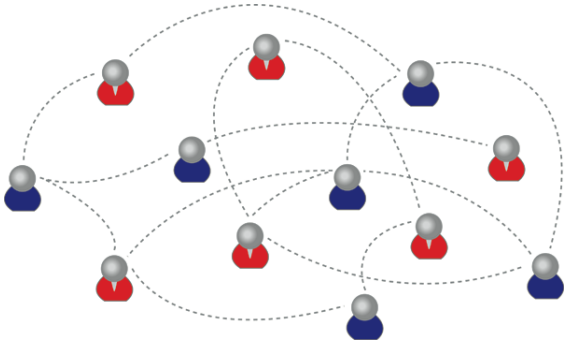
SOCIAL NETWORK MINING

– Aliasgar Lanewala

Contents

Introduction

- Social Network



- Social Network Analysis

Types of Social Network Analysis

- Sociocentric Network Analysis
  - Used in sociology
  - Focus is on measuring the structure of the organization
  - These patterns explain outcomes
  - Involves quantification of interaction among a socially well defined group of people
  - Results are generalized
  - Most SNA research in organizations employ the sociocentric approach

Types of Social Network Analysis

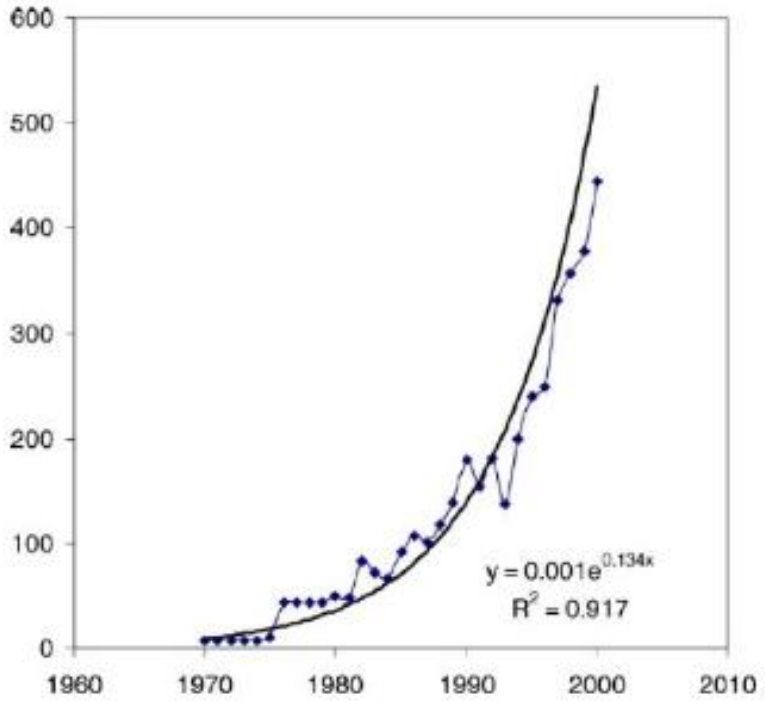
- Egocentric Social Analysis
  - Used in anthropology and psychology
  - *Ego* and *Alters*
  - Involves quantification of interactions between ego and alters related to the ego
  - Make generalizations of features found in personal network
  - Difficult to collect data

Types of Social Network Analysis

- Knowledge Based Network Analysis
  - Used in computer science
  - Involves quantification of interaction between individuals, groups and other entities.
  - Based on entities associated with actors in the social newtwork.

Classical Social Network Analysis

- Social networks have been widely studied since a long time, historically.
- Since 1990s, there is a massive increase in studies in this area.



Exponential growth of publications indexed by Sociological Abstracts containing "social network" in the abstract or title Source: [7]

Terms and Key Concepts

- **Actor:** Nodes in a social network
- Dyad: A pair of actors in the network
- Triad: A subset of three actors or nodes
- Degree Centrality: Degree of node normalized to the interval {0...1}
- Clustering Coefficient: When applied to a single node, it is the measure of how complete the neighborhood of the node is.

