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

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)
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
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.

Employee Surveys → **Synthesis** → Social Network → **Shift in approach** → Analysis

- **Electronic communication**
  - Email
  - Web logs

Shifts in approach:
- Cognitive network for A
- Cognitive network for B
- Cognitive network for C
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 \textit{from} \(A_x\) \textit{to} \(A_y\) is modeled using the Bernoulli distribution \(L(x,y)=[p, 1-p]\)
  - Where,
    - \(p = (# \text{ of emails from } A_x \text{ with } A_y \text{ as recipient})/(\text{total # of emails sent by } A_x)\)
  - 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
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 $J(a,b)_{k,t}$.

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

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,

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

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

- 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,

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

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/](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

Conventional wisdom 1: 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

Conventional wisdom 2: 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
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.

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

<table>
<thead>
<tr>
<th>Ranks</th>
<th>Not Available</th>
<th>Employees</th>
<th>Higher Management</th>
<th>Executive Management</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-10</td>
<td>7.9% (3)</td>
<td>9.75% (4)</td>
<td>0% (0)</td>
<td>10.3% (3)</td>
<td>0% (0)</td>
</tr>
<tr>
<td>11–50</td>
<td>21.1% (8)</td>
<td>17.1% (7)</td>
<td>25% (7)</td>
<td>55.2% (16)</td>
<td>13.33% (2)</td>
</tr>
<tr>
<td>51-151</td>
<td>71% (27)</td>
<td>73.15% (30)</td>
<td>75% (21)</td>
<td>34.5% (10)</td>
<td>86.67% (13)</td>
</tr>
<tr>
<td>Total</td>
<td>100% (38)</td>
<td>100% (41)</td>
<td>100% (28)</td>
<td>100% (29)</td>
<td>100% (15)</td>
</tr>
</tbody>
</table>
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

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 \textit{tf-idf} style scheme for relations
  - an actor’s view of the social network \( \rightarrow \) document in a corpus
  - an (unordered) pair-wise actor relation \( \rightarrow \) a term in a document
  - number of instances of a relation observed by an actor \( \rightarrow \) term frequency (\( tf \)) in a document
  - number of actors that know about a relation \( \rightarrow \) document frequency (\( df \))
  - actual frequency of a relation is used to determine a ‘global’ ranking of concealed relations
Table 1: Top 10 Concealed Relations (October 2000)

<table>
<thead>
<tr>
<th>Relation</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tana Jones (e) ← Sara Shackleton (e)</td>
<td>1.7760794E7</td>
</tr>
<tr>
<td>Richard shapiro (vp) ← Jeff Dasovich (e)</td>
<td>1.3316896E7</td>
</tr>
<tr>
<td>Marie Heard (na) ← Tana Jones (e)</td>
<td>1.2081506E7</td>
</tr>
<tr>
<td>Jeff Dasovich (e) ← Mary Hain (lawyer)</td>
<td>1.0895643E7</td>
</tr>
<tr>
<td>Stephanie Panus (e) ← Sara Shackleton (e)</td>
<td>1.0026255E7</td>
</tr>
<tr>
<td>Stacy Dickson (e) ← Tana Jones (e)</td>
<td>9685016.0</td>
</tr>
<tr>
<td>Matthew Lenhart (e) ← Eric Bass (trader)</td>
<td>8021003.5</td>
</tr>
<tr>
<td>Mark Whitt (na) ← Gerald Nemec (na)</td>
<td>7739389.0</td>
</tr>
<tr>
<td>Richard shapiro (vp) ← Mary Main (lawyer)</td>
<td>5182706.0</td>
</tr>
<tr>
<td>Stephanie Panus (e) ← Tana Jones (e)</td>
<td>4637158.0</td>
</tr>
</tbody>
</table>

Table 2: Top 10 Concealed Relations (October 2001)

<table>
<thead>
<tr>
<th>Relation</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. Steffes (vp) ← Jeff Dasovich (vp)</td>
<td>1.0007403E7</td>
</tr>
<tr>
<td>Richard Shapiro (vp) ← Jeff Dasovich (e)</td>
<td>9063396.0</td>
</tr>
<tr>
<td>D. Steffes (vp) ← Richard Shapiro (vp)</td>
<td>4718486.5</td>
</tr>
<tr>
<td>Marie Heard (na) ← Sara Shackleton (e)</td>
<td>3927464.5</td>
</tr>
<tr>
<td>Kimberly watson (e) ← Mark Mcconnel (na)</td>
<td>3759267.0</td>
</tr>
<tr>
<td>Kimberly watson (e) ← Michelle Loken (e)</td>
<td>3408572.5</td>
</tr>
<tr>
<td>Mike Grigsby (man) ← Barry Rycholiz (vp)</td>
<td>3079402.2</td>
</tr>
<tr>
<td>Mike Grigsby (man) ← Matt Smith (na)</td>
<td>2905096.5</td>
</tr>
<tr>
<td>Mike Grigsby (man) ← Jason Wolfe (na)</td>
<td>2902135.2</td>
</tr>
<tr>
<td>Mike Grigsby (man) ← Jay Reitmeyer (e)</td>
<td>2552143.8</td>
</tr>
</tbody>
</table>

Score of this relation has dropped
### Top actors from top clusters

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

<table>
<thead>
<tr>
<th>Actor</th>
<th>Actor Relative Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tana Jones (e)</td>
<td>1.7760794E7</td>
</tr>
<tr>
<td>Sara Shackleton (e)</td>
<td>1.7760794E7</td>
</tr>
<tr>
<td>Susan Bailey (na)</td>
<td>5729288.5</td>
</tr>
<tr>
<td>Stephanie Panus (e)</td>
<td>5442824.0</td>
</tr>
<tr>
<td>Carol Clair (lawyer)</td>
<td>4583431.0</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Actor</th>
<th>Actor Relative Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. Steffes (vp)</td>
<td>1.0007493E7</td>
</tr>
<tr>
<td>Jeff Dasovich</td>
<td>1.0007493E7</td>
</tr>
<tr>
<td>Richard Shapiro (vp)</td>
<td>5138983.0</td>
</tr>
<tr>
<td>J. Kean (vp)</td>
<td>1700114.6</td>
</tr>
<tr>
<td>B. Sanders (vp)</td>
<td>1661475.6</td>
</tr>
</tbody>
</table>

---

10/17/2007 Jaideep Srivastava 31
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.
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.
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 😊