoidance interventions

Connect & *Dis*-Connect the Dots!

	Degree	Betweenness	Closeness
1	0.417 Mohamed Atta	0.334 Nawaf Alhazm	0.571 Mohamed Atta
2	0.389 Marwan Al-Shehhi	0.318 Mohamed Atta	0.537 Nawaf Alhazmi
3	0.278 Hani Hanjour	0.227 Hani Hanjour	0.507 Hani Hanjour
4	0.278 Nawaf Alhazmi	0.158 Marwan Al-Shehhi	0.500 Marwan Al-Shehhi

Figure 3 Trusted Prior Contacts + Meeting Ties [shortcuts]

Standard Social Network Measures

So – why is this hard?

- The Network
  - Vast quantities of data
  - Multi-mode – people, events, etc.
  - Multi-plex – many connections e.g. financial and authority
- The Information
  - Intentional misinformation – e.g., aliases
  - Inaccurate information – e.g., typos
  - Out-of-date information
  - Incomplete information
- Dynamic
  - Learning
  - Recruitment
  - Attrition
  - ...

Typical Mainstream Data Structure

Variables (attributes)

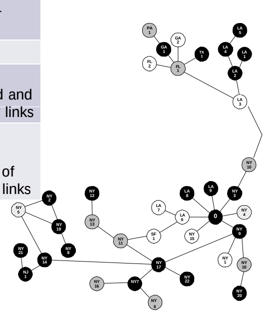
Cases (individuals)

	Age	Gen	Education	Income	State
1001	28	0	< highschool	28,000	PA
1002	35	1	highschool	54,000	MA
1003	26	0	< highschool	26,000	MA
1004	40	1	Bachelors	65,000	MA
1005	24	0	highschool	27,500	PA
1006	55	1	Ph.D.	82,800	PA
1007	31	1	Ph.D.	73,000	MA
1008	M	0	highschool	33,500	PA
…					

Analysis consists of correlating attributes, regression, anova …

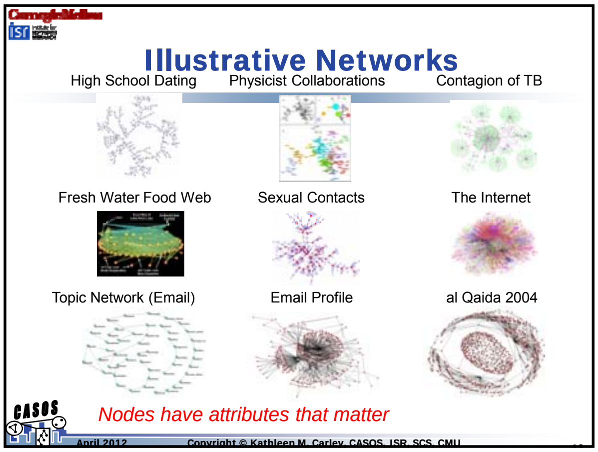
The Network Perspective

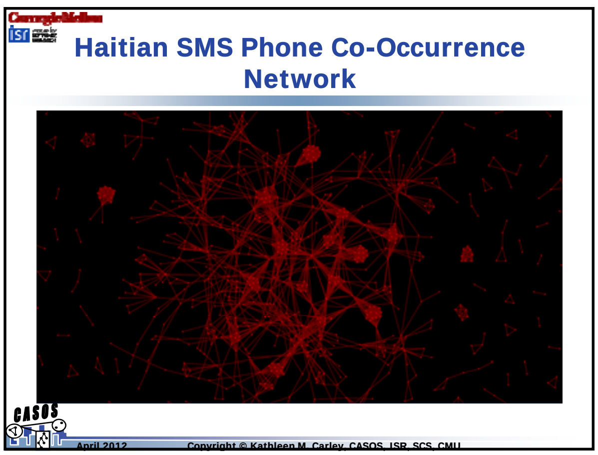
Standard Statistics	Social Network Analysis	Dynamic Network Analysis
Attributes	Relations	Relations + Attributes
Atomistic	Interdependence	
Actors as independent	Actors constrained and enabled by links	Actors constrained and enabled by links
Actor state matters	Actor state irrelevant	Actor state impacts perception of and use of links

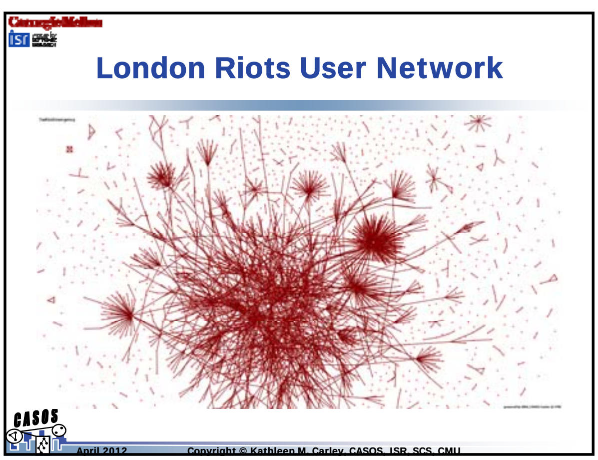


Discovery of HIV: Sexual contacts among gay men w/ unusual cancers, traced by Bill Darrow of the CDC

<Your Name>

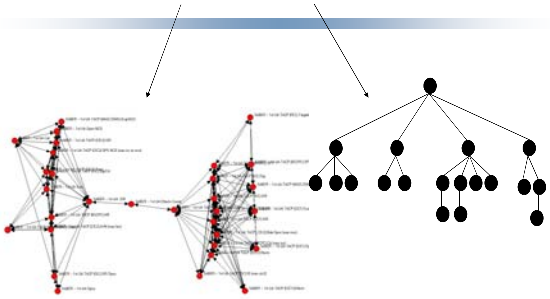

Illustrative Networks
High School Dating
Physicist Collaborations
Contagion of TB
Fresh Water Food Web
Sexual Contacts
The Internet
Topic Network (Email)
Email Profile
al Qaida 2004
Nodes have attributes that matter
CASOS
April 2012
Copyright © Kathleen M. Carley, CASOS, ISR, SCS, CMU


Haitian SMS Phone Co-Occurrence Network
CASOS
April 2012
Copyright © Kathleen M. Carley, CASOS, ISR, SCS, CMU


London Riots User Network
CASOS
April 2012
Copyright © Kathleen M. Carley, CASOS, ISR, SCS, CMU

<Your Name>

Informal and Formal Structure



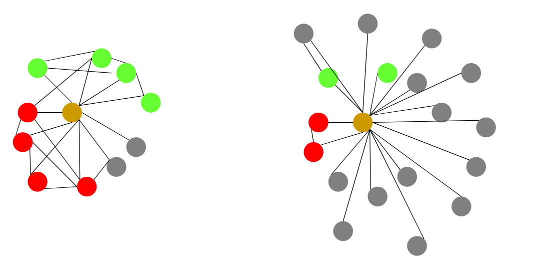
Each person is embedded in many networks

Drug Networks

Family
Work
Friend

Normal Person

Cocaine User

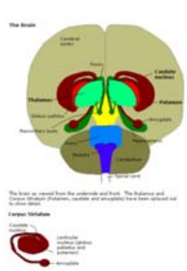


People with Different Roles have Different Networks

Amygdala?

- Controls processing and memory of emotional reactions
- Suggests building ties is an "emotional" act
- Since emotion after an event increases the chance the event will be remembered, recall of alters is higher if
  - Interaction is repeated (chance)
  - Interaction is emotional

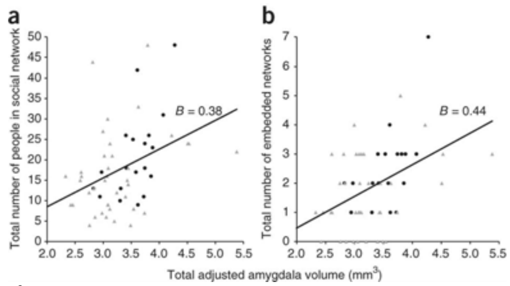
Joy	Sadness
Trust	Disgust
Fear	Anger
Surprise	Anticipation



From WikiPedia

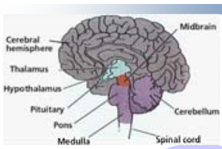
<Your Name>

The Size of Your Network Depends on the Size of Your Brain



Bickart, Wright, Dautoff, Dickerson and Barrett, 2010. "Amygdala volume and social network size in humans." Nature Neuroscience

Dunbar's Number



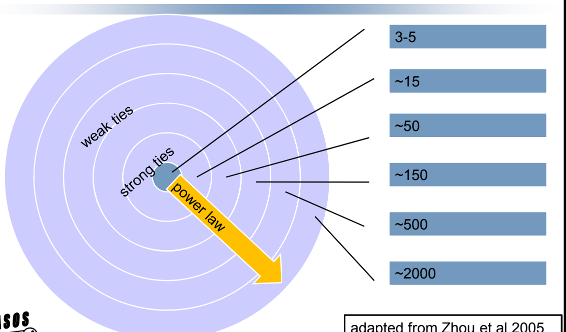
150

(well, actually 147.8)

- Hypothesized in 1993
  - the number of individuals that one person could follow based on extra-polations of neocortex size
- Wide variance
  - the original 95% CI was between ~100 and ~230, and was due to an extra-polation "well beyond" those known for primates

adapted from Dunbar 1993

How Many People Do You Know?



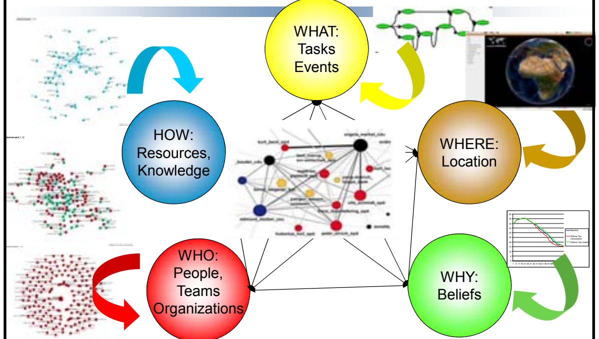
adapted from Zhou et al 2005

<Your Name>

In other words

- Network topology matters
  - Networks are not random
  - All networks are not scale free
- Informal and formal networks differ
- Individual's ego network
  - Constrains and enables behavior
  - Individual differences result in difference in ego net composition
  - Ties are emotion + frequency
  - Psycho physical constraints impose limits
- Overall size
  - Coordination constraints impose limits
  - Vary by media

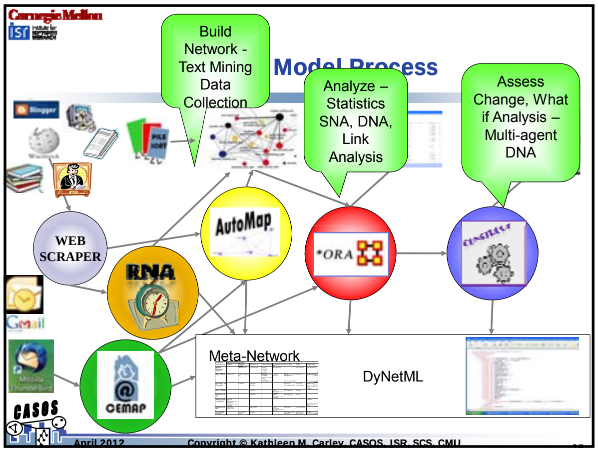
Social Networks are Linked to Other Networks

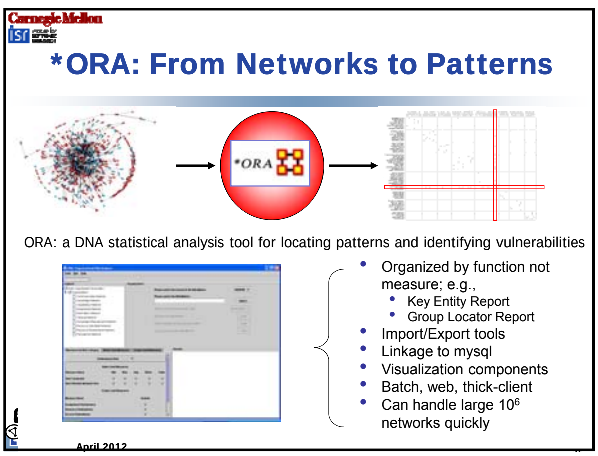


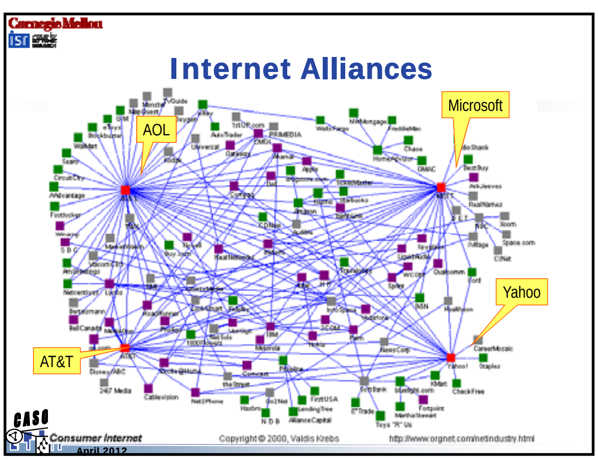
Meta-Network: multi-mode, multi-plex, multi-level

	People	Organizations	Expertise	Activities	Events	Locations
People	Social Network	Affiliation Network	Knowledge Network	Assignment Network	Participation Network	Presence Network
Organizations		Organizational Network	Capability Network	Action Network	Participation Network	Presence Network
Expertise			Information Network	Needs Network	Contributing expertise Network	Availability Network
Activities				Precedence Network	Contributing Acivity Network	Happenings Network
Events					Precedence Network	Happenings Network
Locations						Border Network

<Your Name>


Carnegie Mellon
Model Process
Build Network - Text Mining Data Collection
Analyze – Statistics SNA, DNA, Link Analysis
Assess Change, What if Analysis – Multi-agent DNA
WEB SCRAPER
AutoMap
RNA
*ORA
Meta-Network
DynetML
CASOS
April 2012
Copyright © Kathleen M. Carley, CASOS, ISR, SCS, CMU


Carnegie Mellon
*ORA: From Networks to Patterns
*ORA
ORA: a DNA statistical analysis tool for locating patterns and identifying vulnerabilities
Organized by function not measure; e.g.,
Key Entity Report
Group Locator Report
Import/Export tools
Linkage to mysql
Visualization components
Batch, web, thick-client
Can handle large 10^6 networks quickly
April 2012


Carnegie Mellon
Internet Alliances
AOL
Microsoft
Yahoo
AT&T
Consumer Internet
Copyright © 2000, Valdis Krebs
http://www.orgnet.com/netindustry.html
April 2012

<Your Name>

Network of Critical Actors

Identify: Who are the Key Players?

