Extraction and Analysis of Cognitive Networks from Electronic Communication

NSF Symposium on Next Generation Data Mining and Cyber-Enabled Discovery for Innovation, 2007

October 10 – 12
Baltimore

Jaideep Srivastava
University of Minnesota
srivasta@cs.umn.edu

Collaborators
Nishith Pathak, Sandeep Mane, Muhammad A. Ahmad, *University of Minnesota*
Noshir S. Contractor, *Northwestern University*
Dmitri Williams, *University of Southern California*

Sponsorship
National Science Foundation
University of Minnesota

Outline

- Introduction
- Modelling a cognitive social network
- Quantitative measures for perceptual closeness
- Experiments with the Enron dataset
- Extracting concealed relationships
- Mining MMORPG logs for social science research
- Conclusion

Introduction

Social Networks

- A **social network** is a social structure of people, related (directly or indirectly) to each other through a common relation or interest
- **Social network analysis (SNA)** is the study of social networks to understand their structure and behavior

(Source: Freeman, 2000)

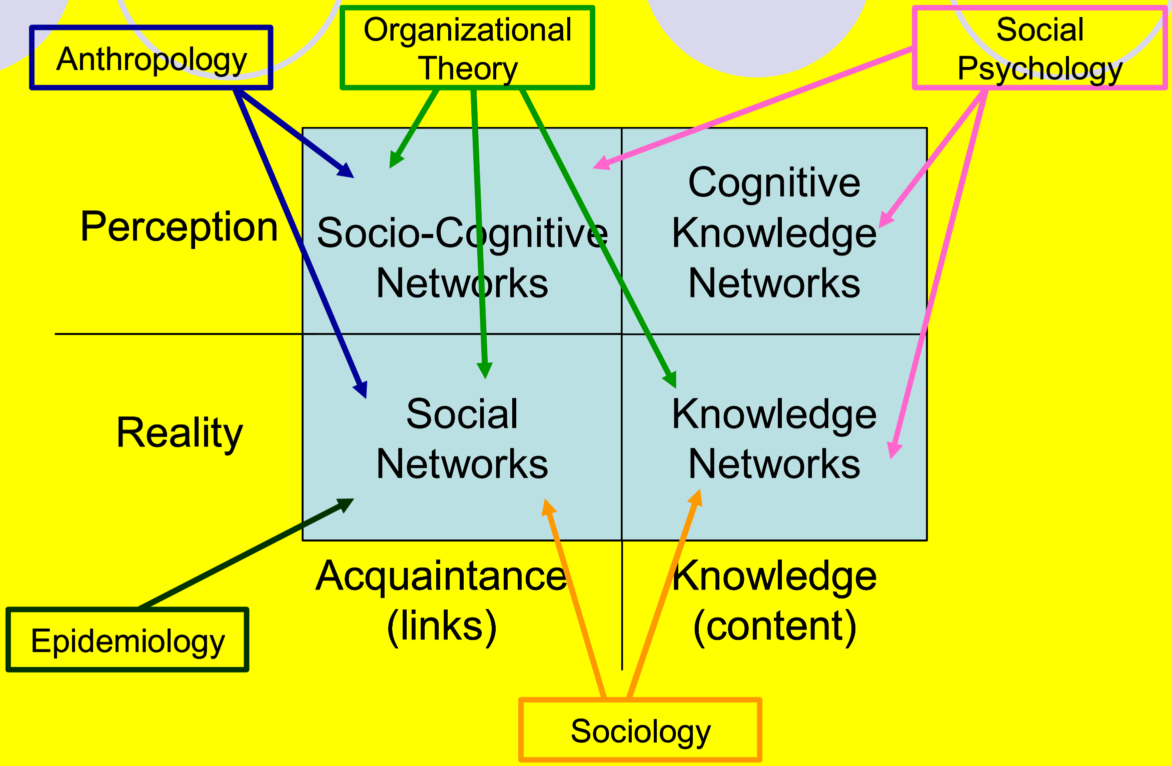
Networks in Social Sciences

- Types of Networks (Contractor, 2006)
  - Social Networks
    - “who knows who”
  - Cognitive Social Networks (also called Socio-Cognitive Networks)
    - “who thinks who knows who”
  - Knowledge Networks
    - “who knows what”
  - Cognitive Knowledge Networks
    - “who thinks who knows what”

Types of Social Network Analysis

- **Sociocentric (whole) network analysis**
  - Emerged in sociology
  - Involves quantification of interaction among a socially well-defined group of people
  - Focus on identifying global structural patterns
  - Most SNA research in organizations concentrates on sociometric approach
- **Egocentric (personal) network analysis**
  - Emerged in anthropology and psychology
  - Involves quantification of interactions between an individual (called *ego*) and all other persons (called *alters*) related (directly or indirectly) to ego
  - Make generalizations of features found in personal networks
  - Difficult to collect data, so till now studies have been rare