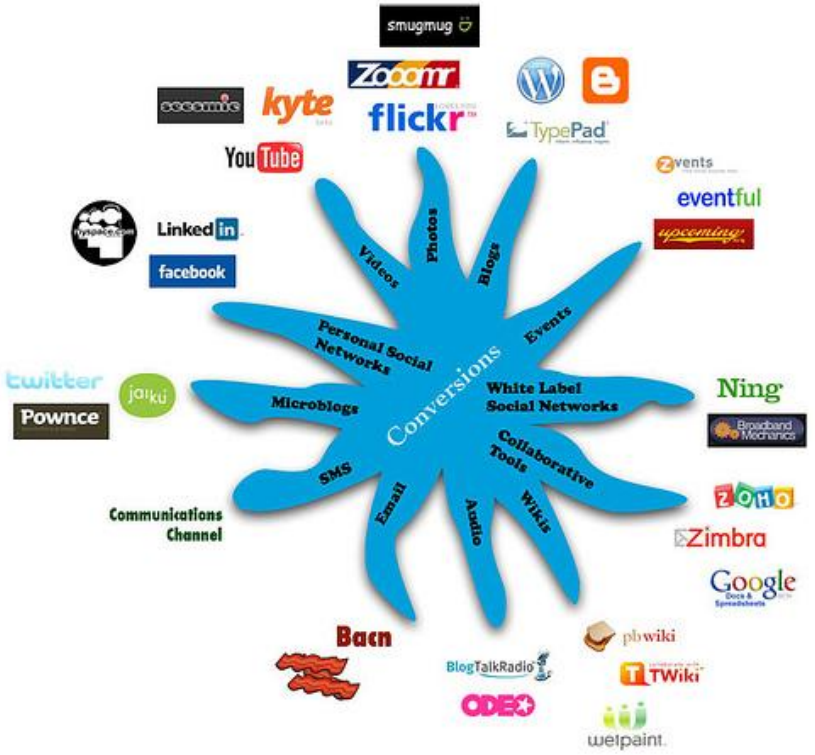
$$C_i = \frac{2|\{e_{jk}\}|}{k_i(k_i-1)} : v_j, v_k \in N_i, e_{ij} \in E$$

Measures of Network Centrality

- Betweenness Centrality
  - Most popular measure of centrality
  - Efficient computation is necessary; best technique is O(mn)
- Closeness Centrality
- Degree Centrality
- Eigenvector Centrality
  - Google's PageRank is an example of this
- Eccentricity

Social Networks in Online Age

- "Computer networks are inherently social networks, linking people, organizations and knowledge."[9]
- Data sources include newsgroups, instant messenger logs, emails, social networks, weblogs, microblogs, etc.



Source: [10]

Key Drivers in SNA

- Infrastructure for
  - Social interaction
  - Knowledge sharing
  - Knowledge discovery
- Ability to capture
  - Difference about various types of social interaction
  - Data at a very fine granularity
  - Without any reporting bias
- Data Mining techniques used

Data Mining for SNA

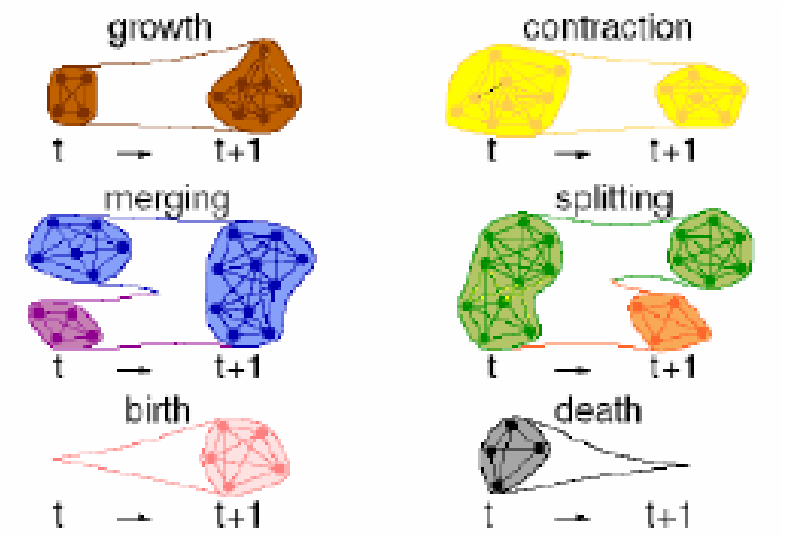
- Community Extraction
- Link Prediction
- Cascading Behavior
- Identifying Prominent Actors
- Search in Social Networks
- Trust in Social Networks
- Characterization of Social Networks
- Anonymity in Social Networks

Community Extraction

- Tyler, J. R., Wilkinson, D. M. and Huberman 2003.
  - The graph is broken into connected components and each component is checked to see if it is a community.
  - If a component is not a community then iteratively remove edges with highest betweenness till component splits. Recompute the betweenness each time an edge is removed.
  - The order in which edges are removed affects the final community structure.
  - Since ties are broken arbitrarily, this affects the final community structure.
  - The entire procedure is repeated several times and the results from each iteration are aggregated to produce the final set of communities.