Agents 286, 652, 97 repeatedly rank in top three in the measures.

In-the-Know (total degree centrality)

Rank	Value	Unscaled	Agent
1	0.181	181	652
2	0.176	176	286
3	0.175	175	97
4	0.165	165	412
5	0.158	158	502
6	0.154	154	273
7	0.139	139	740
8	0.115	115	615
9	0.111	111	825
10	0.105	105	8

Potentially Influential (betweenness centrality)

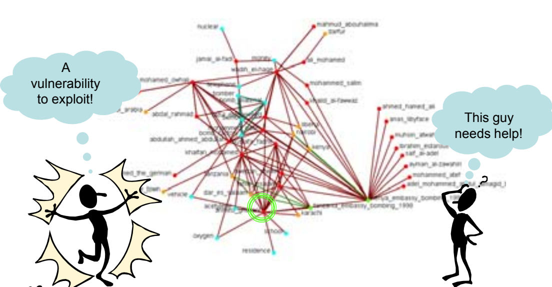
Rank	Value	Unscaled	Agent
1	0.0988664	24667.2	286
2	[illegible]	17600.4	97
3	[illegible]	[illegible]	502
4	[illegible]	[illegible]	[illegible]
5	[illegible]	[illegible]	652
6	0.0542421	13533.4	615
7	[illegible]	[illegible]	412
8	[illegible]	10245.8	501
9	[illegible]	7649.12	[illegible]
10	0.0305522	7622.78	552

Drilling down...

*ORA's **Key Entity Report** shows 3 agents critical to operations.

Narrow our focus from set of interstitial members to small group of leaders.

Network Analysis Enables Management and Disruption



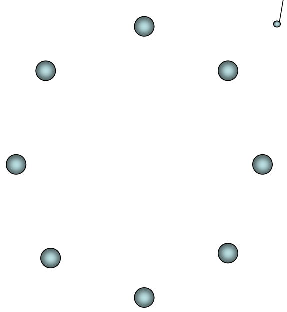
<Your Name>

Centralities

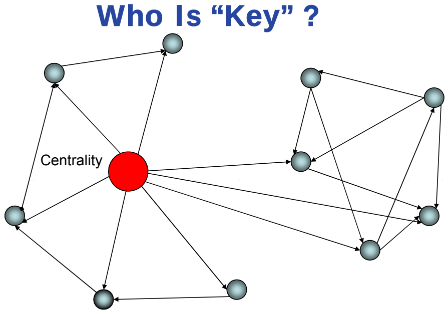
- Degree Centrality
  - Node with the most connections
- Betweenness Centrality
  - Node in the most best paths
    - Requires symmetric data
- Eigenvector Centrality
  - Node connected best overall
    - Doesn't work if there are components
- Closeness Centrality
  - Node that is closest to all other nodes

Issue: Measures are highly correlated

Who Is "Key" ?



Who Is "Key" ?



<Your Name>

Betweenness Centrality

- Frequency with which a node lies along the shortest path between two other nodes
- Computed as: $b_k = \sum_{i,j} \frac{g_{ikj}}{g_{ij}}$

  where g_{ij} is number of geodesic paths from i to j and g_{ikj} is number of those paths that pass through k
- Index of potential for gate-keeping, brokering, controlling the flow, and also of liaising otherwise separate parts of the network
- Interpreted as indicating power and access to diversity of what flows; potential for synthesizing
- Sometimes interpreted as "connecting" groups
- Very "expensive" to compute

Closeness Centrality

- Measured as:
  - Sum of distances to all other nodes
  - Computed as marginals of symmetric geodesic distance matrix
- Closeness is an inverse measure of centrality
- Index of expected time until arrival for given node of whatever is flowing through the network
  - Gossip network: central player hears things first

Eigenvector Centrality

- A node will have a high score if it is connected to many nodes that are themselves highly connected
- Computed as:

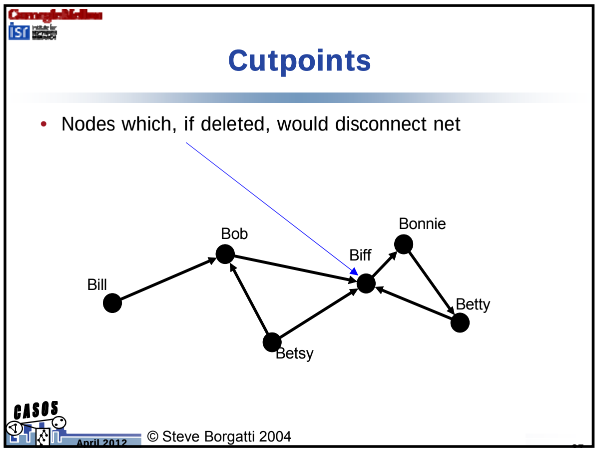
$$\lambda v = Av$$

  where A is adjacency network and V is eigenvector centrality. V is the principal eigenvector of A

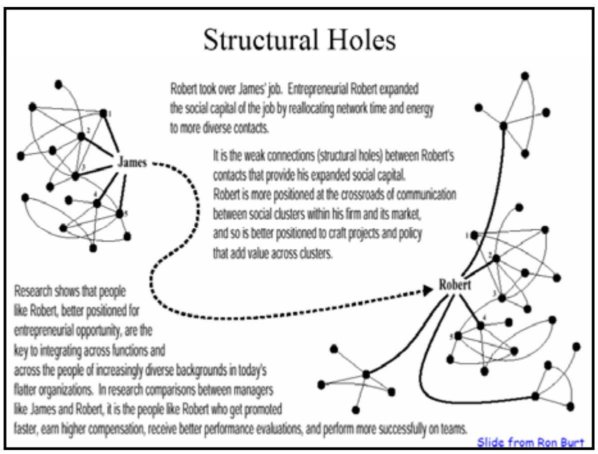
- Indicator of popularity and group-bonding
- Like degree, this is an index of exposure, risk
- Tends to identify centers of large cliques
- Often identified as leader of self-contained group, sometimes referred to as leader of leaders
- Very "expensive" to compute

Adapted from Steve Borgatti 2004

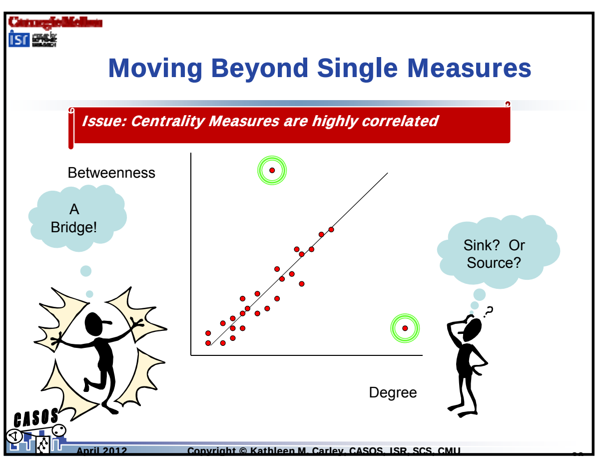
<Your Name>



Carnegie Mellon
Cutpoints
Nodes which, if deleted, would disconnect net
Bob
Bonnie
Biff
Bill
Betty
Betsy
CASOS
April 2012
© Steve Borgatti 2004



Structural Holes
Robert took over James' job. Entrepreneurial Robert expanded the social capital of the job by reallocating network time and energy to more diverse contacts.
It is the weak connections (structural holes) between Robert's contacts that provide his expanded social capital. Robert is more positioned at the crossroads of communication between social clusters within his firm and its market, and so is better positioned to craft projects and policy that add value across clusters.
James
Robert
Research shows that people like Robert, better positioned for entrepreneurial opportunity, are the key to integrating across functions and across the people of increasingly diverse backgrounds in today's flatter organizations. In research comparisons between managers like James and Robert, it is the people like Robert who get promoted faster, earn higher compensation, receive better performance evaluations, and perform more successfully on teams.
Slide from Ron Burt



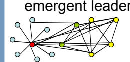
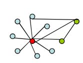
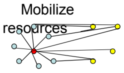
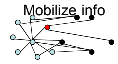
Carnegie Mellon
Moving Beyond Single Measures
Issue: Centrality Measures are highly correlated
Betweenness
A Bridge!
Sink? Or Source?
Degree
CASOS
April 2012
Copyright © Kathleen M. Carley, CASOS, ISR, SCS, CMU

SNA Insufficient

- Centralities
  - Communication
    - Degree – most connected
    - Betweenness – most paths
- Exclusivities
  - Expertise
    - Knowledge – special expertise
    - Task – special experience
- Demands/Loads
  - Roles
    - Cognitive demand – emergent leader
    - Workload

Figure 3 Trusted Prior Contacts + Meeting Ties [shortcuts]

Meta-Network KEY ACTORS

	Cognitive Demand emergent leader 	Task exclusivity critical ability 
	Resource exclusivity Mobilize resources 	Knowledge exclusivity Mobilize info 

Assignment Network
Assignment Redundancy

- Average number of redundant agents assigned to tasks. An agent is redundant if there is already an agent assigned to the task.
- Redundancy occurs only when more than one agent is assigned to a task. Define the assignment redundancy for task j as follows:,

$$d_j = \max\{0, sum(AT(:,j)) - 1\} \quad 1 \le j \le |T|$$

- Then Assignment Redundancy = $\left(\sum_{j=1}^{|T|} d_j\right) / |T|$

<Your Name>

Knowledge Exclusivity Index

- Detects agents who have singular knowledge.
- The Knowledge Exclusivity Index (KEI) for agent i is defined as follows:

$$\sum_{j=1}^{|K|} AK(i,j) * e^{(1-sum(AK(:,j)))}$$

- The values are then normalized to be in [0,1] by dividing by the maximum KEI value.

Additional Specialized Measures Exist Particularly Ones Using Multiple Matrices

- Performance
  - Diffusion
  - Accuracy
- Loads
  - Cognitive demand
  - Workload
  - Potential Work Load
- Congruency – fit
  - Communication
  - Knowledge
  - Resource
- Need for Negotiation
- Under Supply

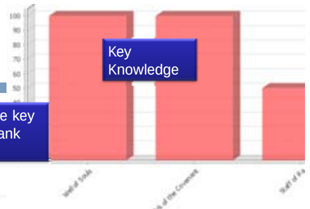
Cognitive Demand

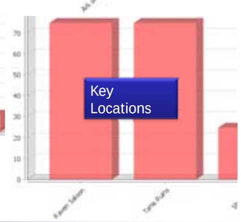
- *The cognitive effort the individual has to do on average*
- How many people do you interact with CENTRALITY
- How many tasks do you do
- How much knowledge do you have
- How much knowledge is needed to do the tasks
- How many people do you need to interact with to do the tasks
- How many other tasks and so people depend on you
- How many other tasks and so people do you depend on

Carnegie Mellon

Key Entities

Indiana Jones & Marion Ravenwood are key characters; Rene Belloq & Salah also rank high in importance.





Carnegie Mellon

Key Entities: Resources

Dominant Resource (total degree centrality)

The Total Degree Centrality of a node is the normalized sum of its row and column degrees. Individuals or organizations who are "in the know" are those who are linked to many others and so, by virtue of their position have access to the ideas, thoughts, beliefs of many others. Individuals who are "in the know" are identified by degree centrality in the relevant social network. Those who are ranked high on this metrics have more connections to others in the same network. The scientific name of this measure is total degree centrality and it is calculated on the agent by agent matrices.

Input: all networks based on the node class(es) Resource

Rank	Resource	Value	Unscaled
1	Ark of the Covenant	0.277	72.000
2	truck	0.127	33.000
3	bullwhip	0.123	32.000
4	torch	0.112	29.000
5	Headpiece for Staff - Ravenwood's half	0.112	29.000
6	pistol	0.104	27.000
7	machine gun	0.085	22.000
8	fire	0.081	21.000
9	rope	0.081	21.000
10	car	0.069	18.000

Ark of the Covenant is a dominant resource.

Carnegie Mellon

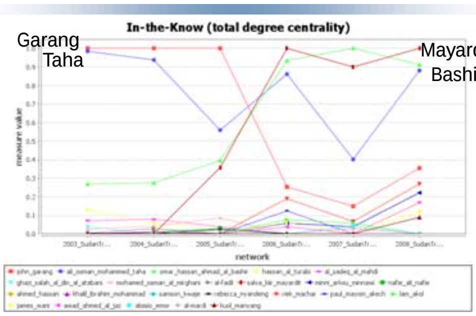
What Have We Learned?

