Mining and Analyzing Social Media: Part 2

Dave King

January 7, 2013

Agenda: Part 2

- Sentiment Analysis & Opinion Mining
  - Defined
  - Business Interest & Software Packages
  - Levels of Analysis
  - Automated Classification
- Social Network Analysis
  - Defined
  - History
  - Basic techniques and measures
  - Ego and Social-Centric Analysis

Sentiment Analysis and Opinion Mining: Interchangeable Terms

Computational study of opinions, sentiments, subjectivity, evaluations, attitudes, appraisals, affects, views, emotions, etc., expressed in text. (Lui, 2012)

Sentiment Analysis and Opinion Mining: Business Interests

Service

Response

Products

Issues and Focus

Marketing

Message

Company

Opinion Mining and Sentiment Analysis: Some Sample Questions of Interest

- Is the sentiment towards my X primarily positive, neutral, or negative? How does it compare to my key competitors? Has it changed overtime?
- What factors are positively and negatively influencing my X's image?
- Are there opportunities and needs my customers are identifying for me through their conversations?

Opinion Mining and Sentiment Analysis: An Offshoot of Social CRM

Social CRM

- Social Media services, techniques and technology for engaging customers
- Sometimes synonymous with Social Media Monitoring
- Gartner's Magic Quadrant has a min of $10M in rev.

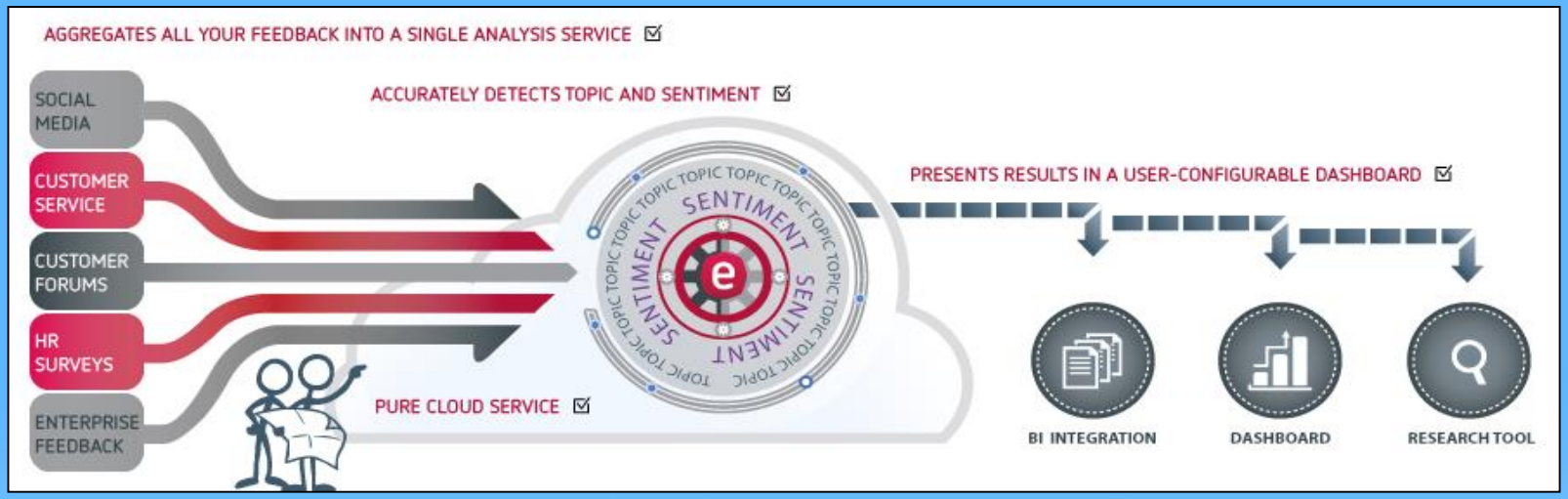
Opinion Mining and Sentiment Analysis: An Offshoot of Social CRM

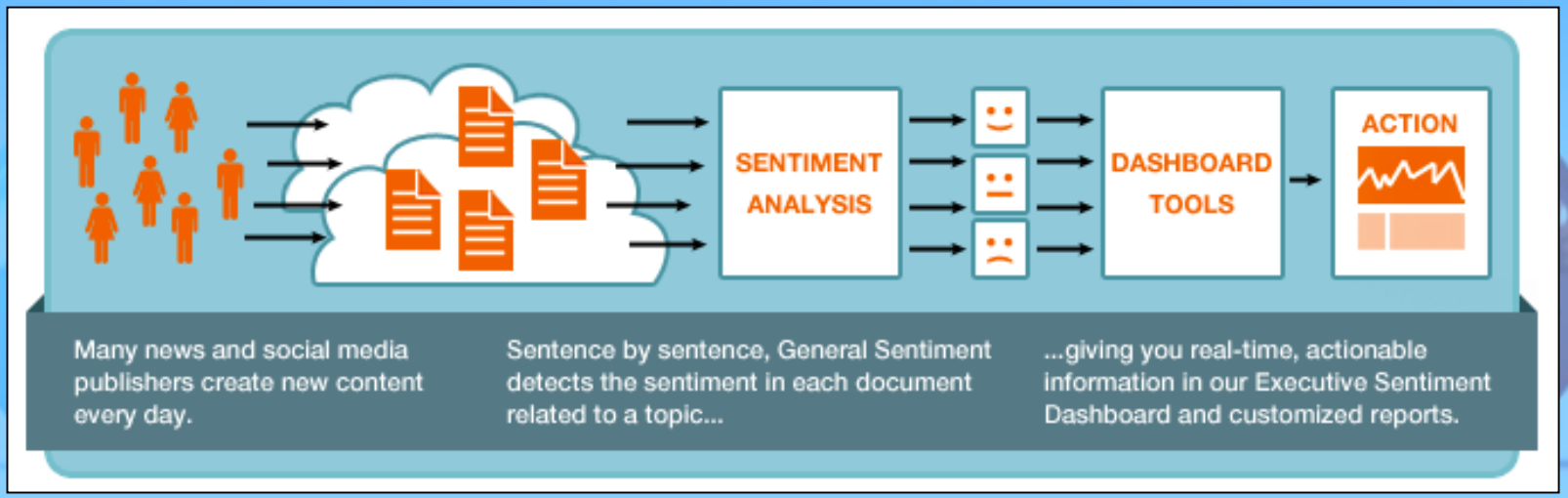
Company	Focus	Parent	Gartner
Alterian	Customer Experience Management (CXM)	SDL	
Artesian	Social CRM and Social Intelligence Platform		Niche
Attensity	Social CRM and Social Analytics and Engagement		Visionary
Bizaarvoice	Social CRM and Voice of the Customer (VoC)		Challenger
Brandwatch	Social Media Monitoring		
Clarabridge	Sentiment and Text Analytics		
Cognos	Customer Insight and Social Media Analytics	IBM	
Collective Intellect	Social CRM and Cloud-Based Social Intelligence Solution	Oracle	Challenger
Crimson Hexagon	Social Media Monitoring Analysis and Analytics		
Demand Media	Social CRM and Customer Insight Analysis		Niche
Digimind	Competitive Intelligence		
DigitalPebble	Open Source Text Engineering		
Etuna	Customer Feedback Analytics		
Evolve24	Market Research	Maritz Research	
General Sentiment	Sentiment Analytics		
Get Satisfaction	Social CRM and Customer Engagement Platform		Niche
Infinigraph	Intelligence Marketing for Social Media Communities		
Involver	Social Markup Language (SML) technology	Oracle	Challenger
Jive	Social CRM		Leader
Kana	Social CRM and Experience Analysis		Niche
Kotagent	User Analytics for Mobile and Social Web		
Lexalytics	Text Analytics, Text Mining, and Sentiment Analysis		

Opinion Mining and Sentiment Analysis: An Offshoot of Social CRM

Company	Focus	Parent	Gartner
Lithium	Social CRM and Social Mobility, Commerce, Support and Innovation		Leader
Lymbix	Sentiment and Tone		
Medallia	Customer Experience Management (CXM)		
Meltwater	Reputation Management for Social Media Monitoring and Analysis		
Meshlabs	Enterprise Text Analytics		
Netbase Solutions	Social Intelligence Platform		
Open Amplify	Natural Lanaguage Processing and Text Analytics		
Overtone	Customer Listening Platform and Sentiment Analysis	Kana	
Quantivo	Marketing Analysis		
Questback	Social CRM and Customer Experience Management		Niche
Radian6	Social CRM and Salesforce Marketing Cloud	Salesforce.com	Leader
SAS	Text and Sentiment Analytics		
Sentiment Metrics	Social Media Monitoring		
SentiMetrix	Opinion Tracking		
SPSS	Text Analytics and Sentiment Analysis of Survey Data	IBM	
Telligent	Social CRM and Social Community Software		Visionary
Traacker	Influencer Tracking and Marketing Program		
Virtue	Social CRM and Marketing Intelligence	Oracle	
Visible Technologies	Social CRM and Social Media Monitoring, Analytics and Engagement		Niche
Wise Window	Mass Opinion Business Intelligence	KPMG	

Sentiment Analysis and Opinion Mining: Commercial Products – General Operation





Sentiment Analysis and Opinion Mining: Some Examples – What do you see?