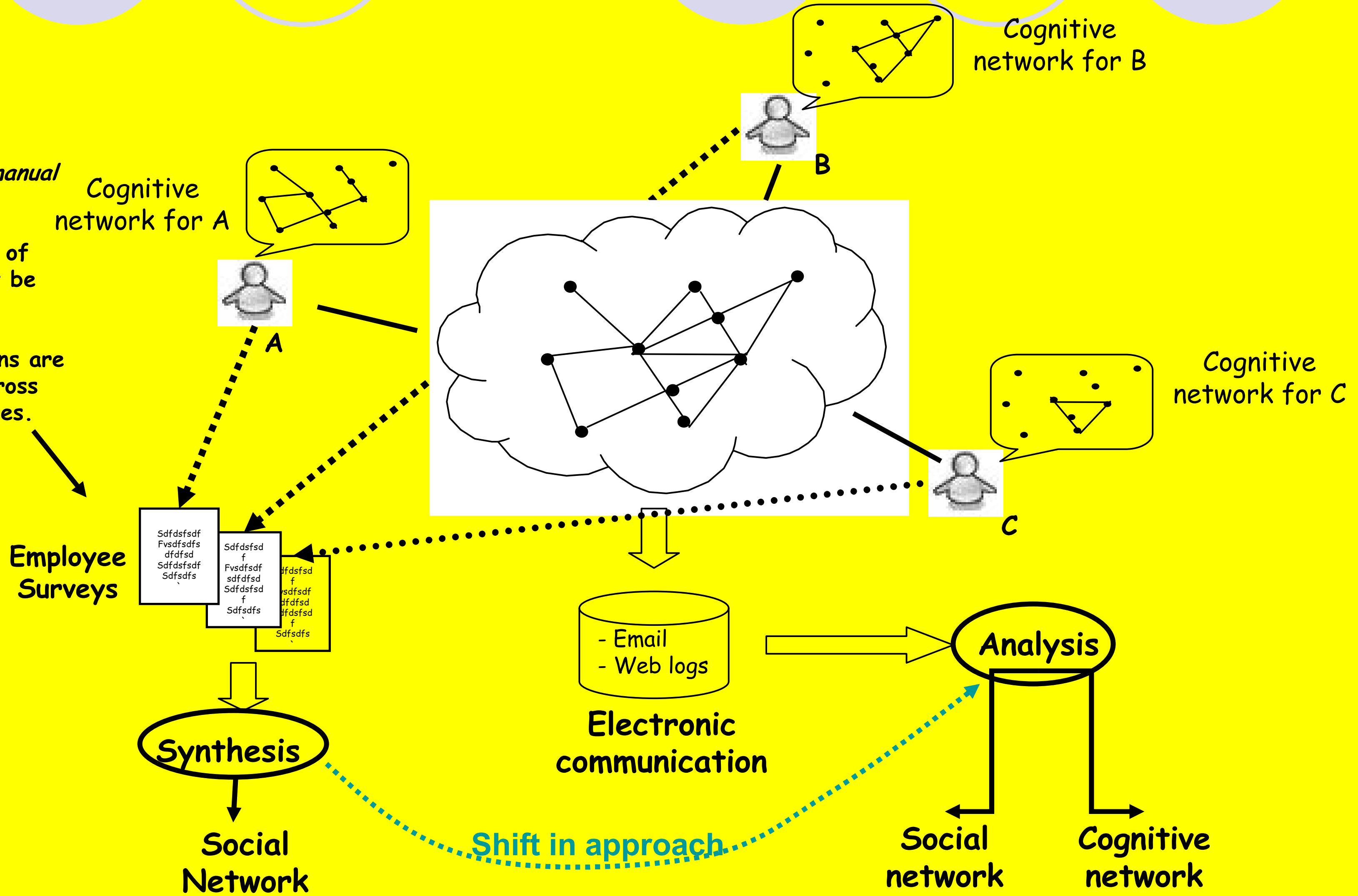
Networks Research in Social Sciences



- Social science networks have widespread application in various fields
- Most of the analyses techniques have come from Sociology, Statistics and Mathematics
- See (Wasserman and Faust, 1994) for a comprehensive introduction to social network analysis

A shift in approach: from ‘synthesis’ to ‘analysis’

Problems
- *High cost of manual surveys*
- *Survey bias*
  - Perceptions of individuals may be incorrect
- *Logistics*
  - Organizations are now spread across several countries.