- Indy is an important character, given a variety of relevant measures
  - Indy ranked in top 3 in 94% of measures calculated
  - Marion Ravenwood, Sallah, & Rene Belloq are also important (i.e., top-ranked in a high percentage of measures)
  - German Agents, while identified as important, is an entity that represents various extras who wore Nazi uniforms in bit parts
- Knowing of the Well of Souls & the Ark of the Covenant is important
- The Ark of the Covenant is the most important resource in the movie
- The Raven Saloon & Tanis Ruins are important locations

<Your Name>

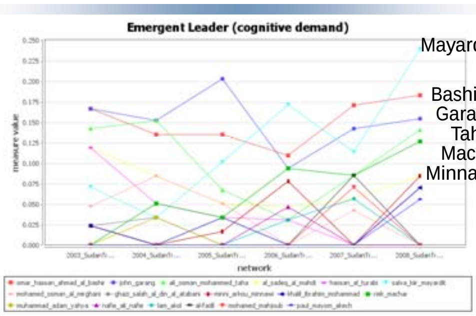
Sudan

Total Degree Centrality for Political Agents of Sudan in 2003-2008 (Sudanese only)



10/21/2010 Data Extraction

Cognitive Demand for Political Agents of Sudan in 2003-2008 (Sudanese only)



10/21/2010 Data Extraction

Sudan – Key Actors

Rank	Degree	Betweenness	Eigenvector
1	omar_al_bashir	omar_al_bashir	omar_al_bashir
2	john_garang	john_garang	salva_kiir_mayardit
3	george_w_bush	george_w_bush	john_garang
4	salva_kiir_mayardit	salva_kiir_mayardit	luis_moreno_ocampo
5	yoweri_museveni	mustafa_fadhil	ali_osman_taha
6	ali_osman_taha	saddam_hussein	george_w_bush
7	joseph_kony	keith_richards	yoweri_museveni
8	kofi_annan	barack_obama	hosni_mubarak
9	barack_obama	ali_osman_taha	joseph_kony
10	hosni_mubarak	usama_bin_laden	thabo_mbeki

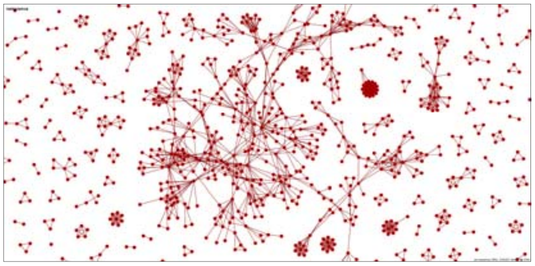
Findings on Sudan

- At the macro-level – little change
- Tribal interactions have coalesced into more formal consistent lines of alliances and conflict
- Bashir's influence increased between 2003-2008 and Minnawi shows as an emergent leader
- Political "brokers" continually changing – situational volatility
- Harbored terrorists show as key actors only from a "global" perspective
- Rise in power of Dinka
- Conflict logic changes to enable creation of S. Sudan

2003	Ecology, Land_Water_Use
2004	Land_Water_Use, Ideology
2005	Ideology, Economy
2006	Ideology, Economy
2007	Economy, Ideology
2008	Land_Water_Use, Economy

Groups!

Motivation



What is a Group?

- Any number of entities considered as a unit
- Nominal group – "named" collective e.g., nurses
- Collection of entities with features in common
- Small Group
  - 3-15 members
  - Able to communicate freely and openly with all of the other members of the group
  - Norms
  - Roles
  - Common purpose

Group Rationales

3 conceptual reasons for why groups matter

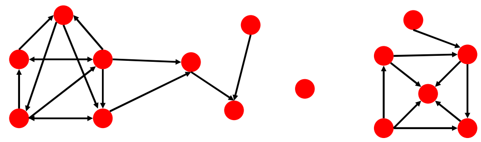
- Cohesion
  - Because the nodes have the same kind of position – relations to same type of other nodes
  - Network region might contain cohesive subgroups
- Equivalence
  - Because the nodes have the same linkages = relationships to the same other nodes
- Distinction
  - Because the nodes are different from other nodes around them, anomalies

NOTE: A group may or may not be a component or a K-Core

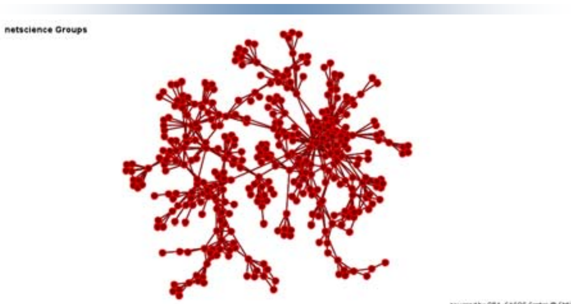
<Your Name>

Terminology: Components

- A subgraph S of a graph G is a component if S is maximal and connected
- If G is a digraph, then
  - S is a weak component if it is a component of the underlying (undirected) graph
  - S is a strong component if for all dyads u,v in S, there is a path from u to v
- Finding components is the first step in analysis of large graphs
  - Analyze each component separately, or discard very small components

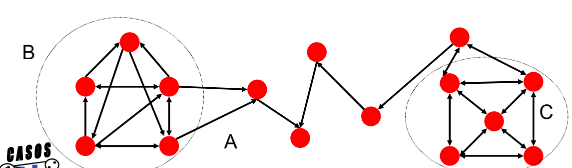


Largest Component



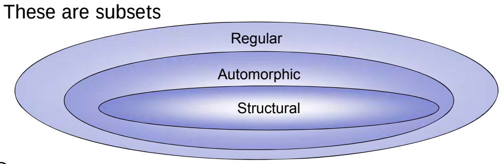
Terminology: K-Cores

- K-CORE
  - A maximal subgraph S such that for all u in S, $\alpha(u,S) >= k$
    - S=A is 1-core & 2-core; B and C 3-core
    - There is no 4-core or higher
  - Finds large regions within which cohesive subgroups may be found
  - Identifies fault lines across which cohesive subgroups do not span



Groups and Equivalences

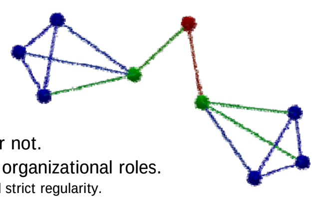
- Many grouping mechanisms are based on equivalences
- Common ones:
  - Structural
  - Regular
  - Automorphic *At least as defined in JMS paper in 1994.
- These are subsets



Structurally Similar Groups

- Classes of node connecting to the same other classes.

0	1	1	0	0	0	0	0	0
1	0	0	1	1	1	0	0	0
1	0	0	0	0	0	1	1	1
0	1	0	1	1	1	0	0	0
0	1	0	1	1	1	0	0	0
0	1	0	1	1	1	0	0	0
0	0	1	0	0	0	1	1	1
0	0	1	0	0	0	1	1	1
0	0	1	0	0	0	1	1	1



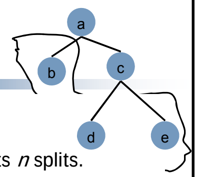
or not.

- Good for detecting organizational roles.
  - Early methods found strict regularity.
- Definition is circular!

Structural Equivalence

- Structurally indistinguishable
  - Same degree, centrality, belong to same number of cliques, etc.
  - Only the label on the node can distinguish it from those equivalent to it.
  - Perfectly substitutable: same contacts, resources
- Face the same social environment
  - Similar forces affecting them – same influencers
  - On average, hear things equally early, influenced similarly, have similar things to cope with

CONCOR



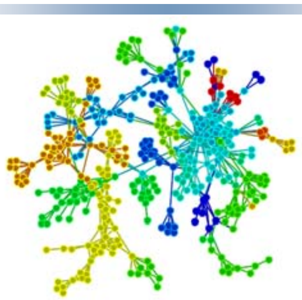
- Works by splitting groups
- Specify number of splits
- Recursively splits partitions, user selects n splits.
  - n splits → 2^n groups
- At each split, divides nodes based on maximum correlation in outgoing connections.
- Builds a hierarchical decomposition
- Calculates correlation between each pair of rows/columns
  - Then the correlation of the correlations …
  - Repeats until reaches "stableness"
  - Then splits the nodes into two groups based on the correlation

CONCOR

- Finds ZERO blocks
- Issues –
  - First correlation does most of the work
  - Heuristic approach
  - Located groups are "cliques" and often only regularly equivalent
- PRO: Only commonly used algorithm detects relaxed structural equivalence. (except arguable PCA)
- CON: Top down splitting of nodes imposes structure
- CON: Requires user to choose a power of 2 for the number of groups.

CONCOR Grouping

netscience Groups



powered by ORA, CASOS Center @ CMU

Girvan-Newman

- Detects groups
  - "community structure"
  - A community consists of a subset of nodes within which the node-node connections are dense, and the edges to nodes in other communities are less dense
- Procedure:
  - Calculate betweenness of all existing edges in the network
  - Remove edge with the highest betweenness is removed
  - Recalculate betweenness of all edges affected by the removal
  - Repeat until no edges remain
- Procedure to find optimal grouping
- Fast
- Groups sometimes difficult to interpret

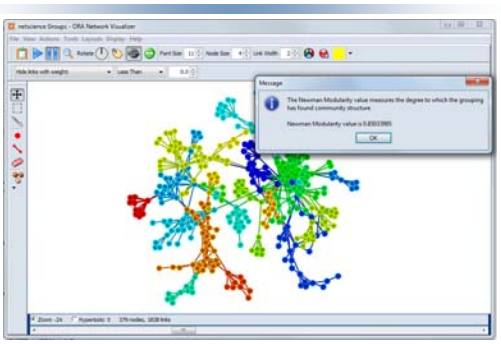
Newman's Algorithms

- Maximize Modularity (diff between expected and observed ties in community)

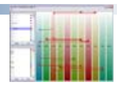
$$Q = \frac{1}{4m}\mathbf{s}^T\mathbf{B}\mathbf{s} = \frac{1}{4m}\sum_i a_i^2 \beta_i$$

- Newman & Girvan [2004]
  - Remove high-betweenness nodes
  - What's left are communities with redundant connections
  - Requires assumption of *k* groups.
- Newman [2006] :
  - Start inside community and search for boundary.
  - Relatively fast for large networks

Newman Grouping



Networks & Trails

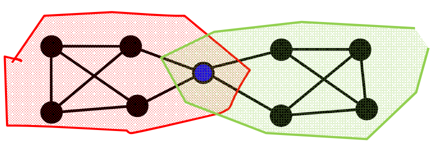


- Can be 1 or 2 mode
- Links occur only once (but can be weighted)
- Time can be modeled by multiple networks or timed attributes
- Entities can maintain many relationships in single snapshot.

- Always involves subjects and locations
- Subject may revisit location
- Time stream integral part of data.
- Subjects at one location at a time (but longer relationship may be implicit in repeated visits).

What is FOG?

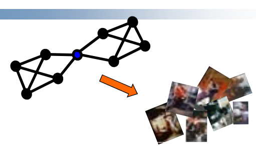
- Fuzzy, Overlapping Groups
  - Multiple group memberships
  - Varying strength of membership
  - No arbitrary assignments on boundary spanners
    - Reveals details of interstitial roles



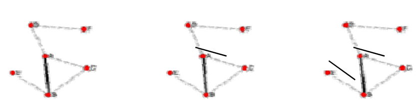
- Designed for Link Data or Network Data
- Generative model (rather than pattern matching)

Sampling Link Data From Networks

- Random tree
- Models iterative interaction
  - Informal gathering
  - Spread of rumor or info



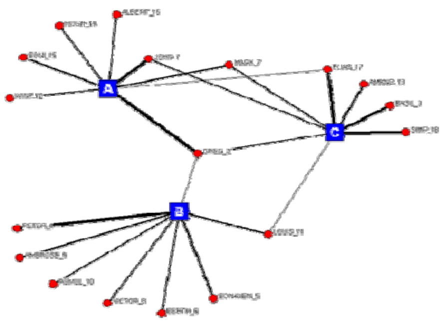
Tree



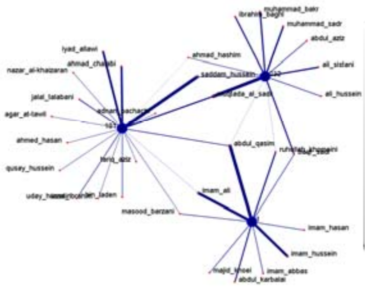
Link {A,B} {A,B,C} {A,B,C,E}

<Your Name>

FOG Groups: Monastery



Assess: How are they organized?



FOG (Fuzzy Group Clustering) shows suspicious entities organized into 5 groups w/shared members.