★★★★☆ **Very good value**, May 28, 2011

By **Cactus Man** (Florida) - See all my reviews

Amazon Verified Purchase (What's this?)

This review is from: **Canon Powershot A1200 12.1 MP Digital Camera with 4x Optical Zoom (Silver) (Camera)**

I needed a cheap pocket camera that takes good pictures and uses AA batteries. I didn't want to mess around with a charger and an extra battery.

The camera is small and does fit right into my pocket. It takes nice pictures, surpisingly good ones indoors without flash.

I turned off the display to conserve battery power and use the built-in view finder. More than 250 pics so far on the original two AA batteries!

Picture quality is good to very good, depending on my ability to hold it steady. It beats other cameras this size from other companies. Pictures are reasonably sharp, detailed. The pics are good for posting on the web, sending to friends some work situations, such as real estate, auto sales and such. Photos are not what you would send to a glossy magazine, but this is not a DSL with an expensive lens, either. Pics could be used for newspapers.

★☆☆☆☆ **Horrible pictures**, April 9, 2012

By **S. Salinas** (Dallas, TX) - See all my reviews

REAL NAME

Amazon Verified Purchase (What's this?)

This review is from: **Kodak Easyshare C1505 12 MP Digital Camera with 5x Digital Zoom - Red (Electronics)**

I gave this camera to my mom for Christmas. I've used it as much as she has and I've been the one to transfer the pictures off the camera to the computer. She hates this camera and I've got to agree with her. The "easy share" is not so easy but she isn't that concerned with that. It's the quality of the pictures that is so horrible. They are so foggy. It looks like there is some type of film over the lens but I've cleaned it several times and it is still there. See the images I added to see what I'm talking about. Don't waste your money on this camera.

Sentiment Analysis and Opinion Mining: What do you see?

- Opinion holders: persons who hold the opinions
- Opinion targets: entities and their features/aspects
- Sentiments: positive and negative
- Time: when opinions are expressed

Sentiment Analysis and Opinion Mining: Opinion Defined

Simply a positive or negative sentiment, view, attitude, emotion, or appraisal about an entity or an aspect of the entity from an opinion holder at a particular point in time.

- Opinion is a quintuple (e_j, a_{jk}, so_{ijkl}, h_i, t_l)
- Sentiment orientation ("so"): +, -, or possibly neutral

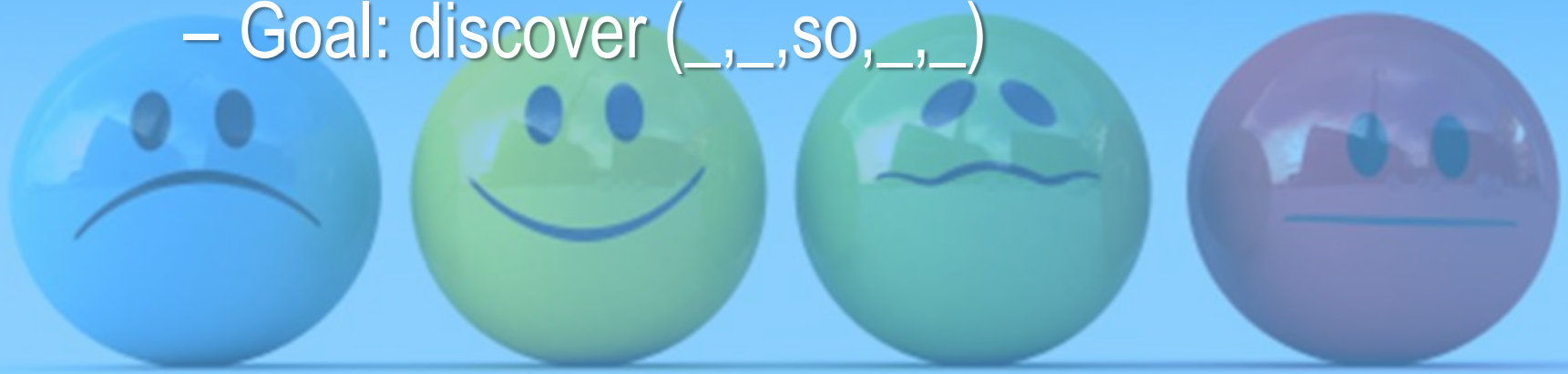
Sentiment Analysis and Opinion Mining: Level of Analysis

- Document Level: +, -, or 0*
- Sentence Level: +, -, or 0*
- Entity and Feature/Aspect Level: +, -, or 0*