Modeling a Cognitive Social Network

Example of E-mail Communication

- A sends an e-mail to B
  - With Cc to C
  - And Bcc to D
- C forwards this e-mail to E
- From analyzing the header, we can infer
  - A and D know that A, B, C and D know about this e-mail
  - B and C know that A, B and C know about this e-mail
  - C also knows that E knows about this e-mail
  - D also knows that B and C do not know that it knows about this e-mail; and that A knows this fact
  - E knows that A, B and C exchanged this e-mail; and that neither A nor B know that it knows about it
  - and so on and so forth …

Modeling Pair-wise Communication

- Modeling pair-wise communication between actors
  - Consider the pair of actors (A_x, A_y)
  - Communication *from* A_x *to* A_y is modeled using the Bernoulli distribution $L(x,y)=[p,1-p]$
  - Where,
    - p = (# of emails from A_x with A_y as recipient)/(total # of emails sent by A_x)
- For N actors there are $N(N-1)$ such pairs and therefore $N(N-1)$ Bernoulli distributions
- Every email is a Bernoulli trial where success for $L(x,y)$ is realized if A_x is the sender and A_y is a recipient

Modeling an agent's belief about global communication

- Based on its observations, each actor entertains certain beliefs about the communication strength between all actors in the network
- A belief about the communication expressed by *L(x,y)* is modeled as the Beta distribution, *J(x,y)*, over the parameter of *L(x,y)*
- Thus, belief is a probability distribution over all possible communication strengths for a given ordered pair of actors (A_x,A_y)

Model for Belief Update

- $J_k(x,y)$ is the Beta distribution maintained by actor A_k regarding its belief about the communication from A_x to A_y
- *a* and *b*, the two parameters of $J_k(x,y)$, are associated with the number of emails observed by A_k which are
  - from A_x to A_y , i.e. number of successes, and
  - from A_x not to A_y, i.e. number of failures
- Initialization
  - *a* and *b* start out with default initial values
  - Many different possibilities
    - For example, values can be chosen to be small so that they do not have much of an impact and can be *"washed out"* by future observations
- Belief update
  - on observing a success or failure, A_k increments *a* or *b* respectively

Belief State of an Actor

- Every actor maintains Beta distributions (or beliefs) for all ordered pairs of actors in the network
- Actor A_k's *belief state* is defined to be the set of all *N(N-1)* Beta distributions (one for every Bernoulli distribution)
- We also introduce a *"super-actor"* in the network
  - The super-actor is an actor who observes all the communication in the network
  - Super-actor is used as the baseline for reality
  - E-mail server is the "super-actor"

Quantitative Measures for Perceptual Closeness

Types of Perceptual Closeness

- We analyze the following aspects
  - Closeness between an actor's belief and reality, i.e. "true knowledge" of an actor
  - Closeness between the beliefs of two actors, i.e. the "agreement" between two actors
- We define two metrics, *r-closeness* and *a-closeness* for measuring the closeness to reality and closeness in the belief states of two actors respectively

Measuring the Closeness Between Beliefs

- For measuring the closeness between two belief states, the KL-divergence across the expected Bernoulli distributions for the two respective beliefs is computed.
  - The expected Bernoulli distribution for a belief is the expectation of the Beta distribution corresponding to that belief
  - If *J(a,b)k,t* is the Beta distribution, then the corresponding expected Bernoulli distribution (denoted by *E[J(a,b)k,t])* is obtained by normalizing the parameters of Beta distribution *J(a,b)k,t*

$$E[J(x,y)_{k,t}] = [\frac{\alpha(x,y)_{k,t}}{\alpha(x,y)_{k,t} + \beta(x,y)_{k,t}}, \frac{\beta(x,y)_{k,t}}{\alpha(x,y)_{k,t} + \beta(x,y)_{k,t}}]$$

Belief Divergence Measures

- The divergence of one belief, expressed by the Beta distribution $J(a,b)_{x,t}$, from another, expressed by $J(a,b)_{y,t}$ at a given time t, is defined as,

$$KL(E[J(a,b)_{x,t}] \,\|\, E[J(a,b)_{y,t}]) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \ldots (4)$$

where, $p = \frac{\alpha(x,y)_{x,t}}{\alpha(x,y)_{x,t} + \beta(x,y)_{x,t}}$ and $q = \frac{\alpha(x,y)_{y,t}}{\alpha(x,y)_{y,t} + \beta(x,y)_{y,t}}$

- The divergence of a belief state $B_{y,t}$ from the belief state $B_{x,t}$ for two actors A_y and A_x respectively, at a given time *t,* is defined as,

$$div(B_{x,t}, B_{y,t}) = \frac{\sum_{\forall(a,b)\in(B_{x,t}\cap B_{y,t})} KL(E[J(a,b)_{x,t}] \,\|\, E[J(a,b)_{y,t}])}{n(B_{x,t}\cap B_{y,t})} \ldots (5)$$

Belief Divergence Measures (contd.)