Community Extraction

- Clique Percolation Method [11]
  - Locate Communities
    - Union of adjacent k cliques
    - Two k-cliques are adjacent if they share (k-1) nodes
    - k is a parameter
  - Identify Evolving Communities



Community Detection

- Community detection in large networks based on label propagation [12]
  - One's label is determined based on the majority of labels of its neighbors
  - Algorithm gives near-linear time complexity

1. Initialize the labels at all nodes in the network. For a given node x, $C_x(0) = x$.
2. Set $t = 1$.
3. Arrange the nodes in the network in a random order and set it to X.
4. For each $x \in X$ chosen in that specific order, let $C_x(t) = f(C_{x_{i1}}(t), ..., C_{x_{im}}(t), C_{x_{i(m+1)}}(t-1), ..., C_{x_{ik}}(t-1))$. f here returns the label occurring with the highest frequency among neighbors and ties are broken uniformly randomly.
5. If every node has a label that the maximum number of their neighbors have, then stop the algorithm. Else, set $t = t + 1$ and go to (3).

Link Prediction

- Different versions
  - Given a social network at time t_i predict the social link between actors at time t_{i+1}
  - Given a social network with an incomplete set of social links between a complete set of actors, predict the unobserved social links
  - Given information about actors, predict the social link between them (this is quite similar to social network extraction)

Link Prediction

- Link Prediction using supervised learning [13]
  - Use machine learning algorithms (decision tree, k-NN, SVM)
  - Identify a group of features that are most helpful in prediction
  - Best Predictor Features: Keyword Match count, Sum of neighbors, Shortest Distance

Link Prediction

- Prediction of Link Attachments [14]
  - Given a network at time t, the goal is to predict k potential links that are most likely to be converted to real links after a certain period of time.
  - Top k links are predicted to be the real links.
  - Pick two nodes v and w such that edge (v,w) does not exist and d(v,w) = 2
  - An edge is created between v and the adjacent nodes of w if information propagation between the two is successful.
  - In the dataset only a small fraction (0.0002) of the potential links are converted to real links. The proposed method outperformed all the other comparison methods.

Identifying Prominent Actors

- Compute scores/rankings over the set (or a subset) of actors in the social network which indicate degree of importance

- Centrality measures
  - Degree Centrality
  - Closeness Centrality
  - Betweenness Centrality

Identifying Prominent Actors

- Based on Betweenness Centrality [15]
  - High betweenness value means Prominence
  - An efficient algorithm for computing for betweenness centrality
  - Betweenness centrality requires computation of number of shortest paths passing through each node
  - Compute shortest paths between all pairs of vertices
  - Trivial solution of counting all shortest paths for all nodes takes O(n3) time
  - A recursive formula is derived for the total number of shortest paths originating from source *s* and passing through a node *v*

$$\delta_{s\bullet}(v) = \sum_{w\,:\,v \in P_s(w)} \frac{\sigma_{sv}}{\sigma_{sw}} \cdot (1 + \delta_{s\bullet}(w))\,.$$

  - The time complexity reduces to O(mn) for unweighted graphs and O(mn + log2n) for weighted graphs
  - The space complexity decreases from O(n2) to O(n+m)

Search in Social Networks

- Searching or querying for information
- Technique of query routing
  - A user can send out queries to neighbors
  - If the neighbor knows the answer then he/she replies else forward it to their neighbors. Thus a query propagates through a network
- Greedy traversal algorithm
  - At each step the query is passed to the neighbor with the most number of neighbors
  - A large portion of the graph is examined in a small number of hops

Trust in Social Networks

- Trust Propagation
  - A user trusts some of his friends, his friends trust their friends and so on.
- TrustMail [16]
  - Consider research groups X and Y headed by two professors such that each professor knows the students in their respective group
  - If a student from group X sends a mail to the professor of group Y then how will the student be rated?
  - Use the rating of professor from group X who is in professor Y's list of trusted list and propagate the rating

Anonymity in Social Networks

- Anonymized Social Networks [17]
  - In order to preserve the privacy of the participants of a social network, names are replaced with meaningless unique identifiers
  - Is this sufficient? No!
  - Various attacks can reveal the true identities of the users

- Types of Attacks:
  - Active
  - Passive

Types of Attacks

- Active Attack
  - Add a node to the graph G
  - Add an undirected edge to graph G
  - Discover edges and targeted nodes

- Passive Attack
  - Attackers are part of the network
  - They discover themselves
  - Thus, compromise the privacy of their neighbors

SNA from Online Networks

- Study of Facebook messages [18]
  - Poking and messaging patterns are extremely similar.
  - Activity on the online social network varies depending upon the time of the day.