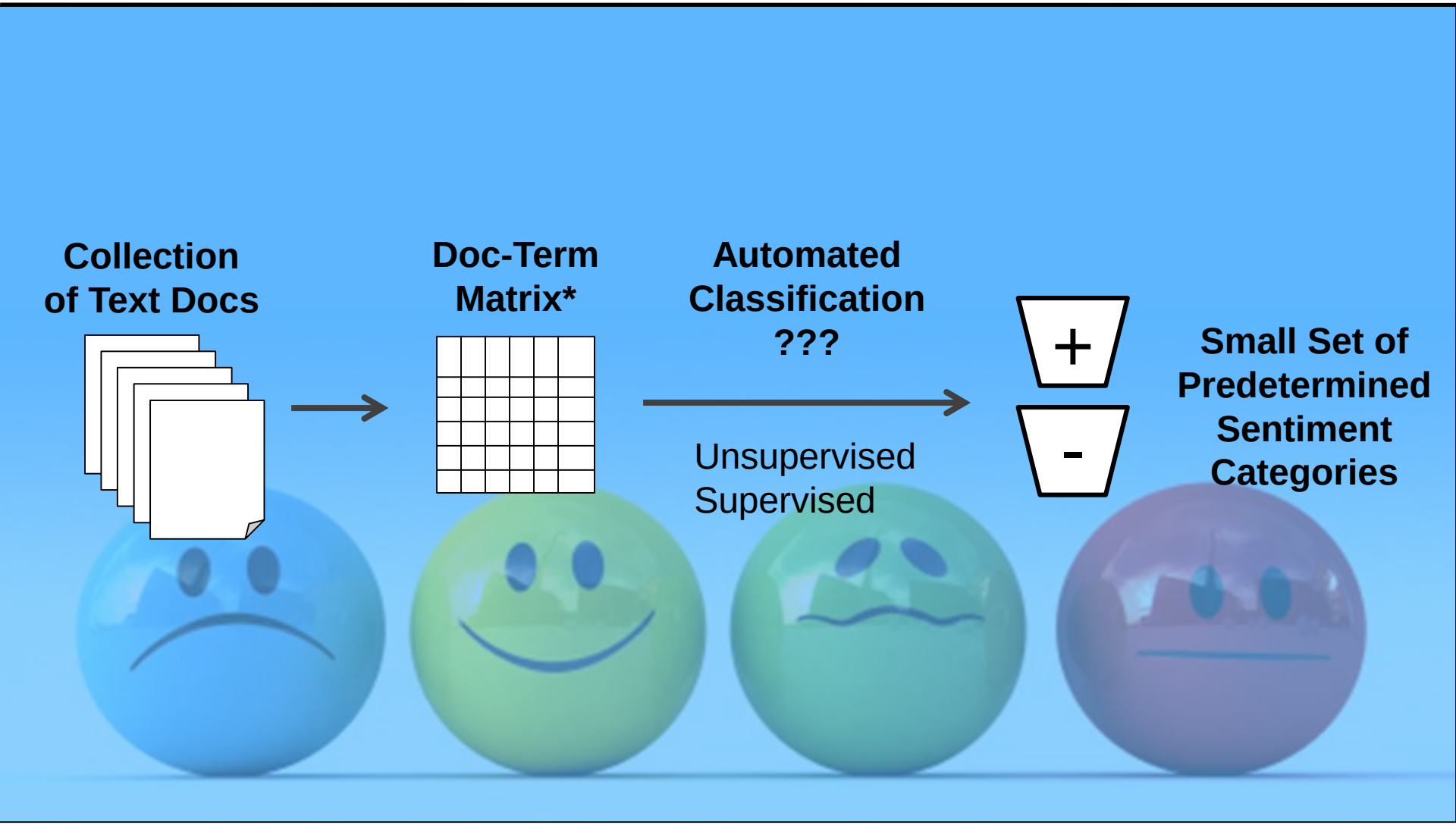
0* ~ possibly neutral

Sentiment Analysis and Opinion Mining: Example – Document Sentiment Classification

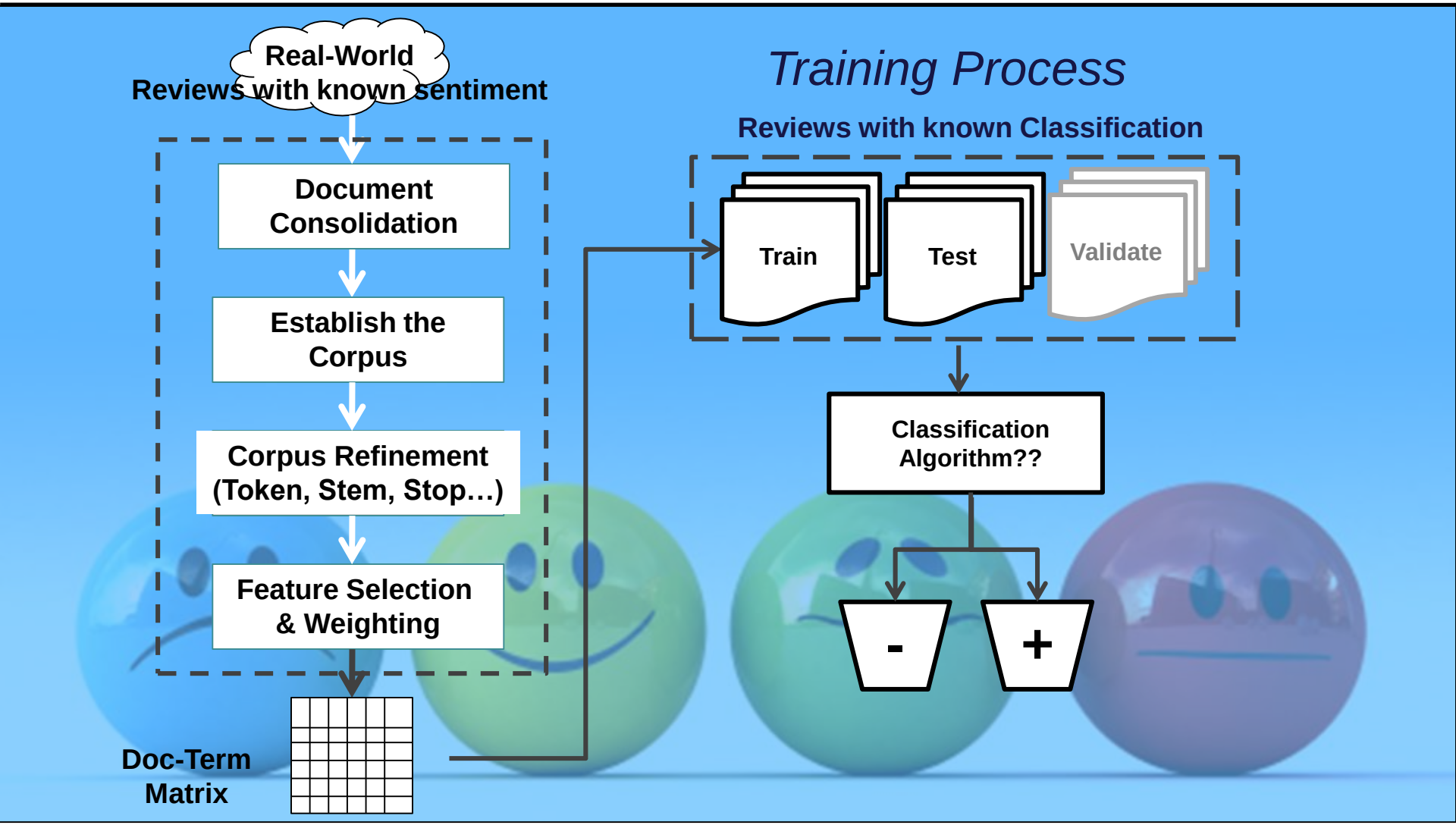
- Basically a Text Classification Problem
- Assumptions
  - Each document written by single person
  - About single entity
  - Goal: discover (_,_,so,_,_)



Sentiment Analysis and Opinion Mining: Example – Document Sentiment Classification



Sentiment Analysis and Opinion Mining: Example – Document Sentiment Classification

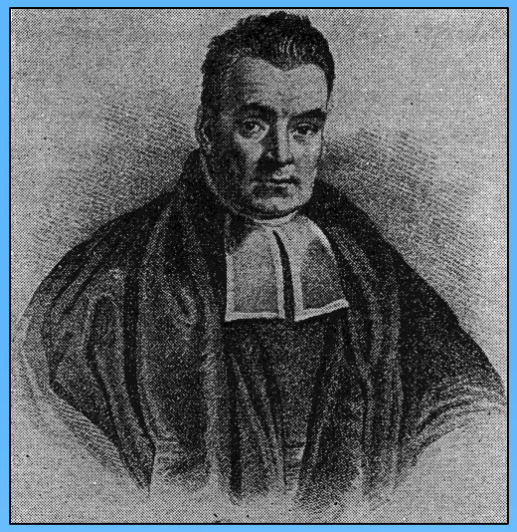


Sentiment Analysis and Opinion Mining: Example – Document Sentiment Classification

Supervised Classification Algorithms

- Naïve Bayes
- Support Vector Machine
- Decision Trees
- Nearest Neighbor (k-NN)
- Neural Nets (e.g. SOM)
- …

Sentiment Analysis:
Doing Simple Sentiment Analysis



Thomas Bayes

$$P(H/D) = P(D/H) * P(H)/P(D)$$

H is the hypothesis and **D** is the data

P(H) is the **prior probability** of *H*: the probability that *H* is correct before the data *D* are seen

P(D/H) is the **conditional probability** of seeing the data *D* given that the hypothesis *H* is true. This conditional probability is called the *likelihood*.

P(D) is the **marginal probability** of *D*.

P(H/D) is the **posterior probability**: the probability that the hypothesis is true, given the data and the previous state of belief about the hypothesis.