- The a-closeness measure is defined as the level of agreement between two given actors A_x and A_y with belief states $B_{x,t}$ and $B_{y,t}$ respectively, at a given time t and is given by,

$$a-closeness(B_{x,t}, B_{y,t}) = \frac{1}{1 + div(B_{x,t}, B_{y,t}) + div(B_{y,t}, B_{x,t})} \quad \ldots(6)$$

- The r-closeness measure is defined as the closeness of the given actor A_k's belief state $B_{k,t}$ to reality at a given time t and it is given by,

$$r-closeness(A_k) = \frac{1}{1 + div(B_{S,t}, B_{k,t})} \quad \ldots(7)$$

Where $B_{S,t}$ is the belief state of the super-actor A_S at time t

Interpretation of the Metrics

- The r-closeness measure
  - An actor who has accurate beliefs regarding only few communications is closer to reality than some other actor who has a relatively large number of less accurate beliefs
  - Thus, accuracy of knowledge is important

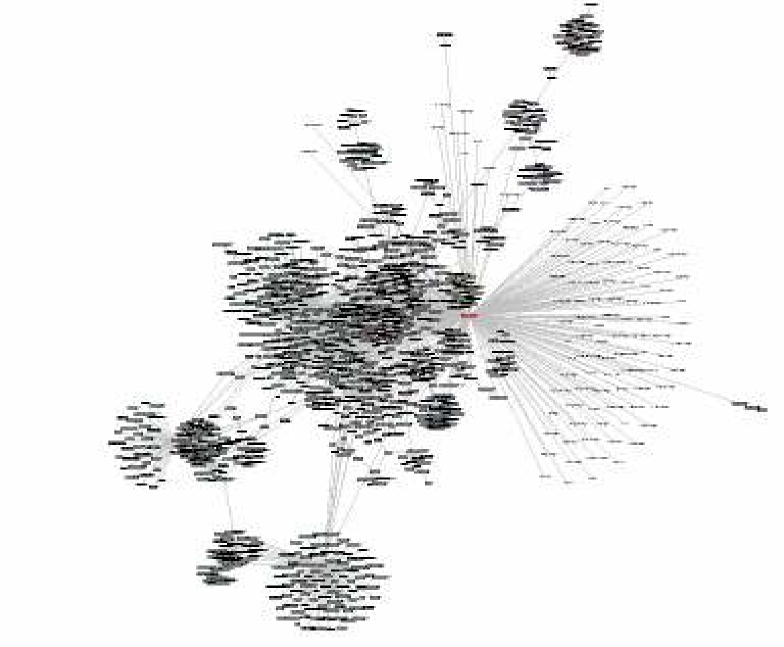
- The a-closeness measure between actor pairs
  - Consider three actors A_x, A_y and A_z
  - Suppose we want to determine how divergent are A_y*'s* and A_z*'s* belief states from that of A_x*'s*
  - If A_y and A_x have few beliefs in common, but low divergence for each of these few common beliefs, then their belief states may be closer than those of A_z and A_x, who have a relatively larger number of common beliefs with greater divergence across them

- a-closeness measure can be used to construct an "agreement graph" (or a who agrees with whom graph)
  - Actors are represented as nodes and an edge exists between two actors only if the agreement or the a-closeness between them exceeds some threshold *t*

r-closeness and a-closeness experiments with Enron E-mail logs

Enron Email Logs

- Publicly available: http://www.cs.cmu.edu/~enron/
- Cleaned version of data
  - 151 users, mostly senior management of Enron
  - Approximately 200,399 email messages
  - Almost all users use folders to organize their emails
  - The upper bound for number of folders for a user was approximately the log of the number of messages for that user
  - A visualization of Enron data (Heer, 2005)
- For experiments emails exchanged between users for the months of October 2000 and October 2001 were used



Testing 'conventional wisdom' using r-closeness

- *<u>Conventional wisdom 1:</u> As an actor moves higher up the organizational hierarchy, it has a better perception of the social network*
  - It was observed that majority of the top positions were occupied by employees
- *<u>Conventional wisdom 2:</u> The more communication an actor observes, the better will be its perception of reality*
  - Even though some actors observed a lot of communication, they were still ranked low in terms of r-closeness.
  - These actors focus on a certain subset of all communications and so their perceptions regarding the social network were skewed towards these "favored" communications
  - Executive management actors who were communicatively active exhibited this "skewed perception" behavior
    - which explains why they were not ranked higher in the r-closeness measure rankings as expected in 1