Interstitial members are likely to contain coordinators & leaders

FOG Algorithms

Algorithm	Based On	Pros	Cons
H-FOG	Hierarchical Clustering	• Nested Groups • Run once; explore tree to determine # of groups.	Scales poorly $O(n^4)$
k-FOG	K-Means	Scales well	Must guess # of groups, *k*
α-FOG	Dirichlet Process	Fast, Does not require guessing number of groups (α parameter is expected concentration)	Data-hungry

<Your Name>

Key Actors are Interstitial

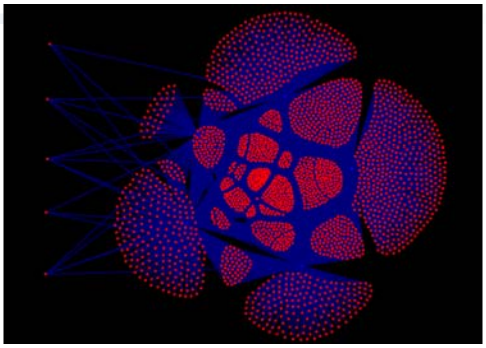
Jeff Skilling

Kenneth Lay

Tanya Jones

Veronica Espinoza

Jeff Dasovich



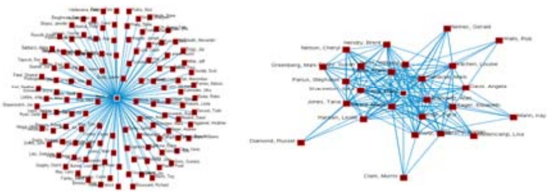
Summary

- Why Group?
  - Reconstruct “real” groups
  - Find individuals who might be or act similarly
  - Find individuals who have unusual community ties/
- CONCOR: Structural Similarity
  - Finds groups with similar roles in network, even if dispersed
- Newman: Cohesive Communities
  - Finds unusually dense clusters, even in large networks
- FOG: Fuzzy, Overlapping Groups
  - Gives better understanding of individuals spanning groups
  - Analyzes network data or raw link data

Motifs

- Specialized patterns

Two Types of HUbbs



Iris Mack

Similar for
2 other

Other 5 are in
a concentrated
core

Illustrative Problem

Gang Interaction

- Two gangs are in the regions
- Recent reports suggest they are working together
- Potential problems
  - Increased drug trafficking
  - Coordinated response to law enforcement activity
  - More dangerous
- *What is the connection?*
- *Who is connecting them?*

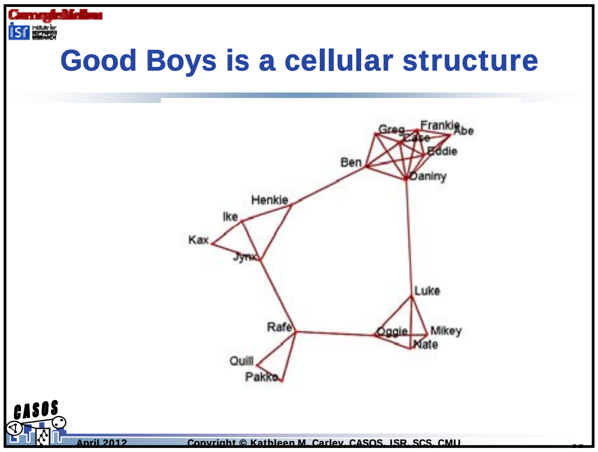
Are Two Critical Actors Linked? Path Finder

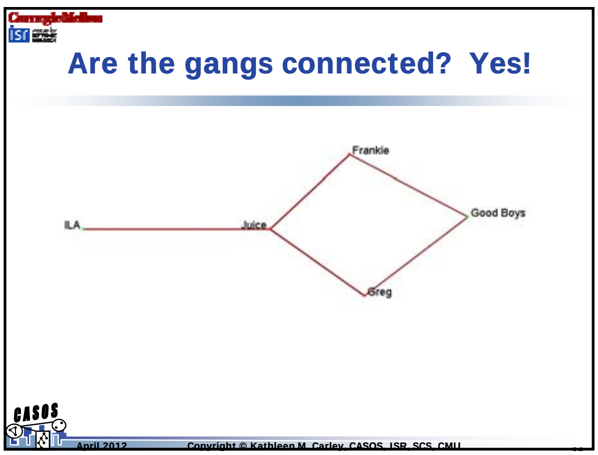
Example

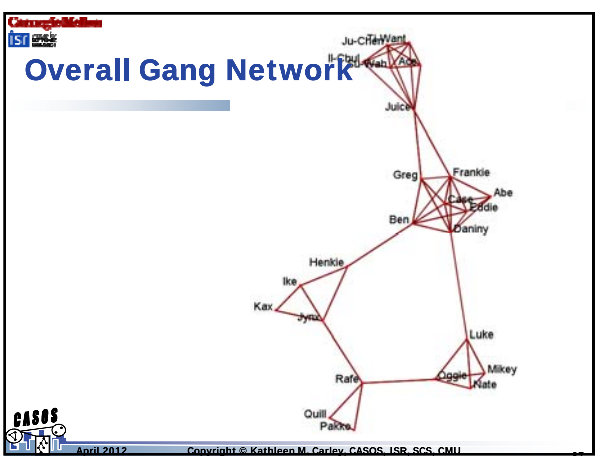
A Tale of Two Gangs

Two Gangs

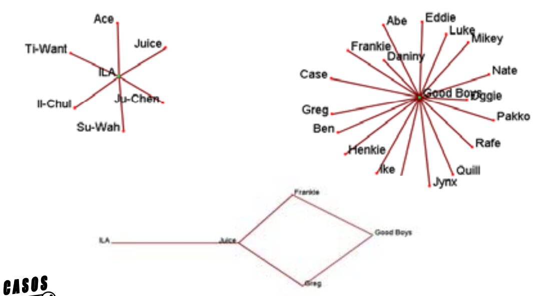
ILA is a tightly structured cell


Good Boys is a cellular structure
Greg
Frankie
Abe
Case
Eddie
Ben
Daniny
Henkie
Ike
Kax
Jynx
Luke
Rafe
Oggie
Mikey
Nate
Quill
Pakko
April 2012
Copyright © Kathleen M. Carley, CASOS, ISR, SCS, CMU


Are the gangs connected? Yes!
Frankie
ILA
Juice
Good Boys
Greg
April 2012
Copyright © Kathleen M. Carley, CASOS, ISR, SCS, CMU


Overall Gang Network
Juice
Greg
Frankie
Abe
Case
Eddie
Ben
Daniny
Henkie
Ike
Kax
Jynx
Luke
Rafe
Oggie
Mikey
Nate
Quill
Pakko
April 2012
Copyright © Kathleen M. Carley, CASOS, ISR, SCS, CMU

From Gang Membership to Key Bridge



Real-world Win! – The Context

- CASOS was training members of Tulsa Police Department Special Investigations Division and Oklahoma Bureau of Narcotics with support from the North Texas HIDTA
- Unknown to the CASOS team
  - The gang unit of a major-city police force had evidence of drug supply-chain that indicated some abnormal cooperation between two dissimilar street-gangs
  - Law enforcement had no leads on who was behind the connection
- The CASOS team was demonstrating ORA
  - to a few officers
  - utilizing a small sample of data from the department's live arrest and surveillance database

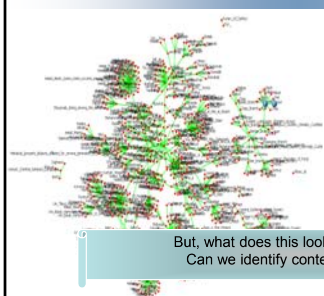
Real-world Win! Event!

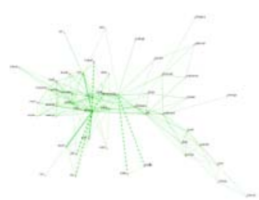
- During the demonstration of ORA's pathfinder feature
  - Officers asked the CASOS team to show the path between two gangs
- Voila!
  - Using ORA a human connection between the two gangs was quickly found
  - This human had not been readily-apparent to the officers
  - But the information was buried in their database
- The investigating officer was called into the room to see the newly discovered finding
- This first link proved to be useful!
- In addition
  - Other links have been found in follow-up analyses
  - Gang investigators from his Special Investigations Division are following these additional leads

Network Comparison

- Are two networks similar?
- What is the difference of two networks?
- How to compare more than two networks?
- How to compare predicted networks to the actual future observed networks?
- Can we use standard statistics (e.g. correlations)?

Compare Networks: Basic Tribal Structure





But, what does this look like on a map?
Can we identify contested borders?

Sudan Tribal Network - Hereditary

Tribal Network in the News

Levels of Comparison

- Node measures:
  - key entities
  - rank of entities
- Network measures:
  - distribution of node measure values
  - network centralization, density, ...
- Network structure:
  - Hamming, Euclidean
  - Correlation of networks – QAP
- "Motifs":
  - local effects (transitivity, reciprocity,...)

<Your Name>

Comparison Report: Nodes

In-the-Know (total degree centrality)

The Total Degree Centrality of a node is the normalized sum of its row and column degrees. Individuals or organizations who are "in the know" are those who are linked to many others and so, by virtue of their position have access to the ideas, thoughts, beliefs of many others. Individuals who are "in the know" are identified by degree centrality in the relevant social network. Those who are ranked high on this metrics have more connections to others in the same network. The scientific name of this measure is total degree centrality and it is calculated on the agent by agent matrices.

Input network(s): eAgent2eAgent

Rank	2008/09/01 00:00:00		2008/09/02 00:00:00		2008/09/03 00:00:00	
1	Agent56	0.110	Agent28	0.277	Agent11	0.175
2	Agent12	0.105	Agent17	0.131	Agent26	0.175
3	Agent60	0.093	Agent21	0.107	Agent7	0.135
4	Agent51	0.087	Agent59	0.102	Agent4	0.127
5	Agent3	0.081	Agent66	0.092	Agent13	0.127
6	Agent18	0.076	Agent80	0.092	Agent28	0.127
7	Agent1	0.064	[illegible]	0.092	Agent1	0.103
8	Agent2	0.058	Agent2	0.087	Agent22	0.095
9	Agent15	0.058	Agent1	0.083	Agent6	0.079
10	Agent30	0.058	Agent73	0.083	Agent11	0.071

Back to measure list

Number of Cliques (clique count)

Comparison Report: Networks

Performance Measures

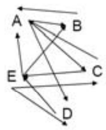
Measure	interaction_tp01	interaction_tp02	interaction_tp03
Overall Complexity			
	0.078	0.080	0.073
The density of the meta-network as a whole.			
Social Density			
agent x agent	0.07777778	0.08000000	0.07333333
Density of the Agent x Agent network.			
Social Fragmentation			
agent x agent	0.000	0.000	0.000
Fragmentation of the Agent x Agent network.			
Avg Communication Speed			
agent x agent	0.282	0.259	0.237
The average speed with which any two nodes can interact. This is based on the inverse of the shortest path lengths between node pairs.			

Simple Measures of Differences

- Hamming distance
  - Sum of differences between networks
  - For binary edges, how many differences/changes between two networks
  - Simple, intuitive, meaningful for binary data
- Correlation
  - Calculate the correlation between the edge values in two networks
  - Useful in standard statistics for independent identical samples
  - Doesn't mean much with binary data

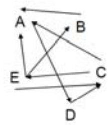
Hamming Distance

1) Picture



2) Matrix

01110
10000
10001
00001
11110



00010
10000
10001
00100
11100

3) String

0111010000100010000111110
0001010000100010010011100

4) Calculate

Distance = 5
5/20, .25, 25%

The Problem with Statistics on Networks

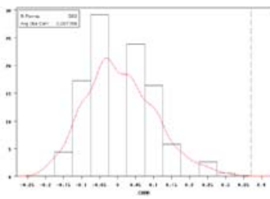
- There are row/column dependency's
- In other words – each entry is a dyad and dyads are not independent
- The basic assumptions of standard statistics are violated:
  - Independent
  - Identically distributed
- Why does this matter?
  - Statistical hypothesis tests require these pre-conditions
  - Statistical guarantees (p-values) don't hold
- We need a better significance value!

QAP / MRQAP

- Problem: Did the network structure cause similarity or did the identity of the nodes?
- Solution: QAP Permutations - quadratic assignment procedure
- QAP tests an arbitrary graph-level statistic against a QAP null hypothesis, via Monte Carlo simulation of likelihood quantiles

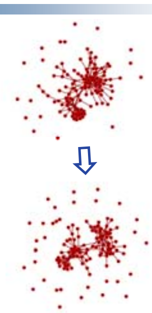
QAP Algorithm

- Randomly permute the dependent network
- Calculate the observed correlation with the independent network
- Instead of all possible N! permutations just n random samples
- In how many of the runs was the observed >= new?
- That's the approximate p-value (significance)



Example in ORA

- Predicted/Observed Networks
- Two networks:
  - Week 2: eAgent2eAgent D08
  - Week 3: eAgent2eAgent D15
- The networks are different
- What happened?
- Hypotheses:
  - a) Caused by transitivity
  - b) Caused by reciprocity

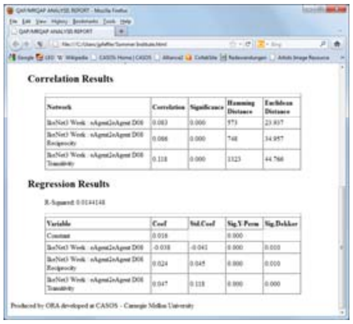


Overtime

- Node level
- Group level
- Structural level

- What – If – immediate impact
- Trends
- Change Detection
- Forecasting – Near term

QAP / MRQAP



Correlation Results

Network	Correlation	Significance	Hamming Distance	Euclidean Distance
[illegible]	0.083	0.000	573	23.937
[illegible] Reciprocity	0.066	0.000	748	34.957
[illegible] Transitivity	0.118	0.000	1323	44.766

Regression Results

R-Squared: 0.0144148

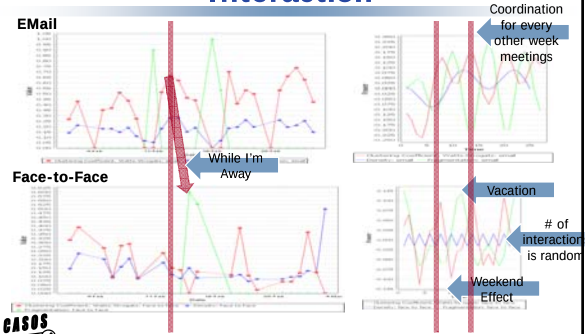
Variable	Coef	Std.Coef	Sig.Y Perm	Sig.Dekker
Constant	0.016		0.000	
[illegible]	-0.038	-0.042	0.000	0.000
[illegible] Reciprocity	0.024	0.045	0.000	0.000
[illegible] Transitivity	0.047	0.118	0.000	0.000

Produced by ORA developed at CASOS - Carnegie Mellon University

Reports with Multiple Networks

Temporal Report

Trail Differences Due to Mode of Interaction



What happens if a change occurs?