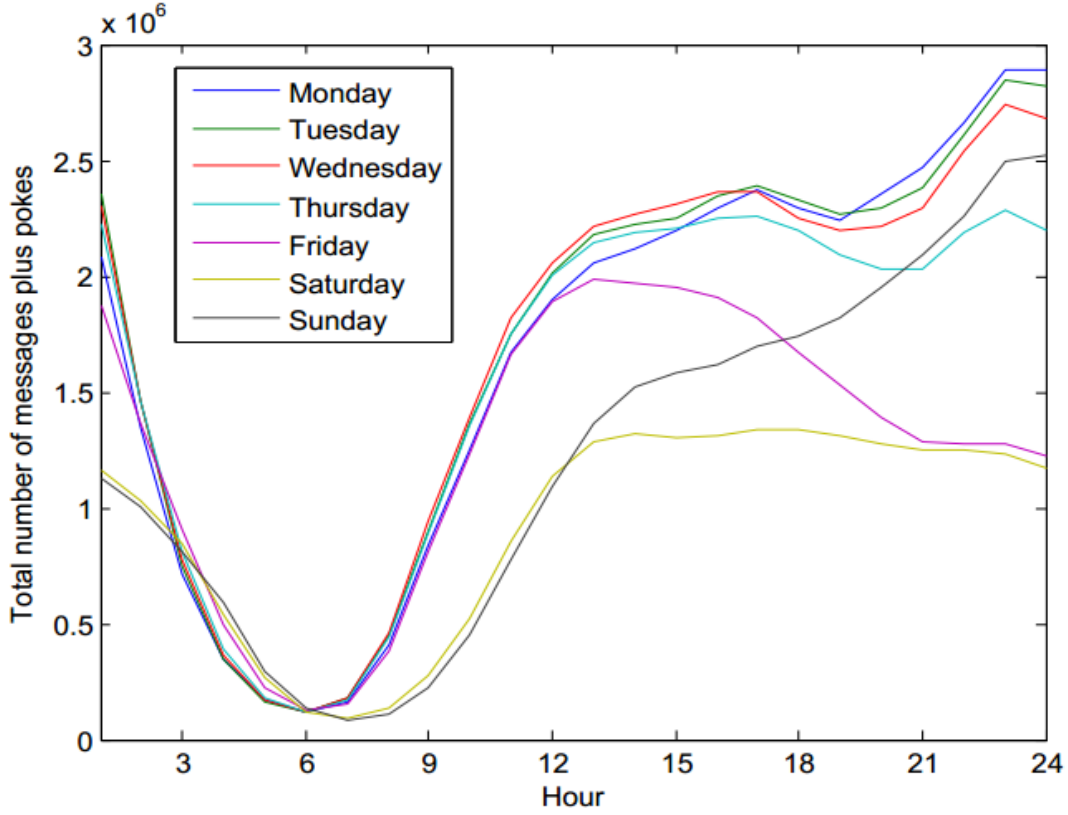


Fig. 3. Message plus pokes sent by hour in Facebook (color)

SNA from Online Networks

- Study of Facebook messages (continued)
  - Different patterns are observed in a corporate messaging network as compared to Facebook suggesting different nature of interaction.
  - Interaction on Facebook does not represent leisure time but rather interaction in parallel with other work.