Experiment with r-closeness – Oct 2000

- For October 2000, based on their r-closeness rankings actors can be roughly divided into three categories
  - Top ranks: Actors who are communicatively active and observe a lot of diverse communications
  - Mid ranks: Actors who also observe a lot of communication but had skewed perceptions
  - Low ranks: Actors who are communicatively inactive and hardly observe any of the communication

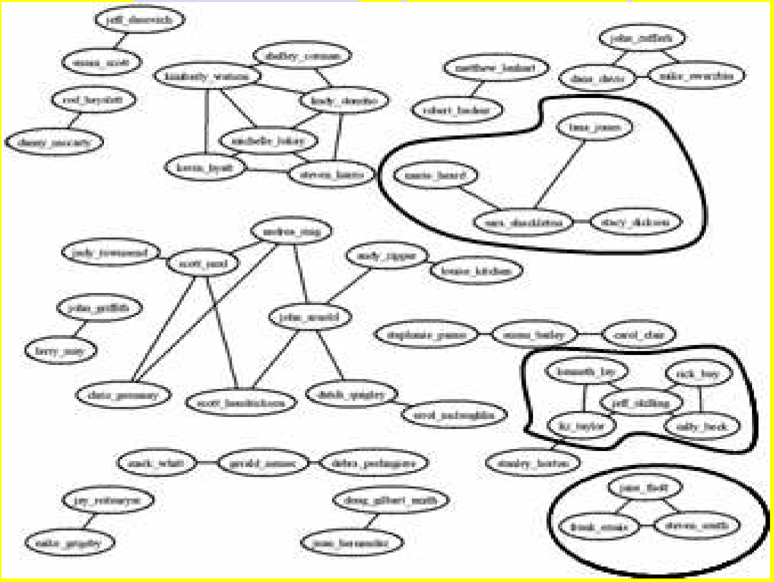
Ranks	Not Available	Employees	Higher Management	Executive Management	Others
1-10	2.6% (1)	14.6% (6)	0% (0)	6.9% (2)	6.67% (1)
11–50	28.9% (11)	34.1% (14)	21.4% (6)	24.1% (7)	13.33% (2)
51-151	68.5% (26)	51.3% (21)	78.6% (22)	69% (20)	80% (12)
Total	100% (38)	100% (41)	100% (28)	100% (29)	100% (15)

Experiment with r-closeness – Oct 2001

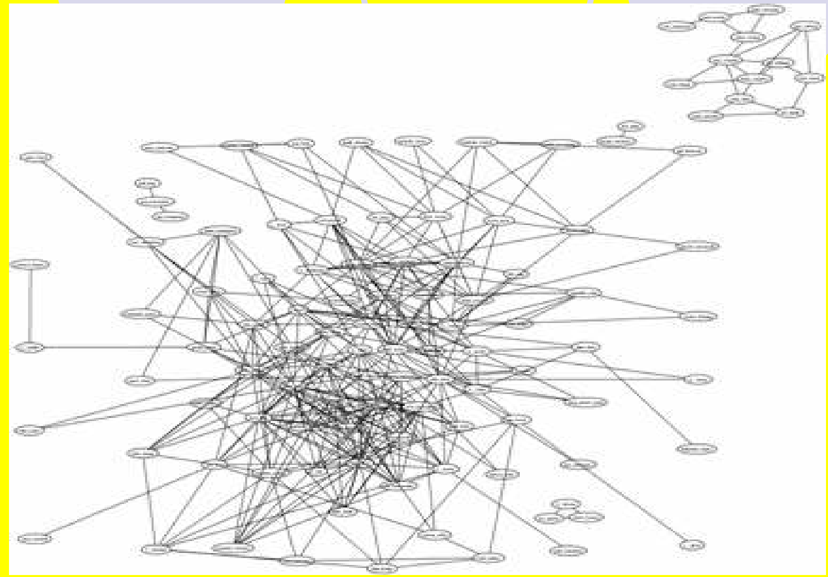
- r-closeness rankings for the crisis month Oct, 2001 show a significant increase (31% to 65.5%) in the percentage of senior executive management level actors in the top 50 ranks, with employees moving down

Ranks	Not Avail-Able	Emplo-yees	Higher Manage-ment	Executive Manage-ment	Others
1-10	7.9 % (3)	9.75% (4)	0% (0)	10.3% (3)	0% (0)
11–50	21.1 % (8)	17.1% (7)	25% (7)	55.2% (16)	13.33% (2)
51-151	71% (27)	73.15% (30)	75% (21)	34.5% (10)	86.67% (13)
Total	100% (38)	100% (41)	100% (28)	100% (29)	100% (15)