- What if
  - You fire someone
  - A group of people retire
  - You arrest members of a cell
  - You use up a resource
- There are two key questions
  - What happens immediately?
  - What will happen after the dust settles in the near term?
- The **Immediate Impact** Report helps answer what happens immediately
- **Near-Term Analysis** helps answer what happens after the dust settles

Purpose of Immediate Impact Report

- Supports what-if analysis of strategic interventions on organizational performance & individuals within
  - Intervention = remove one or more nodes / links
  - Two types of analyses
    - Impact of n specific node removals
    - Impact of n random node removals averaged over r replications
  - Report includes network- & node-level statistics for pre- & post-intervention organizations
    - Specific node removals yield Reports that include network- & node-level measures related to individual agents, tasks, resources
    - Random node removals yield Reports that include only network-level metrics.
- Convenience relative to other Reports' comparison modes

<Your Name>

Illustrative Example

Targeting

- Which individual or group to isolate to achieve maximal effect
- How to influence
  - Are there important connections
- Who to target (vulnerabilities)
  - What groups or individuals stand out
- What is the immediate effect of a COA
  - On Diffusion, Performance, Leadership

What do these things tell us

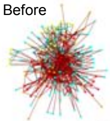
Network-level Metric	Meaning
Number of Nodes	Will go down – anchors how big is the change
Overall Complexity	Impact beyond that node – remember this is a meta-network
Performance as Accuracy	Likelihood the group will make mistakes
Diffusion	How fast does information flow through the group
Clustering Coefficient	Tendency to groupiness
Social Density	Density in the social network
Communication Congruence	The higher the more effective the group
Average Communication Speed	Typical communication speed
Number of Isolated Agents	Who's alone
Fragmentation Overall Fragmentation	Are there subgroups and level of subgroups

What do these things tell us

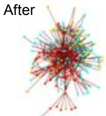
<u>Specific Node Removal Report</u> also includes 3 node-level metrics, rankings & visualization

Metric	Meaning
Emergent Leader (Cognitive Demand)	Who will be calling the shots
Potentially Influential (Betweenness Centrality)	Who will work behind the scenes
Centrality (Total Degree Centrality)	Who will know what is going on

Before



After



Dynamic Network Analysis Locates Effective Single Strike Target

Commander's Choice (in the news)

Model Choice

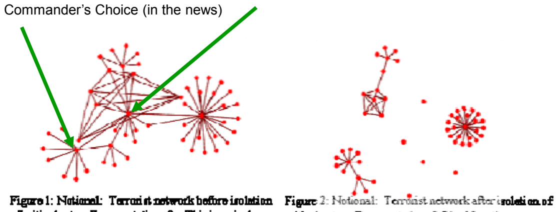
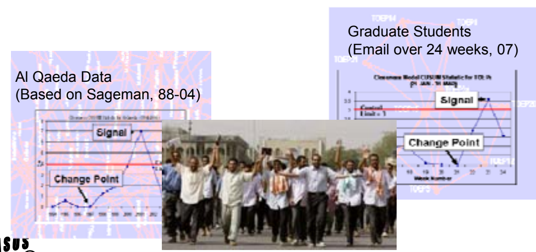


Figure 1: Notional: Terrorist network before isolation of critical actor. Fragmentation: 0 – This is a single unified organization. Diffusion: 0.93 – Messages and resources passed from actor to actor can move quickly.

Figure 2: Notional: Terrorist network after isolation of critical actor. Fragmentation: 0.71 – Now there are many isolated sub-groups. Diffusion: 0.26 – Messages or resources are passed from actor move slower.

Change Detection - Objective

How can we quickly identify changes in social networks subject to a specified risk of false alarm?



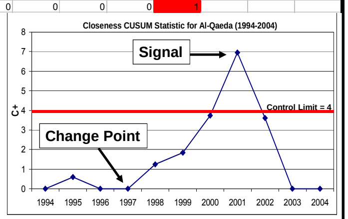
Al-Qaeda Application

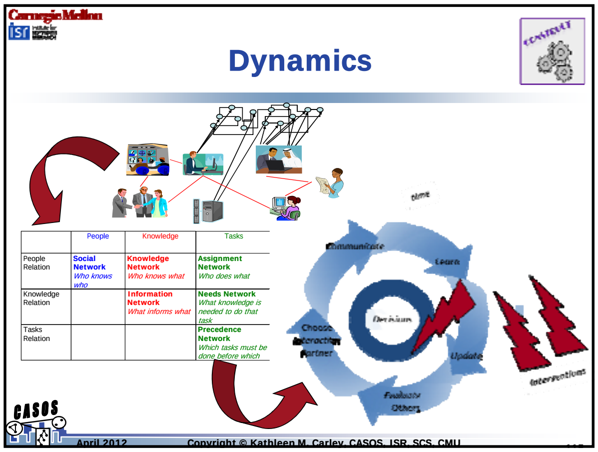
The control chart signals a change in the network in 2001.

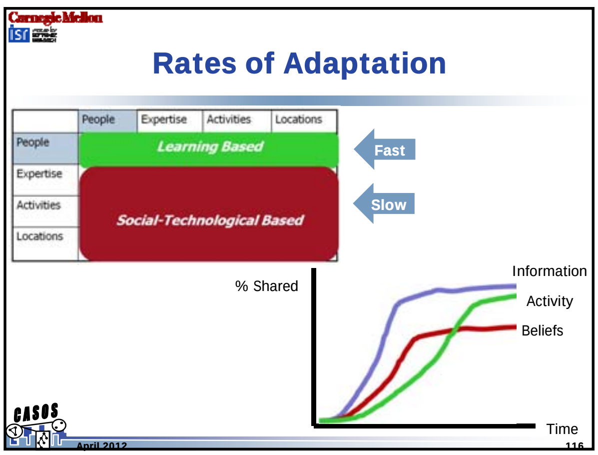
	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Closeness	0.0027	0.003	0.0028	0.0028	0.0031	0.003	0.0032	0.0034	0.0024	0.0015	0.0004
Z	-0.8729	1.0911	-0.2182	-0.2182	1.7457	1.0911	2.4004	3.7097	-2.8368	-8.7287	-15.9299
C+	0	0.5911	0	0	1.2457	1.8368	3.7372	6.9469	3.6101	0	0
C-	0.3729	0	0	0	0	0	0	0	2.3368	10.5655	25.9955
Signal	0	0	0	0	0	0	0	1			

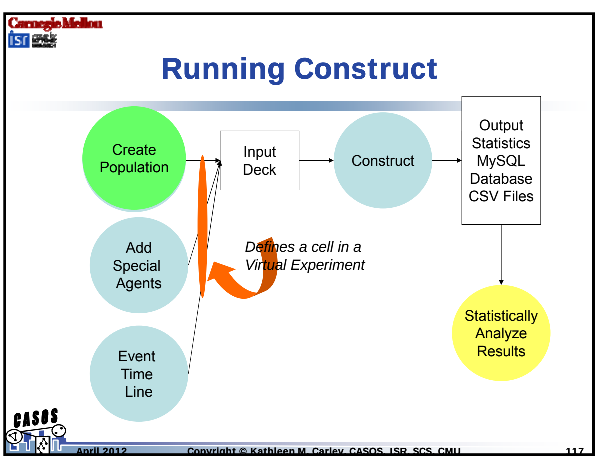
Most Likely Estimate of the change point is 1997:
- Re-establish base in Afghanistan
- Bright Star '97 cut short
- Feb '98 Islamic Front
- Embassy bombings in '98

1997 was a critical planning year for Al-Qaeda



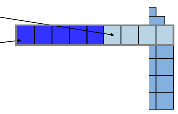

Carnegie Mellon
Dynamics
People
Knowledge
Tasks
People Relation
Social Network
Who knows who
Knowledge Network
Who knows what
Assignment Network
Who does what
Knowledge Relation
Information Network
What informs what
Needs Network
What knowledge is needed to do that task
Tasks Relation
Precedence Network
Which tasks must be done before which
time
Communicate
Learn
Decisions
Choose Interaction Partner
Update
Evaluate Others
Interventions
April 2012
Copyright © Kathleen M. Carley, CASOS, ISR, SCS, CMU


Carnegie Mellon
Rates of Adaptation
People
Expertise
Activities
Locations
Learning Based
Social-Technological Based
Fast
Slow
% Shared
Information
Activity
Beliefs
Time
April 2012
116

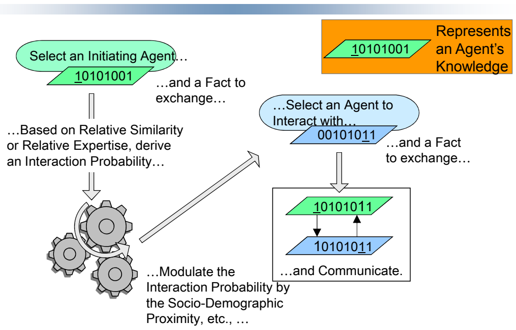

Carnegie Mellon
Running Construct
Create Population
Input Deck
Construct
Output Statistics MySQL Database CSV Files
Add Special Agents
Defines a cell in a Virtual Experiment
Event Time Line
Statistically Analyze Results
April 2012
Copyright © Kathleen M. Carley, CASOS, ISR, SCS, CMU
117

Transactive Memory

- Task memory
  - Knowledge of how to do task (the what)
  - Who knows what
- Social memory
  - Knowledge of who
  - Who knows who
- Transactive memory
  - Knowledge of who knows what, who has done what, who knows whom
  - Who knows who knows who
  - Who knows who knows what
  - AKijk = i thinks j knows k, s.t., k = #agents + # task facts

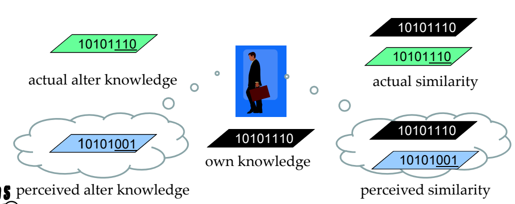


What Drives Interaction



Interaction Is Driven By Transactive Memory

- Agents have transactive memory (perceptions) of others
  - transactive memory may be inaccurate or incomplete
  - agents will not necessarily behave in an "optimal" fashion
  - tm can increase or decrease the probability of interaction

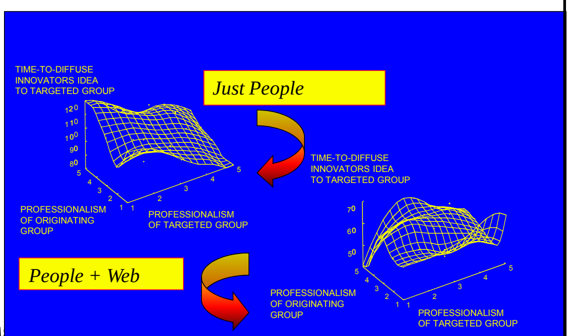


<Your Name>

Ideas Going Viral

- 2 groups
- Population 30
- Group ratio: 1:1, 1:2, 1:3
- Professionalism
  - Amount of knowledge shared by group members - 20%, 40%, 80%
  - Amount of knowledge shared across groups – 10%, 20%, 30%
  - Professionalism increases with higher in-group and lower out-group
- Mode of communication - 1:1 or 1:N+DB
- Monte Carlo 100 repetitions, 500 time periods
- Time to diffusion

Impact of Moving to a Digital Economy



Hartford Sample Network – Notice Issues of Occupation and Location

Network Size

The number of links (relations) for a particular node (agent) depends probabilistically on:

Living quarters

Work status

Gender

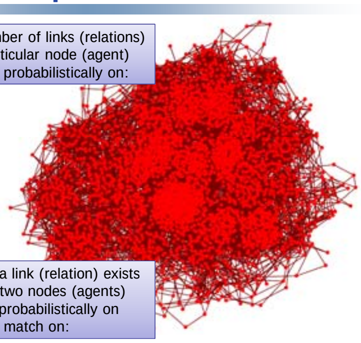
Age Education

Occupation

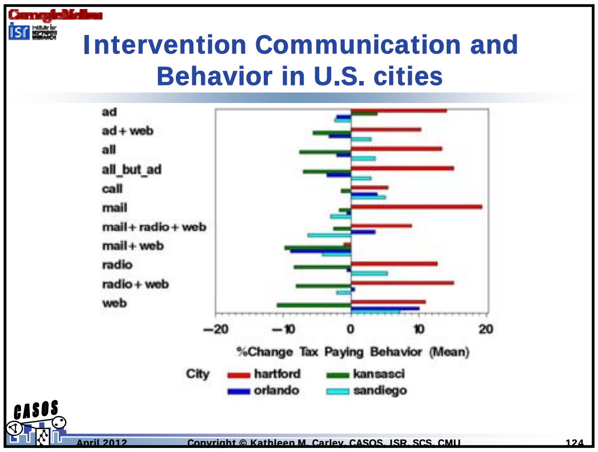
Race

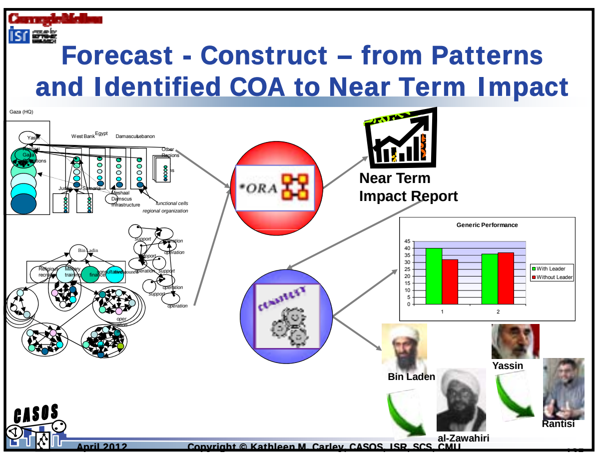
Network Composition

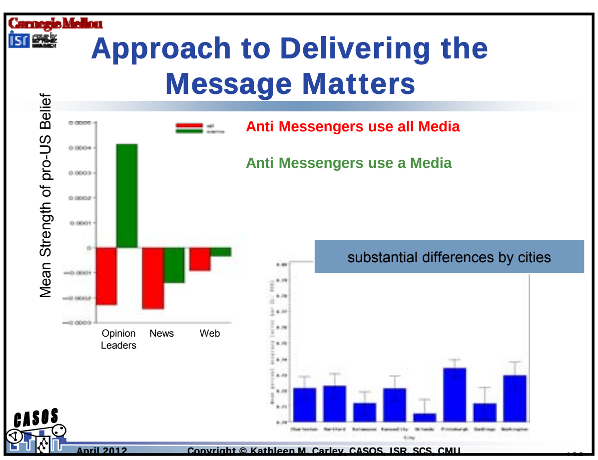
Whether a link (relation) exists between two nodes (agents) depends probabilistically on how they match on:



<Your Name>

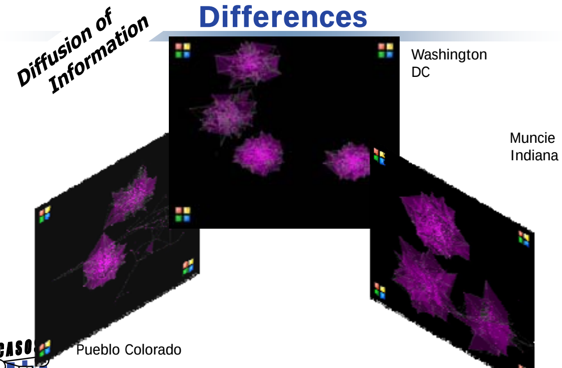

Carnegie Mellon
Intervention Communication and Behavior in U.S. cities
ad
ad+web
all
all_but_ad
call
mail
mail+radio+web
mail+web
radio
radio+web
web
−20
−10
0
10
20
%Change Tax Paying Behavior (Mean)
City
hartford
kansasci
orlando
sandiego
CASOS
April 2012
Copyright © Kathleen M. Carley, CASOS, ISR, SCS, CMU
124


Carnegie Mellon
Forecast - Construct – from Patterns and Identified COA to Near Term Impact
*ORA
Near Term
Impact Report
Generic Performance
Bin Laden
Yassin
al-Zawahiri
Rantisi
CASOS
April 2012
Copyright © Kathleen M. Carley, CASOS, ISR, SCS, CMU


Carnegie Mellon
Approach to Delivering the Message Matters
Mean Strength of pro-US Belief
Anti Messengers use all Media
Anti Messengers use a Media
Opinion Leaders
News
Web
substantial differences by cities
CASOS
April 2012
Copyright © Kathleen M. Carley, CASOS, ISR, SCS, CMU

Carnegie Mellon

Social Results Vary Due to These Differences



Carnegie Mellon

Monetary Exchange Process

- Canonical example:
  - specific dollar bill moving through the economy
- Single object in only one place at a time
- Can travel between same pair more than once
  - A--B--C--B--C--D--E--B--C--B--C ...
- This is a walk

- Use link analysis on walks in multi-mode networks
  - To find money launderers
  - To find possible covert activity

Carnegie Mellon

Gossip Process

- Canonical Example:
  - rumor moving through informal network
- Multiple copies exist simultaneously
- Person tells only one person (or a small number) at a time
- Information or good doesn't travel between same pair twice
- Information or good can reach same person multiple times
- This is a trail

- Use link analysis on trails on uni-mode networks:
  - to find rumor source
- Use link analysis on trails on uni-mode networks:
  - To identify providers of specialty items

Infection Process

- Canonical Example:
  - virus which activates effective immunological response
- Multiple copies may exist simultaneously
- Cannot revisit a node
  - A--B--C--E--D--F...
- This is a path

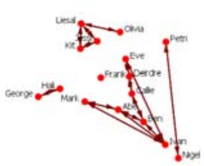
- Use link analysis on paths in uni-mode network
  - to identify key infectious individuals and points for blocking epidemic spreads
- Use link analysis on paths in multi-mode network
  - to identify key locations that might be quarantined or closed to block epidemic spreads

ORA Micro-Simulation Overview

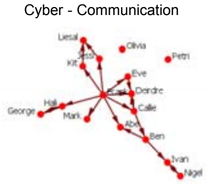
	I can give it to others	I keep it after sharing	I lose it after some time	I can get it back
Ideas	YES	YES	no	n/a
Disease	YES	YES	YES	no
Money	YES	no	no	YES
Tech	YES	YES	Sometimes	YES

Which Node is Critical?

Physical Interaction

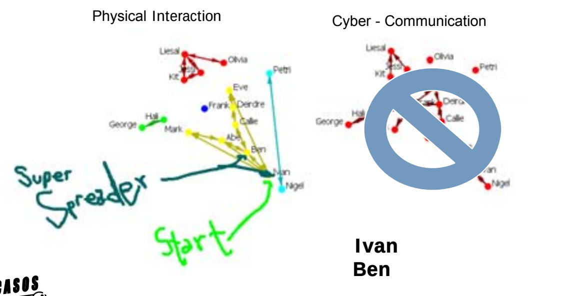


Cyber - Communication

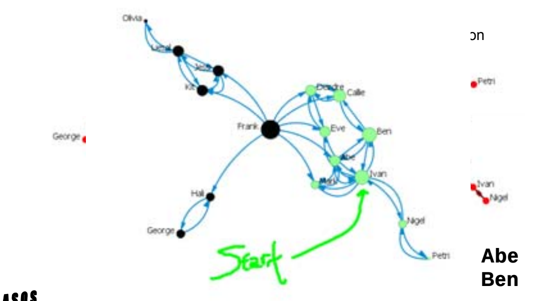


<Your Name>

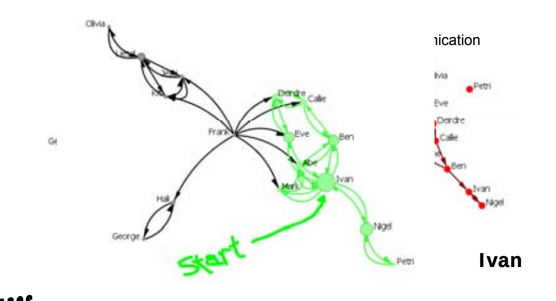
Infection



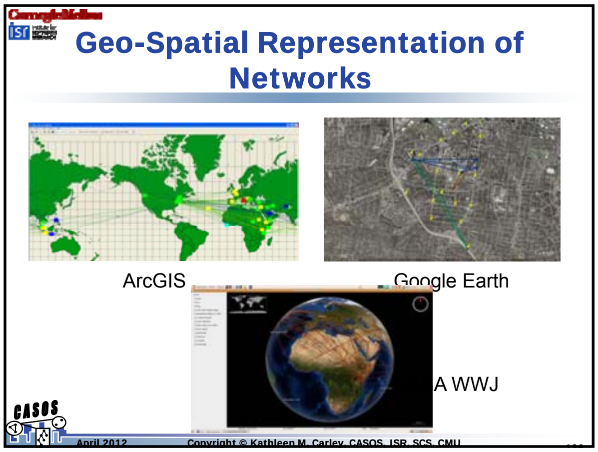
Diffusion



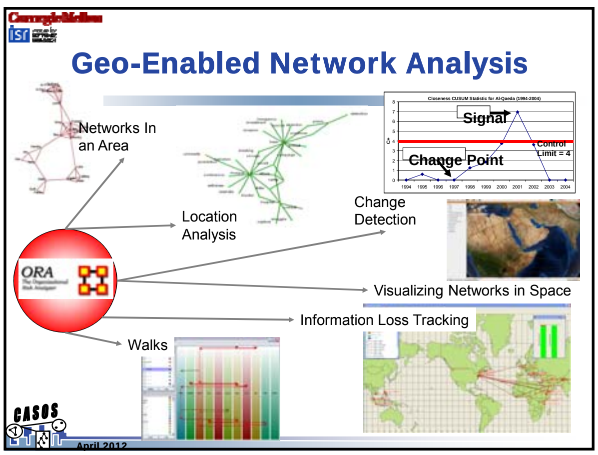
Gifting



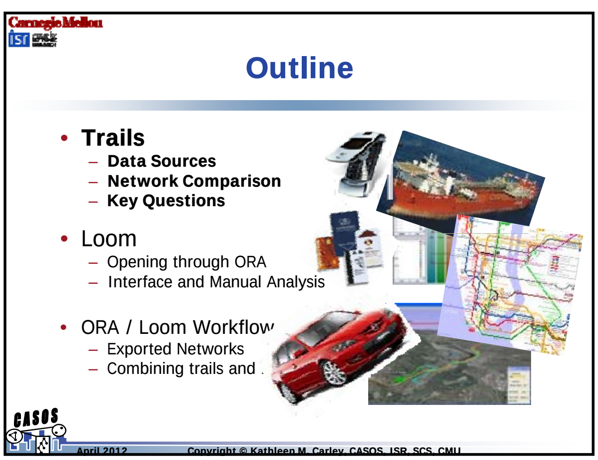
<Your Name>



Geo-Spatial Representation of Networks
ArcGIS
Google Earth
A WWJ
April 2012
Copyright © Kathleen M. Carley, CASOS, ISR, SCS, CMU



Geo-Enabled Network Analysis
Networks In an Area
Location Analysis
Signal
Change Point
Control Limit = 4
Change Detection
Visualizing Networks in Space
Information Loss Tracking
Walks
April 2012



Outline
• Trails
– Data Sources
– Network Comparison
– Key Questions
• Loom
– Opening through ORA
– Interface and Manual Analysis
• ORA / Loom Workflow
– Exported Networks
– Combining trails and
April 2012
Copyright © Kathleen M. Carley, CASOS, ISR, SCS, CMU

<Your Name>

Carnegie Mellon

Trail Data

- <Agent, Location, Time>
  - Evolving 1:Many Relation
- Sensors
  - Cameras
  - GPS
  - AIS
  - Key Card
- Mechanisms
  - Map queries
  - Address registration
  - …

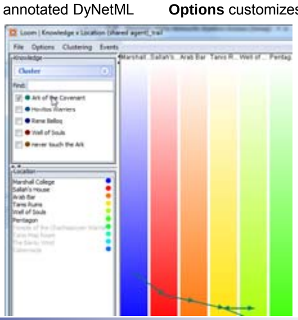
Carnegie Mellon

The Loom Interface

File exports annotated DyNetML **Options** customizes visualization

Select **Subjects** (IE agents) to view

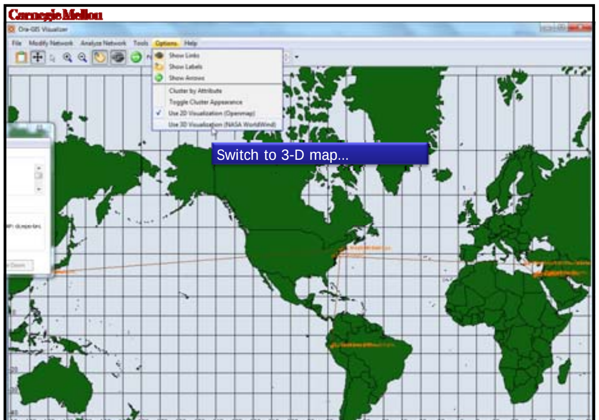
Locations are shown automatically



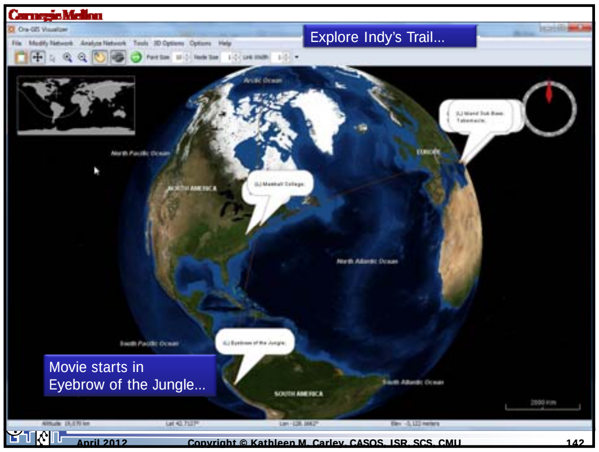
Vertical bars correspond to Locations

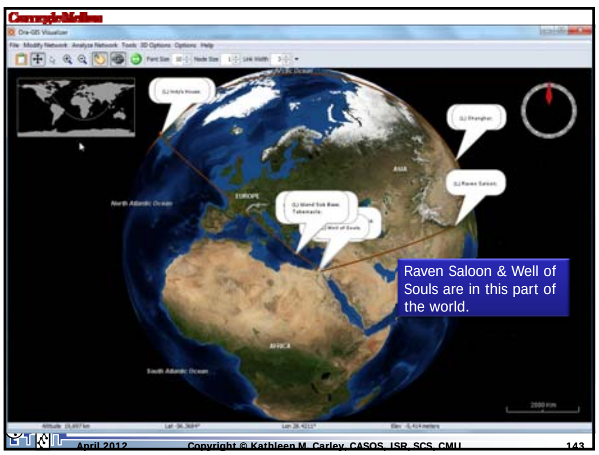
Colored lines Indicate each subject's path over time (down → later)