Sentiment Analysis:
Doing Simple Sentiment Analysis

P(Positive | Tweet) compared to P(Negative | Tweet)

Training Set

Review	Category
Never fails and easy to use.	Positive
Great camera, very easy to use	Positive
Very disappointed. Battery failed.	Negative
Poor picture quality	Negative
Quality was poor. Failed a lot.	Negative

P(Pos | Word) = P(Pos) * P(W1/Pos) / ~~P(M)~~
P(Pos| fail) = P(Pos) * P(great/Pos)
P(Pos | fail) = (2/5) * (1/2) = .2

P(Neg | Word) = P(N) * P(W1/N) / ~~P(M)~~
P(Neg | fail) = P(Neg) * P(great/Neg)
P(Neg| fail) = (3/5)*(2/3) = .4

Sentiment Analysis:
Doing Simple Sentiment Analysis

P(Positive | Tweet) compared to P(Negative | Tweet)

Training Set

Review	Category
Never fails and easy to use.	Positive
Great camera, very easy to use	Positive
Very disappointed. Battery failed.	Negative
Poor picture quality	Negative
Quality was poor. Failed a lot.	Negative

P(Pos | Words) = P(Pos) * P(W1/Pos) * P(W2/Pos) * ...
P(Pos | poor & fail) = P(Pos) * P(poor/Pos) * P(fail/Pos)
P(Pos | poor & fail) = .4 * 0 * .5 = 0

P(Neg | Words) = P(Neg) * P(W1/Neg) * P(W2/Neg) * ...
P(Neg | poor & fail) = P(Neg) * P(poor/Neg) * P(fail/Neg)
P(Neg | poor & fail) = .6 * .67 * .67 = .27

Sentiment Analysis:
Doing Simple Sentiment Analysis

How do you know if your model works? Depends on your Goal?

Confusion Matrix

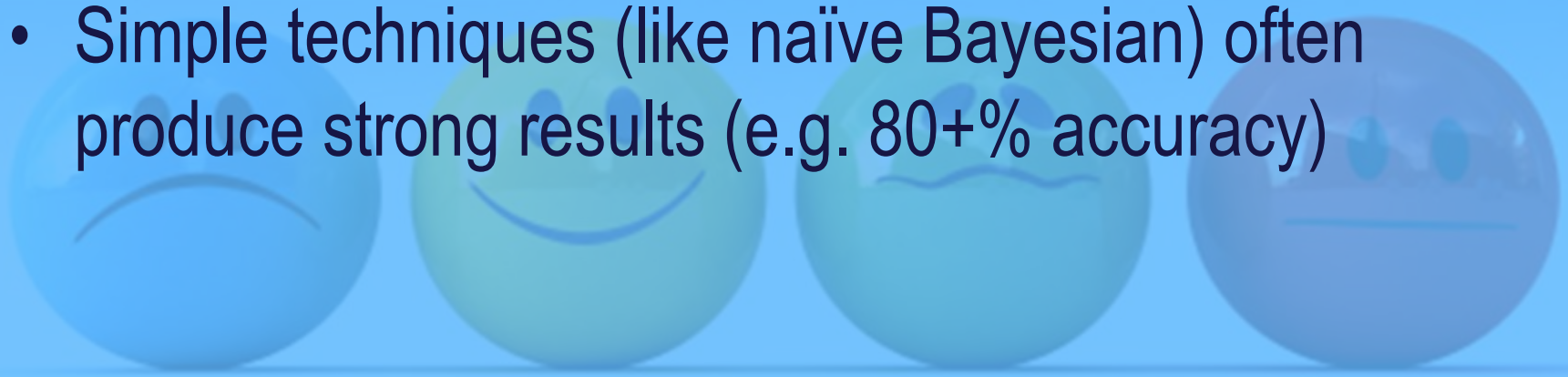
	Computer	
Human	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Accuracy = (TP + TN)/N
Recall = TP / (TP + FN)
Precision = TP / (TP + FP)
Error = (FP + FN)/N
F1 = 2*Recall*Precision/(Recall + Precision)

Where N = TP+FP+FN+TN

Sentiment Analysis: Summary

- From one type to the next (classification, features, comparisons), it becomes more complex to extract the information.
- Once extracted, standard text mining techniques can be used to classify and compare the opinions
- Simple techniques (like naïve Bayesian) often produce strong results (e.g. 80+% accuracy)



Sentiment Analysis: Comparing Techniques

Method	RT-s	MPQA	CR	Subj.
MNB-uni	77.9	85.3	79.8	**92.6**
MNB-bi	**79.0**	**86.3**	80.0	**93.6**
SVM-uni	76.2	86.1	79.0	90.8
SVM-bi	77.7	**86.7**	80.8	91.7
NBSVM-uni	**78.1**	85.3	80.5	92.4
NBSVM-bi	**79.4**	**86.3**	**81.8**	**93.2**
RAE	76.8	85.7	–	–
RAE-pretrain	77.7	**86.4**	–	–
Voting-w/Rev.	63.1	81.7	74.2	–
Rule	62.9	81.8	74.3	–
BoF-noDic.	75.7	81.8	79.3	–
BoF-w/Rev.	76.4	84.1	**81.4**	–
Tree-CRF	77.3	86.1	**81.4**	–
BoWSVM	–	–	–	90.0

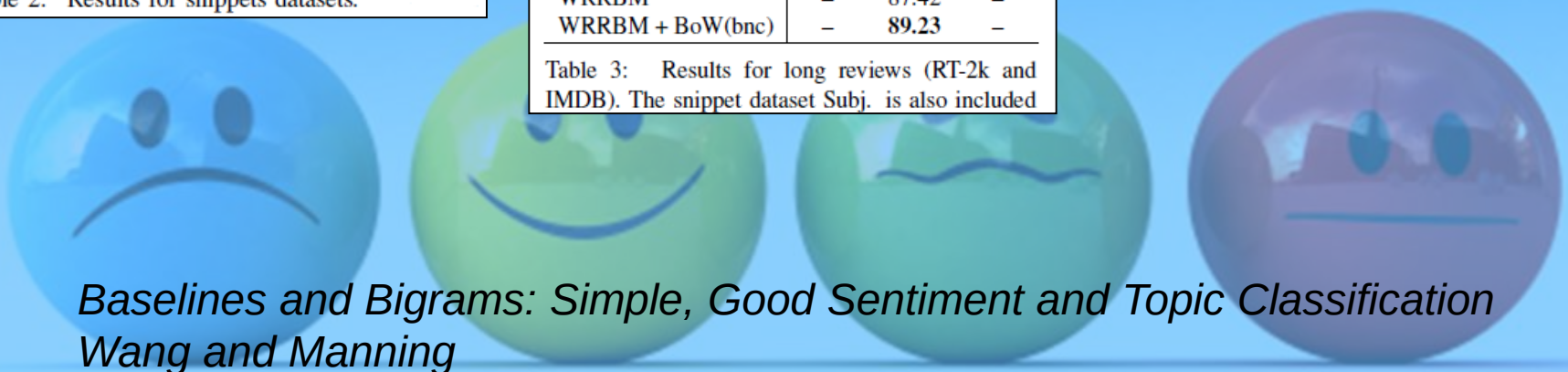
Table 2: Results for snippets datasets.

Our results	RT-2k	IMDB	Subj.
MNB-uni	83.45	83.55	**92.58**
MNB-bi	85.85	86.59	**93.56**
SVM-uni	86.25	86.95	90.84
SVM-bi	87.40	**89.16**	91.74
NBSVM-uni	87.80	88.29	92.40
NBSVM-bi	**89.45**	**91.22**	**93.18**
BoW (bnc)	85.45	87.8	87.77
BoW (bΔt′c)	85.8	88.23	85.65
LDA	66.7	67.42	66.65
Full+BoW	87.85	88.33	88.45
Full+Unlab'd+BoW	**88.9**	88.89	88.13
BoWSVM	87.15	–	90.00
Valence Shifter	86.2	–	–
tf.Δidf	88.1	–	–
Appr. Taxonomy	**90.20**	–	–
WRRBM	–	87.42	–
WRRBM + BoW(bnc)	–	**89.23**	–

Table 3: Results for long reviews (RT-2k and IMDB). The snippet dataset Subj. is also included