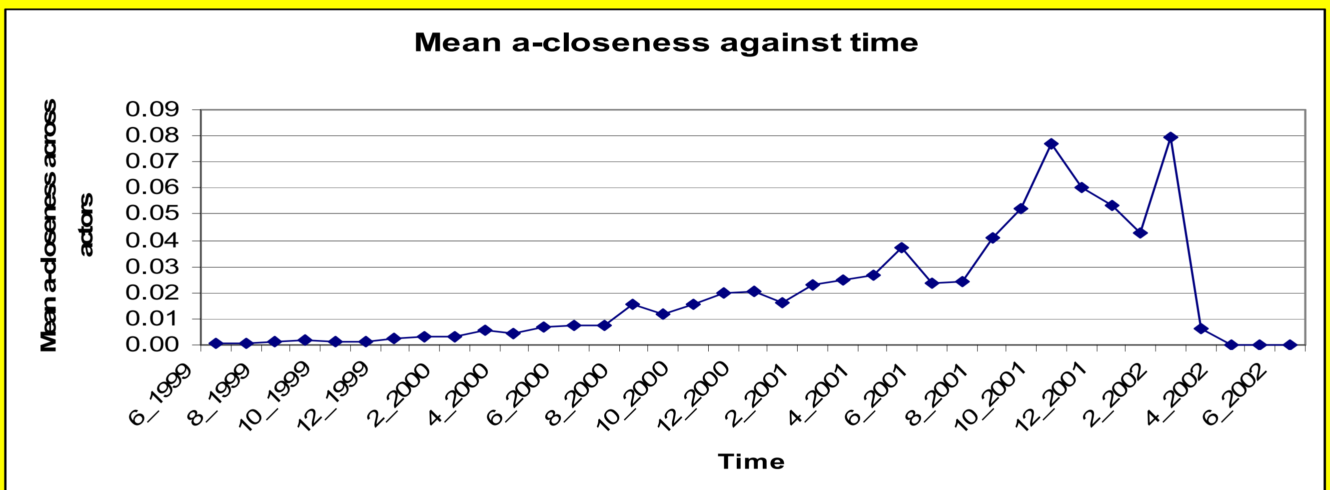
Experiment with a-closeness



Agreement Graph for Oct 2000 (threshold = 0.95)



Agreement Graph for Oct 2001 (threshold = 0.95)



Mean a-closeness against time

Automatic Extraction of Concealed Relations

Concealed Relations

- Concealed/Covert Relations: Relations between groups of actors that
  - have high strength
  - but are known to very few actors in the network outside the group
- Problem: Given email log data for all actors, extract the concealed relations from this data

An IR-Motivated Approach

- Use an approach motivated by informational retrieval
- Use a *tf-idf* style scheme for relations
  - an actor's view of the social network ➔ document in a corpus
  - an (unordered) pair-wise actor relation ➔ a term in a document
  - number of instances of a relation observed by an actor ➔ term frequency (*tf*) in a document
  - number of actors that know about a relation ➔ document frequency (*df*)
  - actual frequency of a relation is used to determine a 'global' ranking of concealed relations

Top 10 Concealed Relations

Table 1: Top 10 Concealed Relations (October 2000)

Relation	Score
Tana Jones (e) ↔ Sara Shackleton (e)	1.7760794$E7$
Richard shapiro (vp) ↔ Jeff Dasovich (e)	1.3316896$E7$
Marie Heard (na) ↔ Tana Jones (e)	1.2031506$E7$
Jeff Dasovich (e) ↔ Mary Hain (lawyer)	1.0895643$E7$
Stephanie Panus (e) ↔ Sara Shackleton (e)	1.0026255$E7$
Stacy Dickson (e) ↔ Tana Jones (e)	9685016.0
Matthew Lenhart (e) ↔ Eric Bass (trader)	8021003.5
Mark Whitt (na) ↔ Gerald Nemec (na)	7739389.0
Richard Shapiro (vp) ↔ Mary Main (lawyer)	5182706.0
Stephanie Panus (e) ↔ Tana Jones (e)	4637158.0

Table 2: Top 10 Concealed Relations (October 2001)