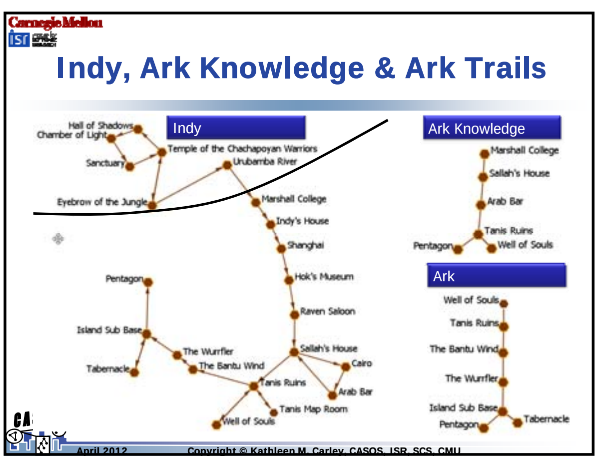
Carnegie Mellon



<Your Name>



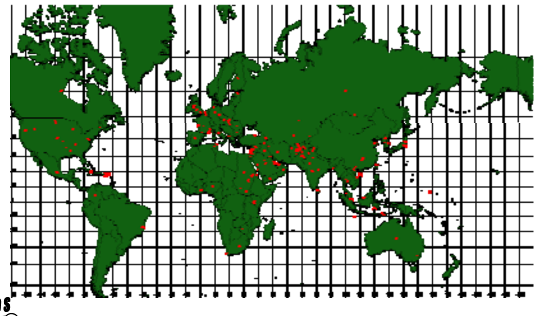




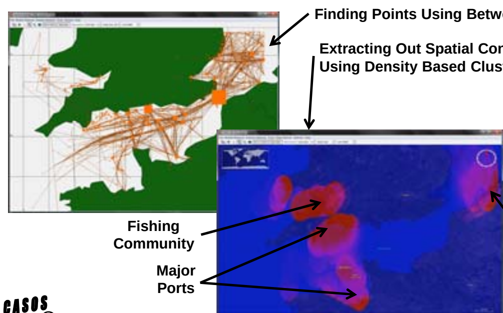
CASOS

<Your Name>

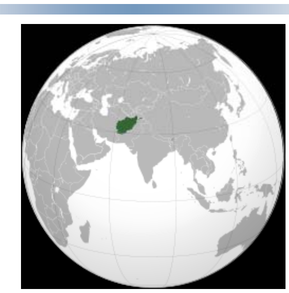
Global Reach of Highly Central Actors



Identify Critical Locations



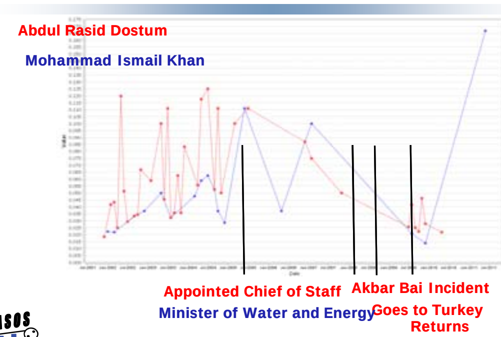
Afghanistan



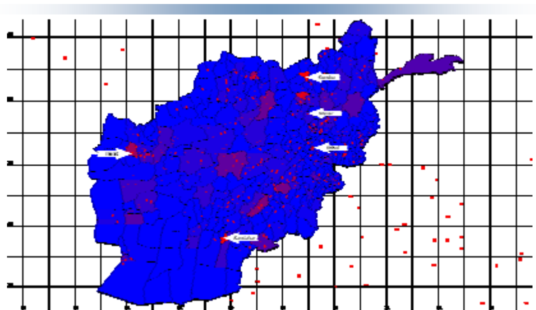
Wikipedia

<Your Name>

Ability of Kahn and Dotsum to act as an Emergent Leader Changes Over Time



Where are the people in Afghanistan Kunduz and Kandahar are hot spots



The Problem

- Traditional network measures were only designed to look at distance with respect to a single relationship.
- Using Meta-Networks we can now model multiple relationships, including geo-spatial relationships.
- This requires us to re-evaluate the concept of “distance” within a network.

<Your Name>

New Metrics

- Spatial betweenness centrality
  - Identifies individuals who allow short paths across long distances
  - Fast, efficient algorithm
- Spatial degree centrality
  - Identifies individuals that are not only high connected, but who's connections cover large geographic areas.
- Spatial eigenvector centrality
  - Identifies Locations who's agent population as a whole has the greatest eigenvector centrality
- Location Relevance
  - Identifies locations by the number of agents that are located at that location and also known to be connected to each other in the agent to agent graph

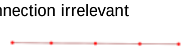
Critical Nodes as Anomalies

- Betweenness centrality
  - Nodes that connect disparate groups
  - Important in diffusion across groups
    (e.g. knowledge, illicit drugs, etc.)

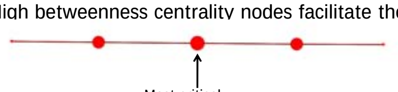
- Eigenvector centrality
  - Social capital
  - Related to PageRank
  - Important in diffusion within groups:
    fast cascades from high eigenvector centrality nodes to rest of a group

Centrality Measure: Betweenness

- Captures the extent to which a node lies along shortest paths between other nodes
  - Often interpreted as having power
  - In simple random networks, these paths get long for large networks
  - Length of connection irrelevant



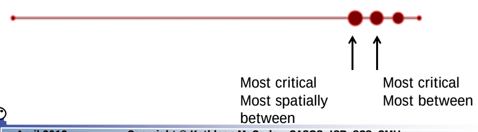
- High betweenness centrality nodes facilitate the fast



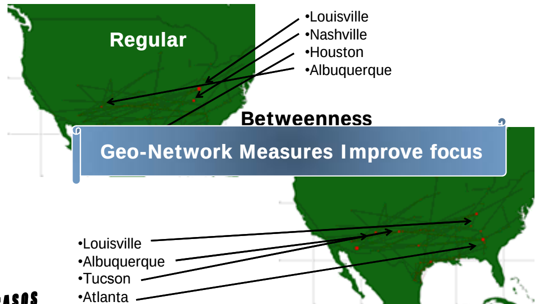
Spatial Betweenness

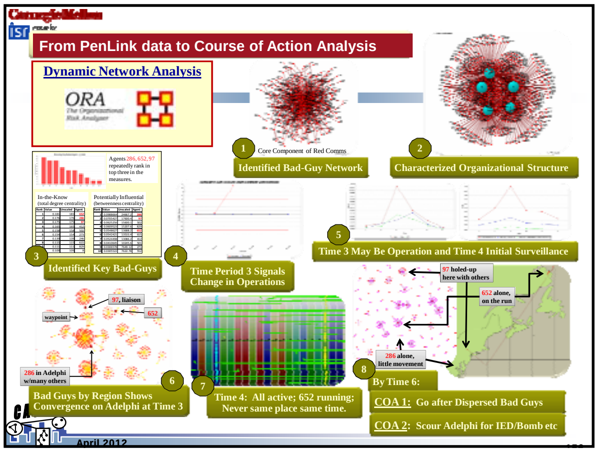
- Propinquity predicts these short paths to exist between spatially proximate nodes.

- Want a new measure to discover the nodes in the network that facilitate short paths across large distance



Using Constraints to Set Strength of Links has Practical Consequences





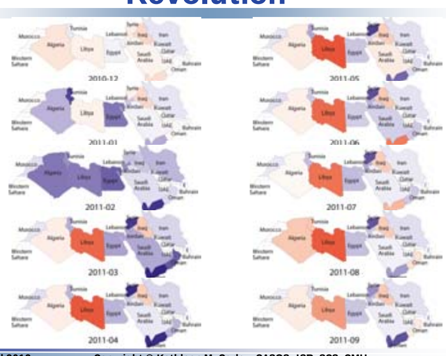
<Your Name>

Arab Spring

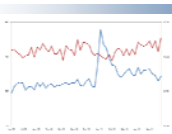
Changing Social Behavior: Terrorism and the Arab Spring

- Using rapid ethnographic assessment at multiple levels
  - High Level
    - Fast processing of lexis nexis tags
    - Examined role of social media in context
    - Fast processing of twitter hashtags
  - Medium
    - Medium processing of content – lexis nexis content
    - Twitter content
- Results
  - 18 arab spring countries for 10 months
  - Spread of revolution not geographically based
  - Mixed effect of social media
  - Conceptual complexity and human rights are dominant factors
  - Media attention to terrorism drops as revolutions begin
  - High power actors and ideas are “secondary”

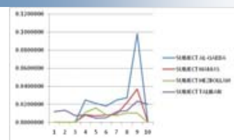
Arab Spring – Rise and Fall of Revolution



Attention to Issues During Arab Spring

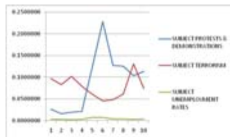


Complexity Increases



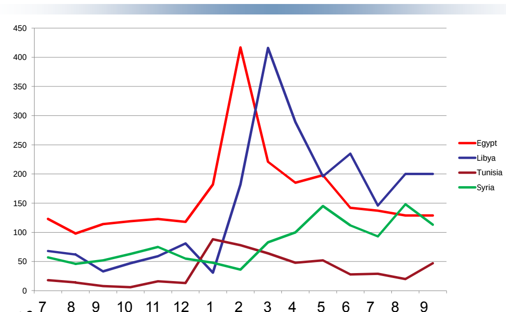
Pattern for all Terror Groups Similar

Activity Increases
32,000+ articles
2 Days



As Revolutions Gets Underway
Focus on Terror Goes Down

Actors of Interest



Familiar Secondary Actors

- Leaders
- Latent leaders – most likely to sway populace when leaders removed
- Gatekeepers
  - Betweenness
  - Even better – high betweenness low degree
  - Individuals with high structural holes
- Critical for impacting
  - Who has access to what information
  - Who gets what job
  - Etc.

<Your Name>

Key Actors- Egypt

Month	Leader	Latent Leader	Gatekeeper
July 10	Hosni Mubark	Baroness Ashton	Michael Hayden
Aug 10	Barack Obama	George J Mitchell	Asif Ali Zardari
Sep 10	Mahmoud Abbas	Hosni Mubark	Dmitry Medvedev
Oct 10	Hosni Mubark	Mohamed Elbaradei	Dmitry Medvedev
Nov 10	Hosni Mubark	Silvio Berlusconi	Muammar Gaddafi
Dec 10	Hosni Mubark	George J Mitchell	John Kerry
Jan 11	Hosni Mubark	Nicholas Sarkozy	Thaddeus G McCotter
Feb 11	Hosni Mubark	George W Bush	Wolfgang Schaeuble
Mar 11	Hosni Mubark	George W Bush	Bill Nelson
Apr 11	Hosni Mubark	Bashar al Assad	Angela Merkel
May 11	Barack Obama	George W Bush	Dick Cheney
Jun 11	Barack Obama	Christine Lagarde	Conan O'Brien
Jul 11	Hosni Mubark	Bashar al Assad	Tzipora Livini
Aug 11	Hosni Mubark	David Cameron	Joe Biden
Sep 11	Barack Obama	Hilary Rodham Clinton	Mark Zuckerberg

Key Actors - Libya

Month	Leader	Latent Leaders	Gatekeeper
July 10	Barack Obama	Hilary Rodham Clinton	Prince Philip
Aug 10	Alex Salmond	Charles Schumer	Peter Mandelson
Sep 10	Alex Salmond	Kirsten E Gillibrand	Ben Cardin
Oct 10	Mahmoud Abbas	George J Mitchell	Lee Myung-Bak
Nov 10	Nicholas Sarkozy	Angela Merkel	Benjamin Netanyahu
Dec 10	Muammar Gaddafi	Julian Assange	Sadam Hussein
Jan 11	Muammar Gaddafi	Ellen Johnson-Sirleaf	Kim Jong II
Feb 11	Muammar Gaddafi	Gordon Brown	Francois Fillon
Mar 11	Muammar Gaddafi	Robert M. Gates	Stephen Colbert
Apr 11	Muammar Gaddafi	Liam Fox	Caroline Spelman
May 11	Muammar Gaddafi	Dmitry Medvedev	Christiane Amanpour
Jun 11	Muammar Gaddafi	Liam Fox	Kevin McCarthy
Jul 11	Muammar Gaddafi	Nicolas Sarkozy	Prince William
Aug 11	Muammar Gaddafi	Nick Clegg	Dalai Lama
Sep 11	Muammar Gaddafi	Ban Ki-Moon	Al Gore

Key Issues - Egypt

Month	First	Second	Third
July 10	Oil & gas	Internat Relations	Religion
Aug 10	Religion	Internat Relations	Economic
Sep 10	Peace Process	Internat Relations	Religion
Oct 10	Religion	Internat Relations	Elections
Nov 10	Religion	Elections	Politics
Dec 10	Religion	Internat Relations	Elections
Jan 11	Protests & Demons	Religion	Internat Relations
Feb 11	Protests & Demons	Religion	Internat Relations
Mar 11	Protests & Demons	Religion	Mubarak Resignation
Apr 11	Protests & Demons	Religion	Mubarak Resignation
May 11	Religion	Terrorism	Internat Relations
Jun 11	Religion	Protests & Demons	Economic
Jul 11	Protests & Demons	Religion	Mubarak Resignation
Aug 11	Religion	Mubarak Resignation	Protests & Demons
Sep 11	Internat Relations	Religion	Protests & Demons

CASOS

<Your Name>

Key Issues - Libya

Month	First	Second	Third
July 10	Oil & gas	Investigations	Internat Relations
Aug 10	Anniversaries	Investigations	Terrorism
Sep 10	Internat Relations	Investigations	Finance
Oct 10	Internat Relations	Peace Process	Terrorism
Nov 10	Internat Econ Org	Internat Relations	Economic Growth
Dec 10	Internat Relations	WikiLeaks	Internat Econ Org
Jan 11	Sports	Internat Relations	Internat Econ Org
Feb 11			
Mar 11	War & Conflict	Internat Relations	Rebellion Insurgent
Apr 11	War & Conflict	Internat Relations	Rebellion Insurgent
May 11	War & Conflict	Internat Relations	Rebellion Insurgent
Jun 11	War & Conflict	Armed Forces	Internat Relations
Jul 11	War & Conflict	Internat Relations	Rebellion Insurgent
Aug 11	War & Conflict	Rebellion Insurgent	Armed Forces
Sep 11	War & Conflict	Internat Relations	Rebellion Insurgent