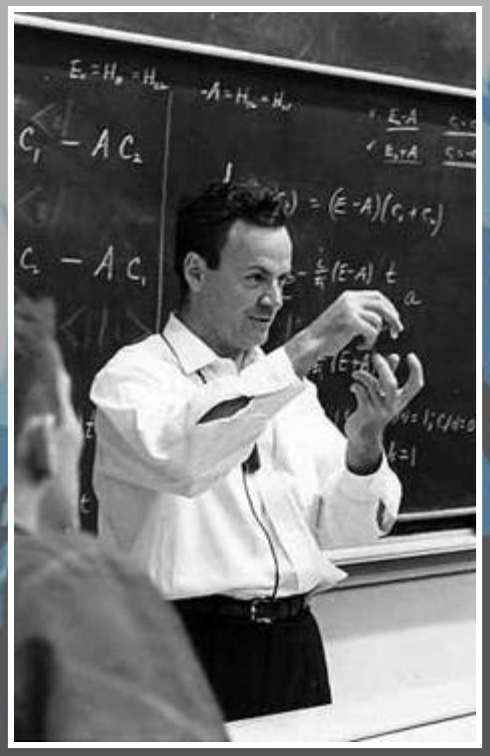
Method	AthR		XGraph		BbCrypt	
MNB-uni	85.0		90.0		**99.3**	
MNB-bi	**85.1**	+0.1	**91.2**	+1.2	99.4	+0.1
SVM-uni	82.6		85.1		98.3	
SVM-bi	83.7	+1.1	86.2	+0.9	97.7	−0.5
NBSVM-uni	**87.9**		**91.2**		99.7	
NBSVM-bi	**87.7**	−0.2	**90.7**	−0.5	99.5	−0.2
ActiveSVM	–		90		99	
DiscLDA	83		–		–	

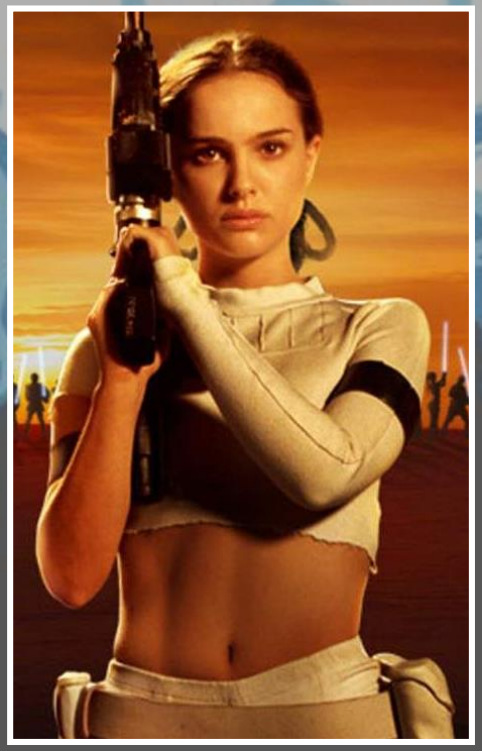
Table 4: On 3 20-newsgroup subtasks, we compare to DiscLDA (Lacoste-Julien et al., 2008) and ActiveSVM (Schohn and Cohn, 2000).

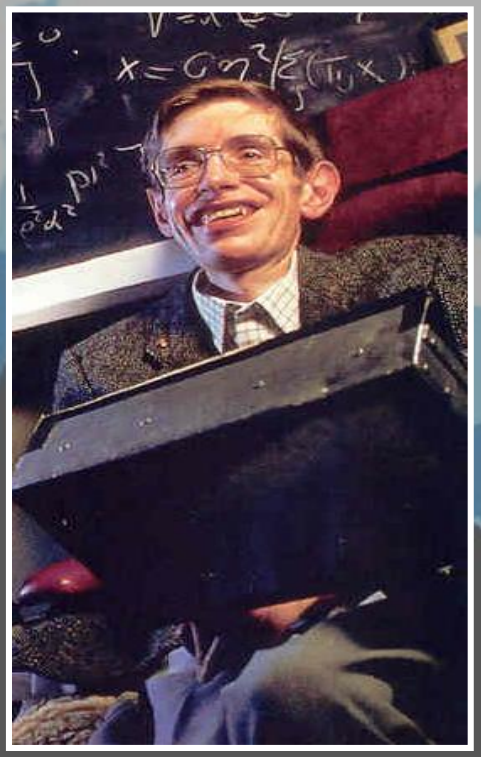


Baselines and Bigrams: Simple, Good Sentiment and Topic Classification
Wang and Manning

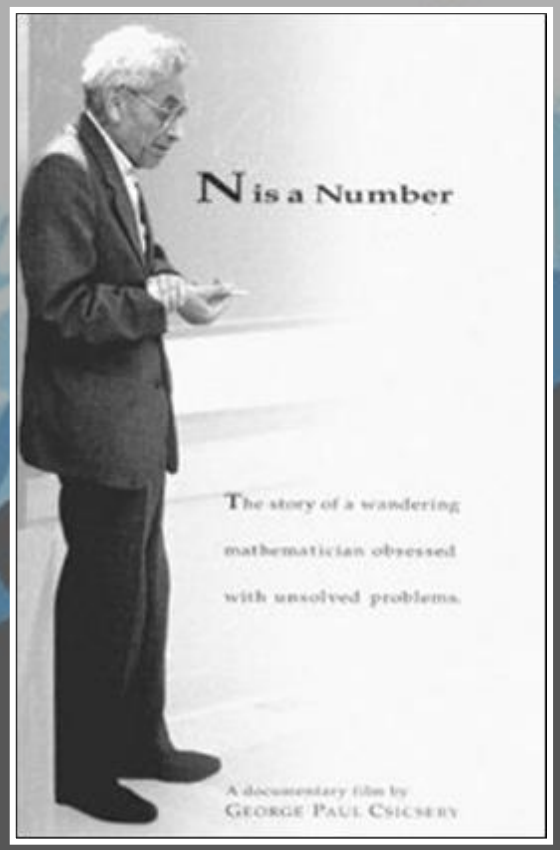
What do they have in common?

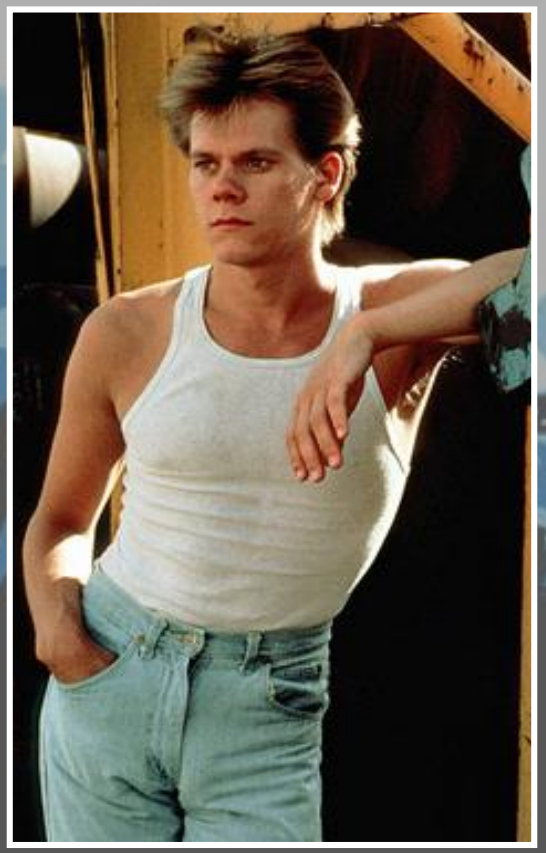




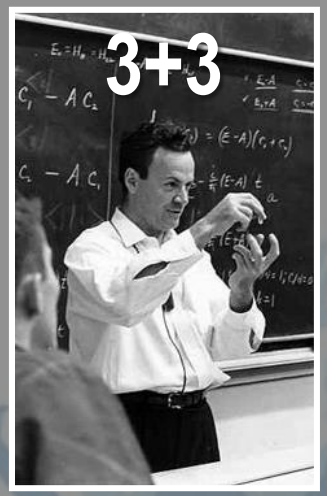


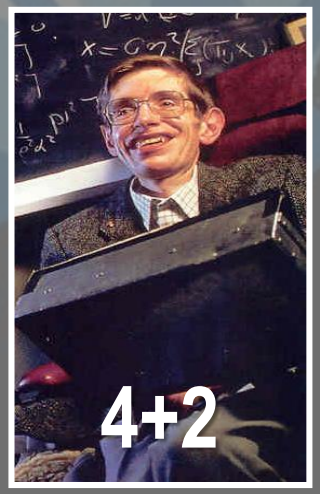
Here's a hint



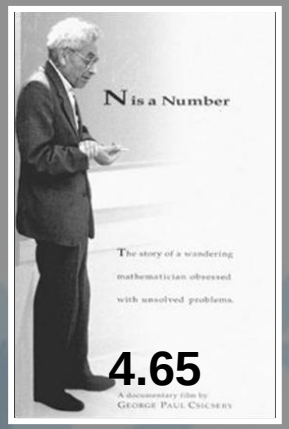


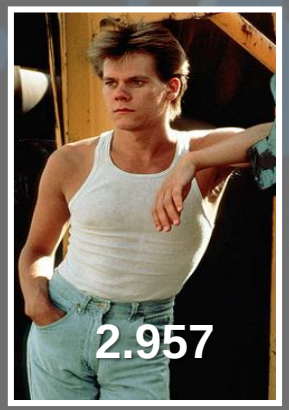
…and the answer is their Erdos-Bacon Number equals