Relation	Score
D. Steffes (vp) ↔ Jeff Dasovich (vp)	1.0007403$E7$
Richard Shapiro (vp) ↔ Jeff Dasovich (e)	5063396.0
D. Steffes (vp) ↔ Richard Shapiro (vp)	4718486.5
Marie Heard (na) ↔ Sara Shackleton (e)	3927464.5
Kimberly watson (e) ↔ Mark Mcconnell (na)	3759267.0
Kimberly watson (e) ↔ Michelle Lokay (e)	3408572.5
Mike Grigsby (man) ↔ Barry Tycholiz (vp)	3079402.2
Mike Grigsby (man) ↔ Matt Smith (na)	2905096.5
Mike Grigsby (man) ↔ Jason Wolfe (na)	2902135.2
Mike Grigsby (man) ↔ Jay Reitmeyer (e)	2852143.8

Score of this relation has dropped

Top actors from top clusters

Table 3: Top 5 actors for the top 3 Concealed Relations (October 2000)

Tana Jones (e) ↔ Sara Shackleton (e)	
Actor	**Actor Relative Score**
Tana Jones (e)	1.7760794E7
Sara Shackleton (e)	1.7760794E7
Susan Bailey (na)	5729288.5
Stephanie Panus (e)	5442824.0
Carol Clair (lawyer)	4583431.0
Richard Shapiro (vp) ↔ Jeff Dasovich (e)	
Actor	**Actor Relative Score**
Richard Shapiro (vp)	1.3316896E7
Jeff Dasovich (e)	1.3316896E7
Mary Hain (lawyer)	2723910.8
Robert Badeer (dir)	302656.75
B. Sanders (vp)	0.0
Marie Heard (lawyer) ↔ Tana Jones (e)	
Actor	**Actor Relative Score**
Marie Heard (lawyer)	1.2031506E7
Tana Jones (e)	1.2031506E7
Stacy Dickson (e)	7448075.5
Stephanie Panus (e)	1432322.1
Susan Bailey (na)	1432322.1

Table 4: Top 5 actors for the top 3 Concealed Relations (October 2001)

D. Steffes (vp) ↔ Jeff Dasovich (e)	
Actor	**Actor Relative Score**
D. Steffes (vp)	1.0007493E7
Jeff Dasovich	1.0007493E7
Richard Shapiro (vp)	5138983.0
J. Kean (vp)	1700114.6
B. Sanders (vp)	1661475.6
Richard Shapiro (vp) ↔ Jeff Dasovich (e)	
Actor	**Actor Relative Score**
Richard Shapiro (vp)	5063396.0
Jeff Dasovich (e)	5063396.0
D. Steffes (vp)	4324214.0
J. Kean (vp)	1921873.0
B. Sanders (vp)	702222.8
D. Steffes (vp) ↔ Richard Shapiro (vp)	
Actor	**Actor Relative Score**
D. Steffes (vp)	4718486.5
Richard Shapiro (vp)	4718486.5
Jeff Dasovich (e)	780501.5
B. Sanders (vp)	496682.78
Louise Kitchen (p)	319296.06

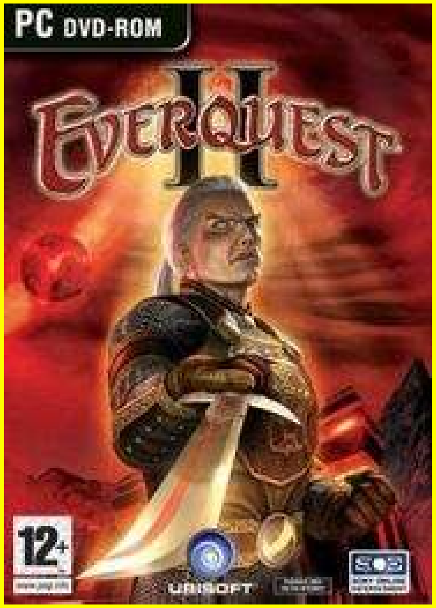
Some observations

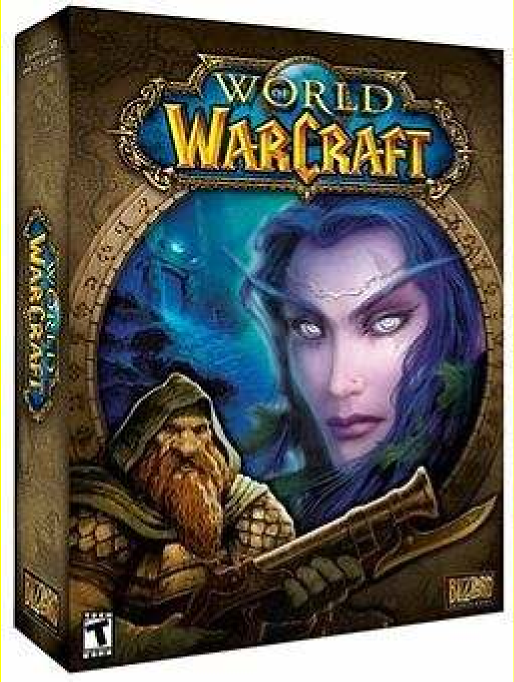
- Actors belonging to the smaller clusters tend to be aware of each others' communications and exhibit community behavior
- One of the strongest clusters of 3 actors consisting of the top 3 concealed relations for the October 2001 crisis period, is made up of actors who held the positions of Vice President (Government Affairs), Government Relations Executive and Vice President (Regulatory Affairs)
- Other statistics
  - October 2001 – 490 nonzero relations
  - October 2000 – 129 nonzero relations
  - Total 11325 relations