Key Betweenness Issues - Egypt

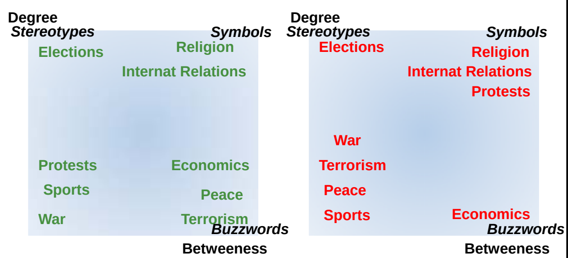
Month	First	Second	Third
July 10	Religion	Investigations	Construction
Aug 10	Religion	Terrorism	Economics
Sep 10	Religion	Deserts	Peace Process
Oct 10	Religion	Penalties	Peace Process
Nov 10	Religion	Economics	Terrorism
Dec 10	Religion	Economics	Vehicles
Jan 11	Peace Process	Religion	Terrorism
Feb 11			
Mar 11	Religion	Internat Relations	Economics
Apr 11	Economics	Anniversaries	Internat Relations
May 11	Religion	Economics	Internat Relations
Jun 11	Economics	Religion	Investigations
Jul 11	Religion	Economics	Internat Econ Org
Aug 11	Religion	Economics	Peace Process
Sep 11	Religion	Economics	Terrorism

Key Betweenness Issues - Libya

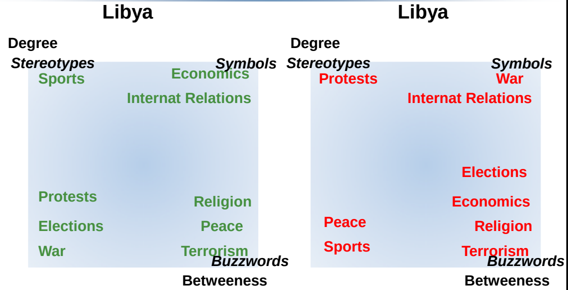
Month	First	Second	Third
July 10	Religion	Internat Relations	Terrorism
Aug 10	Religion	Economics	Investigations
Sep 10	Religion	Terrorism	Oil & Gas
Oct 10	Religion	Internat Relations	Investigations
Nov 10	Economics	Internat Relations	Terrorism
Dec 10	Internat Relations	Religion	Peace Process
Jan 11	Religion	Internat Relations	Economics
Feb 11			
Mar 11	Economics	War & Conflict	Religion
Apr 11	Economics	War & Conflict	Religion
May 11	Religion	War & Conflict	Economics
Jun 11	Religion	Economics	Internat Relations
Jul 11	Religion	Internat Relations	Armed Forces
Aug 11	Religion	Internat Relations	Armed Forces
Sep 11	Religion	Economics	Terrorism

<Your Name>

Communicative Reach



Communicative Reach



Steps in a Structural Analysis

1. Collect network data.
   – Connections among people, knowledge, resources, events …
2. Enter data into ORA.
3. Visualize.
4. Generate Report.
5. If multiple networks create combined measures.
6. If needed look at some measures more indepth.
7. Possibly drop isolates and pendants
8. Check interpretations.

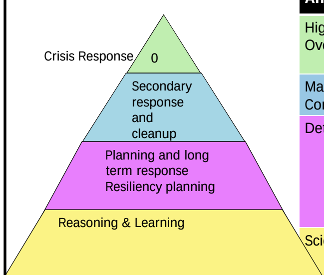
<Your Name>

STEP 1 – Data Collection
Socio-Cultural Data is Every Where

- Unstructured
  - Text – e.g. interviews, news articles, blogs, email
  - Various on-line sources
- Semi-structured
  - Blogs
  - Emails
  - Crowd-sourced
- Structured
  - Government and corporate documents
  - Proceedings

- Unstructured
  - Sudan Tribune Review
    - 2003 - 2932
    - 2004 - 6943
    - 2005 - 3828
    - 2006 - 3828
    - 2007 - 5815
    - 2008 - 9266
  - Archive of Lobban writing
- Semi-structured
  - UN Reports
  - IDA Study 1796 files
- Structured
  - African gazateer

Data Assessment Pyramid



Analysis	Method	Time
High Overview	Rapid Network Assessment	A few hours
Matches Core actors	Network Analytics	A few hours to a day
Detailed	Text Mining Machine Learning Network Analytics Simulation	A few days to a month
Scientific	Mult-method Simulation	Several months

STEP 2 – Cleaning the Corpus

- Cleaning of data
  - Removal of navigation, headers and footers (information not pertinent to the article)
  - Remove non texable files
    - E.g. remove maps/scans
  - Converted PDF to .txt
    - Most but not all can be converted with CASOS tool
  - Convert RTF to .txt
    - Currently semi-manual process
- Semi-automatic cleaning of the corpus which is done manually (optional) and automatically
  - Involves the entire corpus as a whole not individual texts
  - Removed word wrapping
  - Run automated cleaning

STEP 3 - Deduplication

- Deduplication
  - Removal of repeated articles
  - Reduces the number of files and allows a more compact analysis
  - Near Miss procedure is best
- Performed Once
- Time depends on number of texts and length
- No deduplication was done on SNARC data as all files unique

Illustration of Impact of typical Deduplication - Number of texts before and after deduplication applied only to Sudan data		
Sudan	Number before	Number of after
text	32613	18309
concepts	88260	83150
Average frequency per concept	197.417879	88.15785929

Illustration of Impact of Deduplication: Note Deduplication Can Impact Importance of Concepts

Top 10 Concepts Before and After Deduplication - Sudan

Before			After		
Concept	Count	%	Concept	Count	%
Valencia	1141455	6.55	Valencia	853691	11.65
conflict _task	867688	4.98	conflict	585801	7.99
nanuque	500411	2.87	republic_of_the_sudan	332082	4.53
ampere	448679	2.88	conflict_task	207738	2.83
republic_of_the_sudan	385560	2.21	wilayat_darfur	153976	2.1
wilayat_darfur	344036	1.97	political	113992	1.56
valence_task	178629	1.03	Sudanese	94010	1.28
ner_population	178059	1.02	Khartoum	72409	0.99
faouzi_ben_mohamed_be_ahmed_a	172782	0.99	valence_task	62440	0.85
badou	152547	0.88	environment	60138	0.82

STEP 4 – Automated Text Cleaning

- Removed stand-alone numbers
- Removed extra space
- Fixed common typos
- Removed extra white space
- Expanded contraction and abbreviations
- Removed individual letters not in names
- Converted British to American spelling
- Pronoun resolution
- Converted common hyphenated forms and non-hyphenated to common form e.g. Major-General and Major General
- Generalization using standard plus the named entities
- Removed noise words
- Ran standard ngram conversion

This can be done with AutoMap

Additional Aspects of Automated Text Cleaning

Text preparation: a completely automatic process

- Creation of a thesauri of stemmed and non-stemmed version of nouns and verbs
  - Detensing: Reduce all verbs to their present tense
  - Depluralization: Eliminates the plural form and reduces it to its base form
- Apply an n-gram thesauri to convert multi-words to single concepts
- Delete noise/stop words
  - Prepositions
  - helping verbs
  - verb of being
  - remaining pronouns

Illustrative Results for Impact of Specialized Stemmers

Number of concepts after depluralization			
	Original	After	Percentage
Sudan	97492	85758	87.96%
Catnet	24743	22091	89.28%
Singapore	5073	4452	87.76%

Number of texts nouns and verbs before and after depluralization and detensifying				
	Nouns Before	Nouns After	Verbs Before	Verbs After
Sudan	28488	23680	12006	6763
Catnet	7693	6838	4754	3223
Singapore	1677	1445	1213	816

Step 5 – Named Entity Identification

- Extraction and Identification of named entities:
  - Tag by part of speech
  - Identify proper nouns and n-grams that are proper nouns
- Result: thesauri of named entities
- Prior study showed that human processing time reduced by 80% to 99% by using this approach

Number of Named Entities					
	Texts	People	Organizations	Locations	Events
SNARC	265	33,210	5,523	2,807	120

STEP 5 – Includes Ontological Cross Classification and Thesauri Construction

- Apply standard thesauri and ontological categories
- Applying the standard thesauri coverts all multiple-concepts words into a single word/concept already classified
- Ontological classes are suggested using
  - Parts of speech and statistical regularities

Step 6: Named Entity Resolution

- At the same time that the named entities are extracted a meta-network is extracted.
- Named Entity list
  - Contains all concepts/ngrams guessed to be people, organizations, locations, events with best guess
- Meta-network
  - Contains all people, organizations, locations, knowledge, events, activities, resources, beliefs based on existing standard thesauri
  - The specific people, organizations, locations and events in this are viewed as "vetted"
- The vetted list is removed from the named entity list and the vetted class is used
- Humans then go through the remaining named entities to classify those that make sense
- Step 5 and 6 are repeated as needed
- End result is a vetted meta-network

April 2012 Copyright © Kathleen M. Carley, CASOS, ISR, SCS, CMU

Illustrative Results from Named Entity List Before Resolution

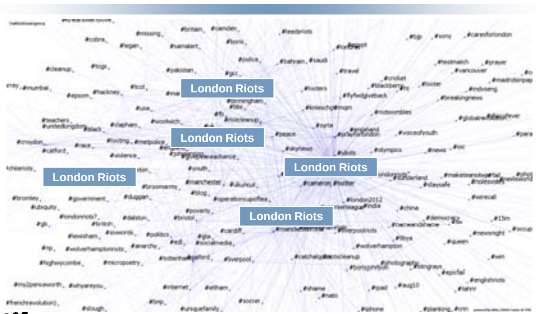
conceptFrom	conceptTo	metaOntology
up missions targeting taliban leaders	up_missions_targeting_taliban_leaders	agent
distribute new	distribute_new	agent
main office california office 1899 l street	main_office_california_office_1899_l_street	location
pay afghan	pay_afghan	agent
military commander	military	organization
once u	once_u	agent
peace press washington	peace_press_washington	agent
david katz	david_katz	agent
david kilcullen	david_kilcullen	agent
david lachapelle	david_lachapelle	agent
david lanz	david_lanz	agent

April 2012 Copyright © Kathleen M. Carley, CASOS, ISR, SCS, CMU

<Your Name>

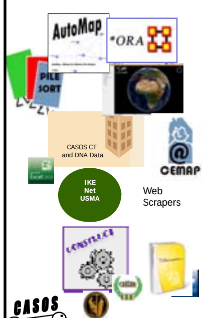
Carnegie Mellon
isr Institute for Software Research

London Riots Hashtag Network



Carnegie Mellon
isr Institute for Software Research

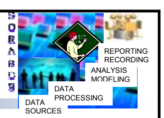
Socio-Cultural Modeling Conundrum!



- There are hundreds (maybe thousands) of services and models in the human socio-cultural behavioral area
- Some are webservices and some need to be on larger systems
- NEED:
  - Increased model reuse
  - Decreased time from data collection to model results
  - Share procedures for going from data to model to policy

Carnegie Mellon
isr Institute for Software Research

SORASCS



What is SORASCS?

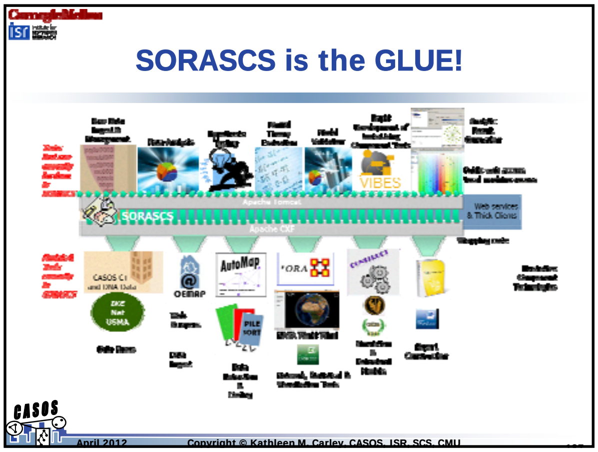
- Platform for building workflows using existing tools
  - key user – analyst
  - tools integrated: (CMU) WebScraper, ORA, Automap (Thick & Thin), PileSort, CMU-front-end for NASA World Wind (GMU) Pythia, Caesar III, (COTS) Word, Powerpoint, Excel.
- Back-end system on which tools are built
  - key user – modelers
  - VIBES (Alion/CMU) now has a SORASCS backend

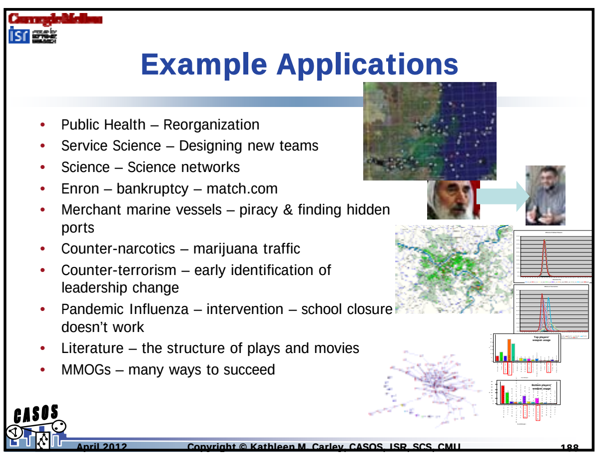
Key Features

- Workflow management (scriptrunner tool and the script2bpel), Simple interface, Integration guidelines, Data Provenance, coherent, flexible, extensible

Key Underlying Technologies

- BPEL, Apache (CXF, ODE, Tomcat), CMU Scriptrunner


SORASCS is the GLUE!
VIBES
SORASCS
Web services & Thick Clients
AutoMap
ORA
April 2012
Copyright © Kathleen M. Carley, CASOS, ISR, SCS, CMU


Example Applications
• Public Health – Reorganization
• Service Science – Designing new teams
• Science – Science networks
• Enron – bankruptcy – match.com
• Merchant marine vessels – piracy & finding hidden ports
• Counter-narcotics – marijuana traffic
• Counter-terrorism – early identification of leadership change
• Pandemic Influenza – intervention – school closure doesn't work
• Literature – the structure of plays and movies
• MMOGs – many ways to succeed
April 2012
Copyright © Kathleen M. Carley, CASOS, ISR, SCS, CMU
188