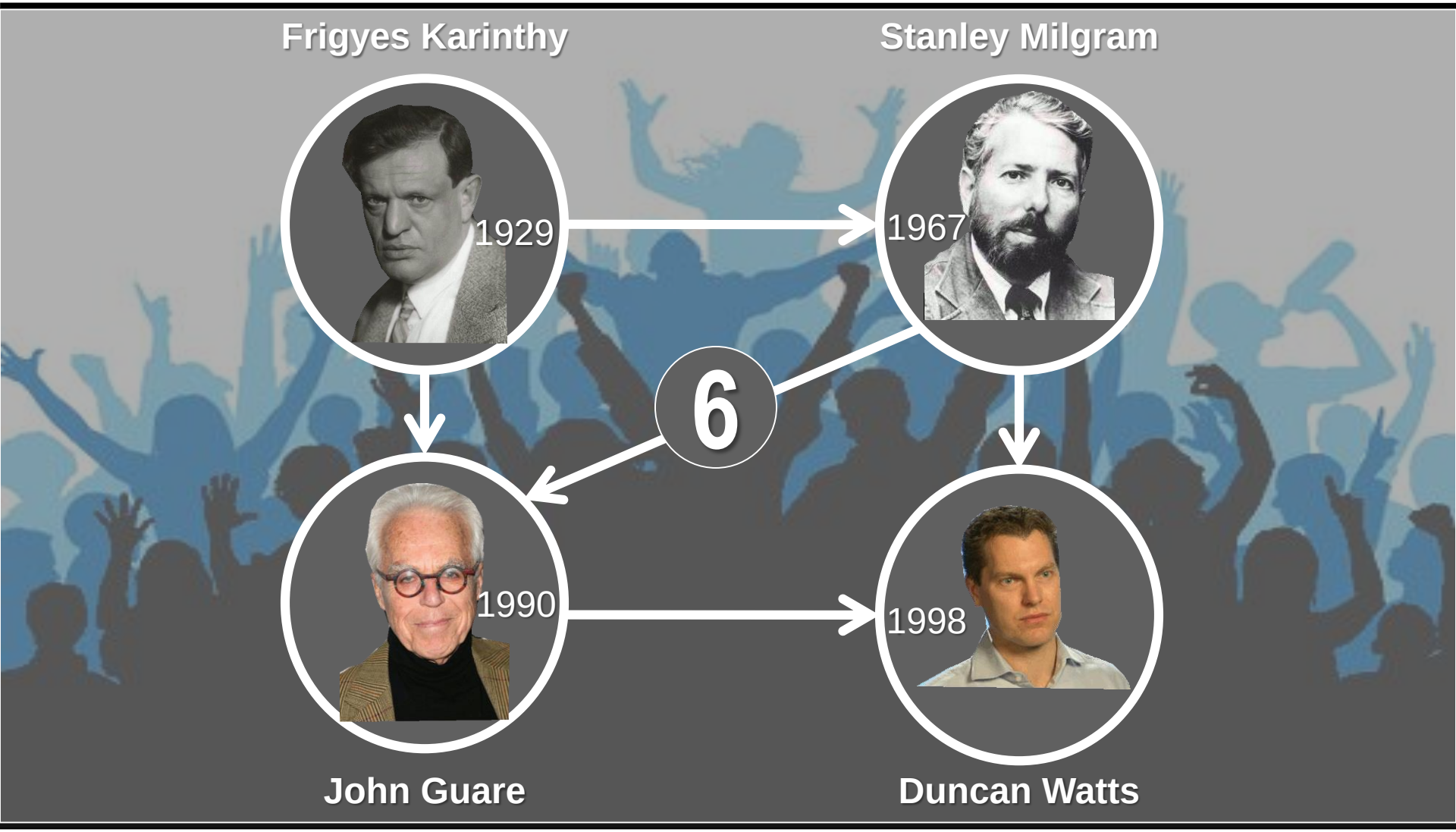
6 = +





Suppose I started with this.
What would you have guessed?

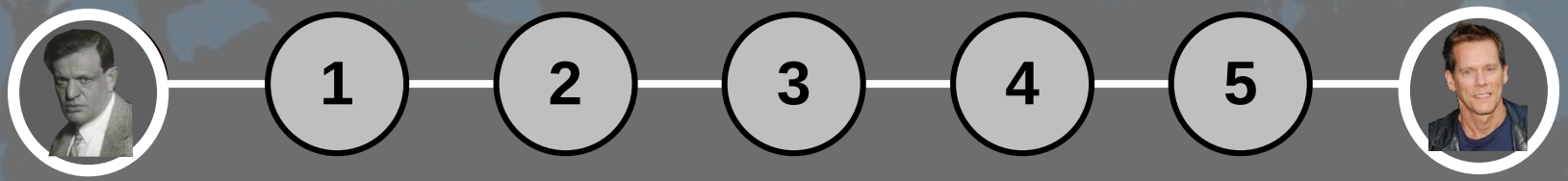
Six Degrees of Separation



Six Degrees of Separation

A fascinating game grew out of this discussion. One of us suggested performing the following experiment to prove that the population of the Earth is closer together now than they have ever been before. We should select any person from the 1.5 billion inhabitants of the Earth—anyone, anywhere at all. He bet us that, using no more than *five* individuals, one of whom is a personal acquaintance, he could contact the selected individual using nothing except the network of personal acquaintances.