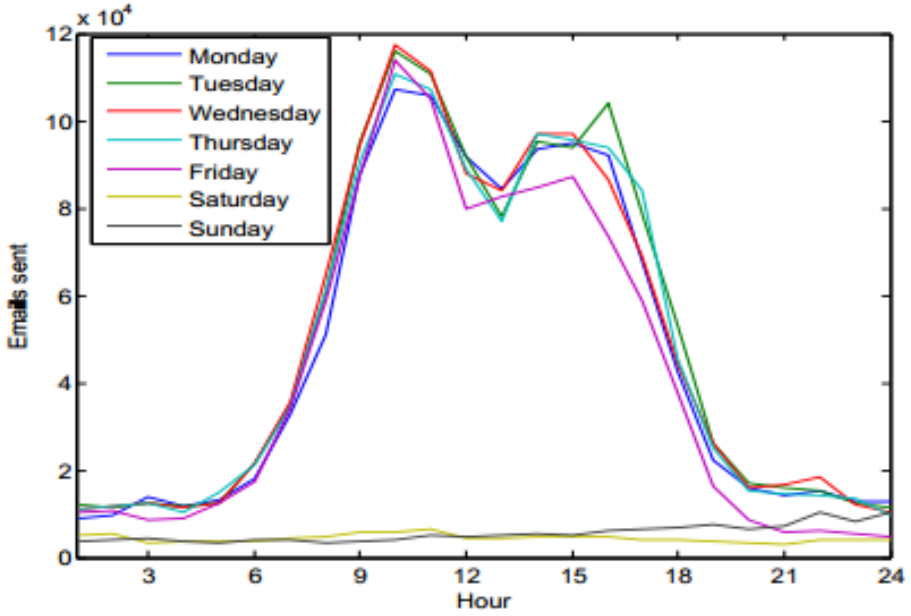


Fig. 4. Message plus pokes sent by hour in a corporate network (color)

Applications of DM based SNA Techniques

- Viral Marketing
- Social Influence and E-Commerce
- Social Computing
- Social Recommendation Systems
- Criminal Network Analysis
- And many more……

Viral Marketing

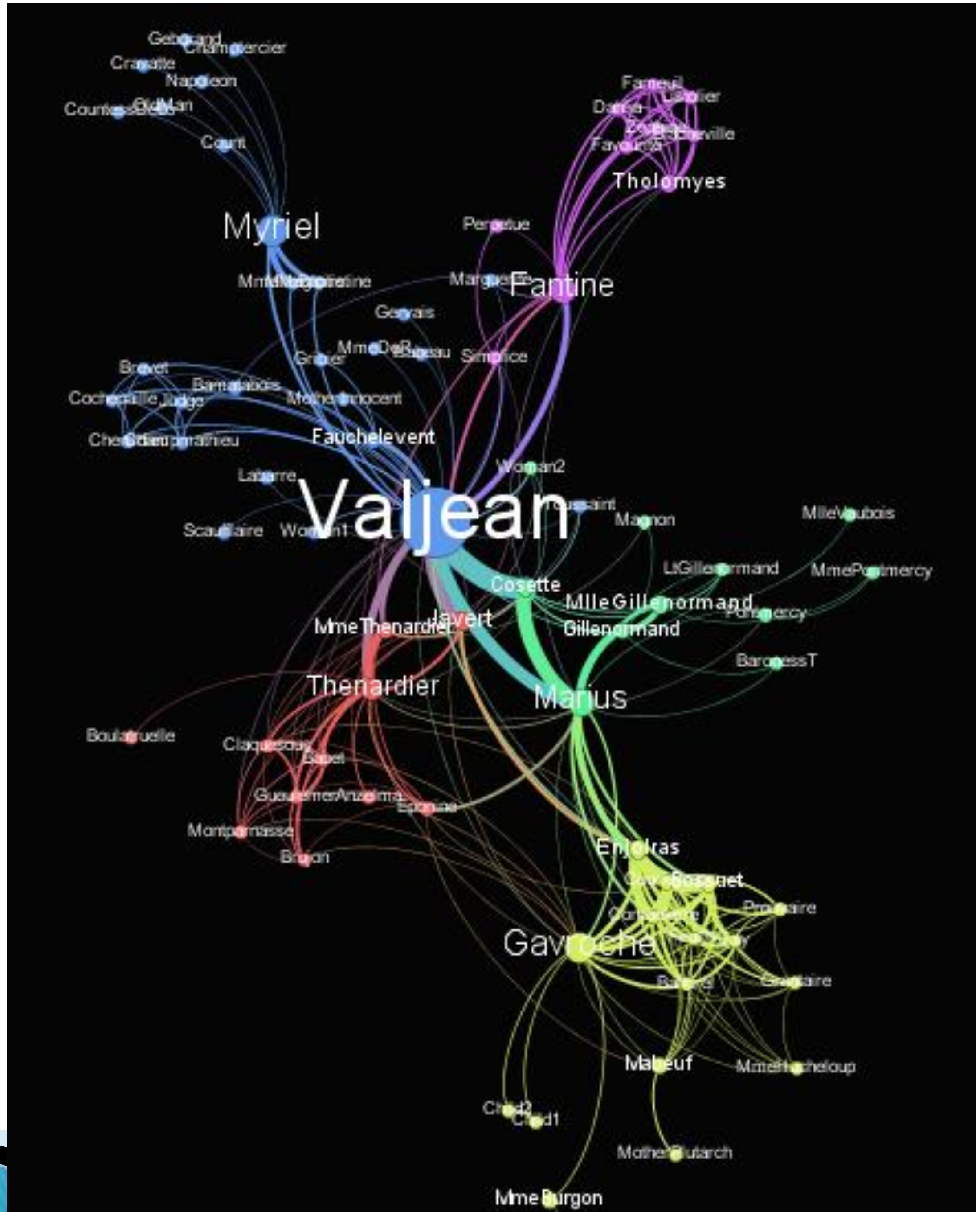
- Customer value is defined as the profit from sales to the customer.
- Customer value determines how much is it worth spending to retain the customer.
- Traditional measures fail to consider the fact that a customer may influence others to buy a product.