Mining MMORPG Logs for Social Science Research

MMO Games

- MMO (Massively Multiplayer Online) Games are computer games that allow hundreds to thousands of players to interact and play together in a persistent online world





Popular MMO Games- Everquest 2, World of Warcraft and Second Life

MMORPG – Everquest 2

- MMORPGs (MMO Role Playing Games) are the most popular of MMO Games
  - Examples: World of Warcraft by Blizzard and Everquest 2 by Sony Online Entertainment
- Various logs of players' behavior are maintained
- Player activity in the environment as well his/her chat is recorded at regular time instances, each such record carries a time stamp and a location ID
- Some of the logs capture different aspects of player behavior
  - Guild membership history (member of, kicked out of, joined, left)
  - Achievements (Quests completed, experience gained)
  - Items exchanged and sold/bought between players
  - Economy (Items/properties possessed/sold/bought, banking activity, looting, items found/crafted)
  - Faction membership (faction affiliation, record of actions affecting faction affiliation)

Impact on Social Science

- Interactions in MMO Gaming environments are real
- MMO Games provide sociologists with a unique source of data allowing them to observe real interactions in the context of a complete environment on a very fine granularity
- Gets around the serious issue of unbiased complete data collection
- Analysis of such data presents novel computational challenges
  - The scale of data is much larger than normally encountered in traditional social network analysis
  - The number of environment variables captured is greater
  - Player interaction data is captured at a much finer granularity
- MMORPG data requires models capable of handling large amounts of data as well as accounting for the many environment variables impacting the social structure

Social Science Research with Everquest 2 Data

- Objective of our research from a social science point of view is to improve understanding of the dynamics of group behavior
- Traditional analysis of dynamics of group behavior works with a *fixed* and *isolated* set of individuals
- MMORPG data enables us to look at dynamics of groups in a new way
  - Multiple groups are part of a large social network
  - Individuals from the social network can join or leave groups
  - Groups are not isolated and some of them can be related i.e. they may be geared towards specific objectives, each of which works towards a larger goal (e.g. different teams working towards disaster recovery)
  - The emergence, destruction as well as dynamic memberships of the groups depend on the underlying social network as well as the environment

DM Challenges for Social Science Research with Everquest 2 Data

- Inferring player relationships and group memberships from game logs
  - Basic elements of the underlying social network such as player-player and layer-group relationships need to be extracted from the game logs
- Developing measures for studying player and group characteristics
  - Novel measures need to be developed that measure individual and group relationships for dynamic groups
  - Novel metrics must also be developed for quantifying relationships between the groups themselves, the groups and the underlying social network as well as the groups and the environment
- Efficient computational models for analyzing group behavior
  - Extend existing group analysis techniques from the social science domain to handle large datasets
  - Develop novel group analysis techniques that account for the dynamic multiple group scenario as well as the data scale

Summary

- Research in Social Network Analysis has significant history
  - **Social sciences:** Sociology, Psychology, Anthropology, Epidemiology, …
  - **Physical and mathematical sciences:** Physics, Mathematics, Statistics, …
- **Late 1990s:** computer networks provided a mechanism to study social networks at a granular level
  - Computer scientists joined the fray
- **2000 onwards:** Explosion in infrastructure, tools, and applications to enable social networking, and capture data about the interactions
  - Opens up exciting areas of data mining research

Impact on Organizational Policy Research

- Data security
  - An absolute must
- Privacy
  - Careful balance between privacy and data analysis
- Impact of SNA on employee-organization relationship
  - Careful thought needed in managing this
  - Should there be ‘opt-in’ or ‘opt-out’ options for employees?
  - Is this too ‘big brother-ish’?
- Bottom line
  - New technologies are radically transforming the workplace, impacting organization information flow like never before
  - Not managed properly, they can lead to serious problems, e.g. employee releasing corporate secrets in blogs (Google)
  - Need to have tools that enable the understanding (and thus management) of organizational information flow

Thank you!

And be careful with that e-mail ☺