Frigyes Karninthy , *Chains*, 1929



Degrees of separation ~ average path length ~ distance

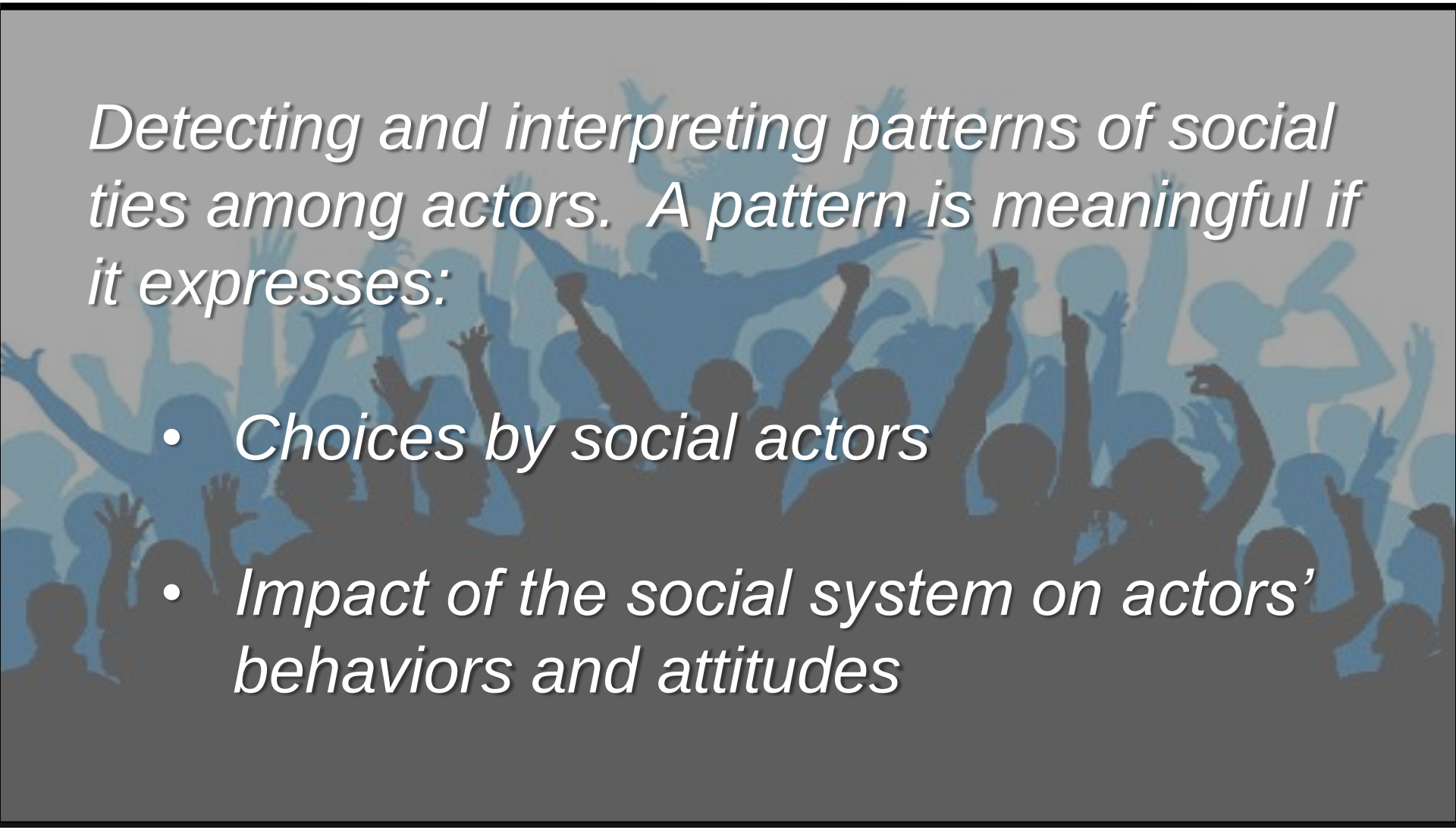
Social Network Analysis Definitions

Network – *Collection of things and their relationships to one another.*

Social Network – *Collection of humans, roles, groups, and/or institutions and their relationships with one another.*

Social Network Analysis (SNA) – *Application of Graph Theory or Network Science to the study of social relationships and connections.*

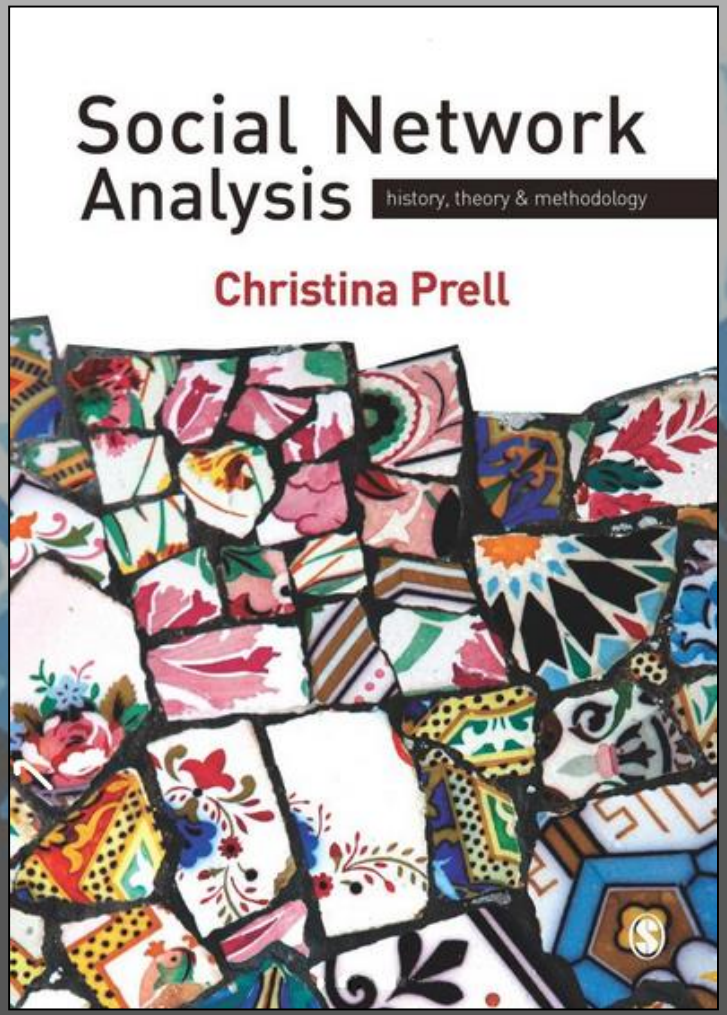
Social Network Analysis
Main Purpose



Detecting and interpreting patterns of social ties among actors. A pattern is meaningful if it expresses:

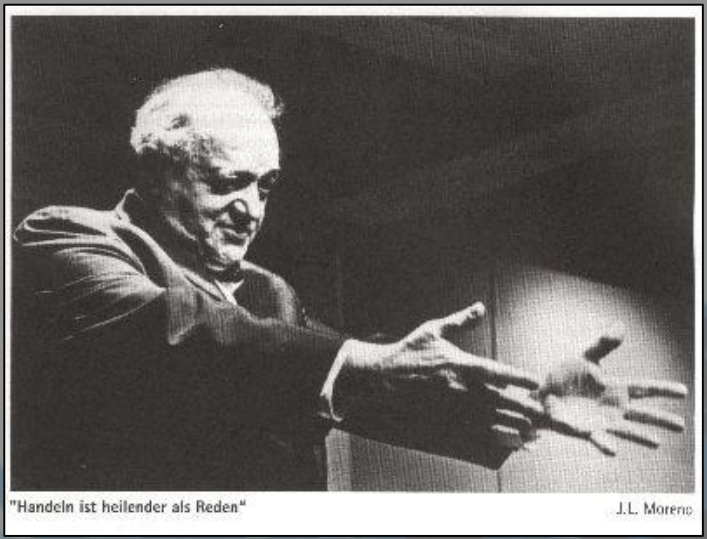
- *Choices by social actors*
- *Impact of the social system on actors' behaviors and attitudes*

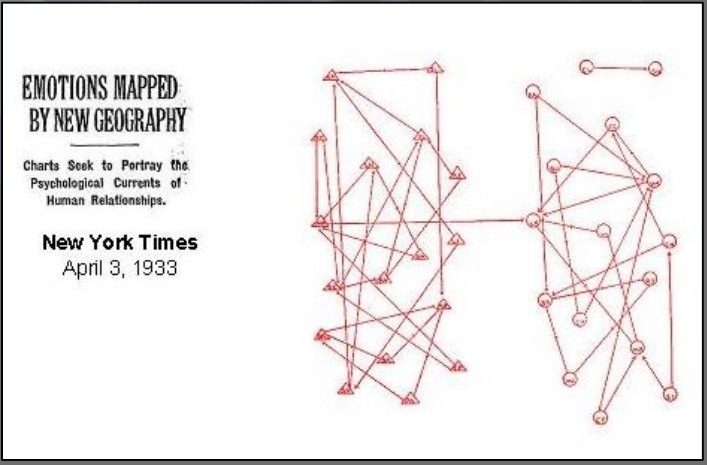
Social Network Analysis: Brief Highlights