Social Influence and E-Commerce

- Purchasing decisions are strongly influenced by people who the consumer knows and trusts.
- Many shoppers tend to wait for review from early adopter before making a purchase.
- Capturing the data about social influence can aid e-commerce to use the power of social interaction
  - Tell a friend at Amazon
  - Customer review discussion board
  - Write or rate reviews

Criminal Network Analysis

- Because SNA techniques are designed to discover patterns of interaction between social actors in a social network, they are appropriate for criminal network analysis.
- Intelligence and law enforcement agencies are interested in finding structural properties
  - What subgroups exist in the network?
  - How do these subgroups interact with each other?
  - What is the overall structure of the network?
  - What are the roles (central/peripheral) network members play?


Myriel
Fantine
Tholomyes
Napoleon
Count
Perpetue
Gervais
Simplice
Fauchelevent
Labarre
Scaufflaire
Valjean
Woman2
Magnon
MlleVaubois
LtGillenormand
MmePontmercy
Cosette
MlleGillenormand
Gillenormand
BaronessT
MmeThenardier
Javert
Thenardier
Marius
Boulatruelle
Montparnasse
Brujon
Enjolras
Bossuet
Prouvaire
Gavroche
Mabeuf
Child1
MotherPlutarch
MmeBurgon

Conclusion

- Computer Science has provided the infrastructure for
  - Fostering social interaction
  - Capture it at very fine granularity
  - Avoid report bias
- Computational social science has the potential to revolutionize social sciences like
  - Gene Sequencing revolutionized study of Genetics
  - The electron microscope revolutionized chemistry

References

- [1] https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201111fa4.html
- [2] Cover page picture, http://www.readwriteweb.com/archives/the_inner_circles_of_10_geek_heroes_on_twitter.php
- [3] SN Picture, http://www.relenet.com
- [4] SNA Picture, http://thetalentcode.com/
- [5] Chung, Hossain, Davis. "Exploring Sociocentric and Egocentric Approaches for SNA". http://kmap2005.vuw.ac.nz/papers/Exploring%20Sociocentric%20and%20Egocentric.pdf
- [6] IBM, Knowledge Discovery and Data Mining. http://researcher.ibm.com/view_pic.php?id=144
- [7] Borgatti and Foster, "The network Paradigm in Organizational Research". http://www.scribd.com/doc/2530469/Borgatti-S-Foster-P-2003-The-network-paradigm-in-organizational-research-A-review-and-typology
- [8] Gephi:wiki, http://www.wiki.gephi.org
- [9] Barry Wellman, "Computer Networks as Social Networks".
- [10] http://cleantechnica.com/2010/02/18/end-mountaintop-removal-coal-mining-today/
- [11] Palla et al. "Nature 446, 664 (2007)".

References

- [12] Albert et al. “Near linear time algorithm to detect community structures in large scale networks”. http://arxiv.org/pdf/0709.2938v1.pdf
- [13] Hasan et al. “Link prediction using supervised learning”.
- [14] Saito et al. “Prediction of Link Attachments by Estimating Probabilities of Information Propagation”.
- [15] Brandes, “A Faster Algorithm for Betweenness Centrality”.
- [16] Golbeck et al. “Trust Networks on the Semantic Web”.
- [17] Backstrom et al. “Wherefore art thou R3579X? Anonymized social networks, hidden patterns, and structural steganography”. http://www.cs.cornell.edu/~lars/www07-anon.pdf
- [18] Golder et al. “Rhythms of social interaction: Messaging within a massive online network”. http://www.hpl.hp.com/research/idl/papers/facebook/facebook.pdf

Thank You