Decade	Scholar(s)	Innovations
1900	Simmel	Dyads and Triads
1930	Jacob Moreno	Sociometry, Sociograms
1930	Mayo & Warner	Hawthorne Study
1940	Forsyth & Katz	(Adjacency) Matrix
1940	Luce & Festinger	Matrix Algebra, n-cliques
1940	Bavelas	Centrality, Centralization
1950	Radcliff-Browne	Social Structure as a Network of Social Relations
1950	Harary & Norman	Graph Theory, Structural Balance
1950	Manchester School	Ego Networks
1950	Bott	Connectedness, Density
1950	Barnes	Social Network'
1950	Homans	Social Exchange
1960	James Davis	Clustering, Transitivity
1960	Coleman	Diffusion in Social Networks
1960	Milgram	Small world
1970	Blau	Homophily
1970	White	Block models, Vacancy Chains
1970	Granovetter	Weak ties
1980	Holland & Leinhardt	Exponential Random Graph Models
1980	Frank & Strauss	Markov dependency grfaphs
1990	Friedkin	Social Influence Network Theory
1990	Bonacich	Eigenvector centrality, Power centrality
1990	Putnam	Social capital
1990	Watts & Strogatz	Small world simulation
2000	Snijders & Huisman	Longitudinal network data

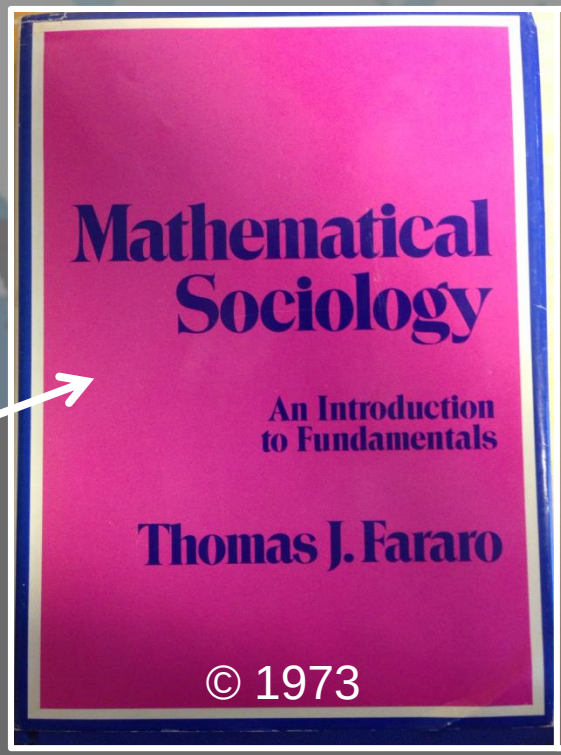
Social Network Analysis: Early Efforts

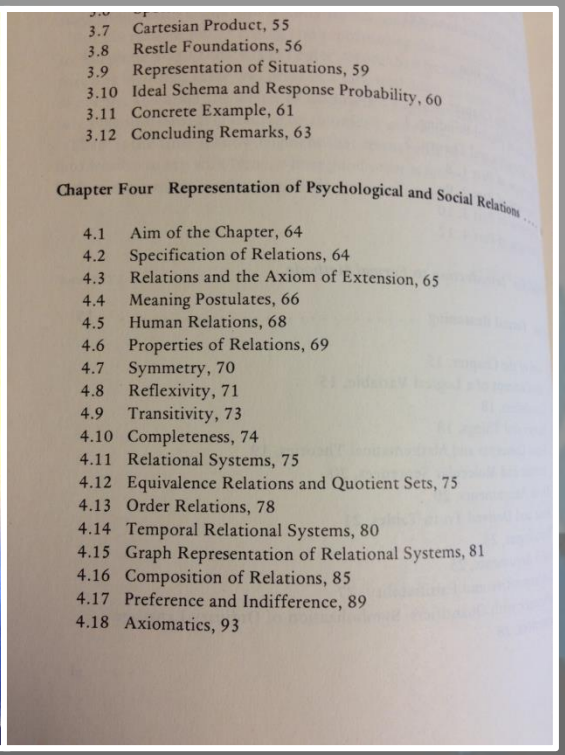

"Handeln ist heilender als Reden" J.L. Moreno



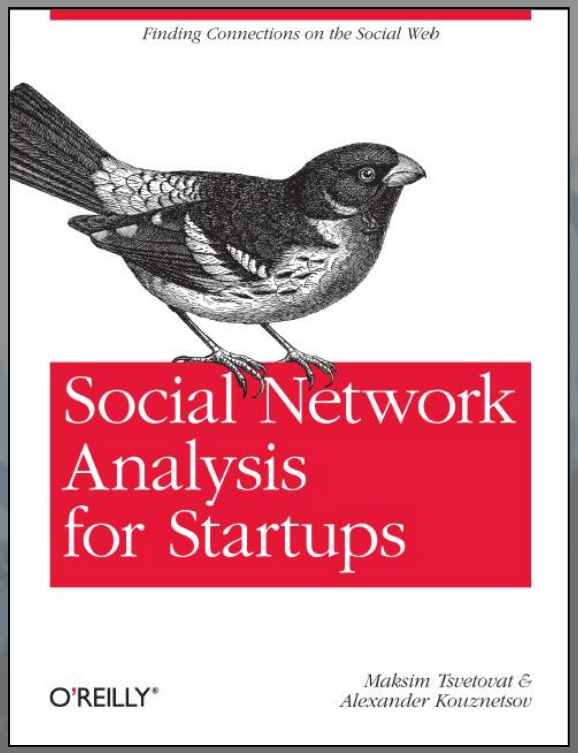
Decade	Scholar(s)	Innovations
1900	Simmel	Dyads and Triads
1930	Jacob Moreno	Sociometry, Sociograms
1930	Mayo & Warner	Hawthorne Study
1940	Forsyth & Katz	(Adjacency) Matrix
1940	Luce & Festinger	Matrix Algebra, n-cliques
1940	Bavelas	Centrality, Centralization
1950	Radcliff-Browne	Social Structure as a Network of Social Relations
1950	Harary & Norman	Graph Theory, Structural Balance
1950	Manchester School	Ego Networks
1950	Bott	Connectedness, Density
1950	Barnes	Social Network'
1950	Homans	Social Exchange
1960	James Davis	Clustering, Transitivity
1960	Coleman	Diffusion in Social Networks
1960	Milgram	Small world
1970	Blau	Homophily
1970	White	Block models, Vacancy Chains
1970	Granovetter	Weak ties
1980	Holland & Leinhardt	Exponential Random Graph Models
1980	Frank & Strauss	Markov dependency grfaphs
1990	Friedkin	Social Influence Network Theory
1990	Bonacich	Eigenvector centrality, Power centrality
1990	Putnam	Social capital
1990	Watts & Strogatz	Small world simulation
2000	Snijders & Huisman	Longitudinal network data

Social Network Analysis: Visualization/Analysis Libraries



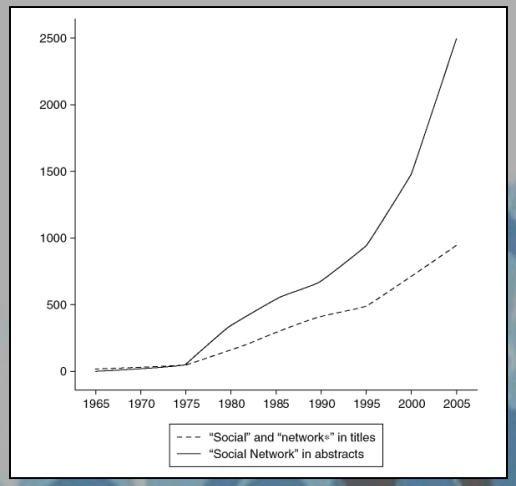


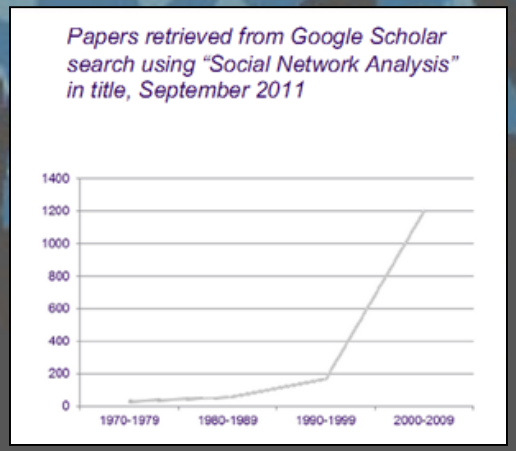
Social Network Analysis: Growing Interest

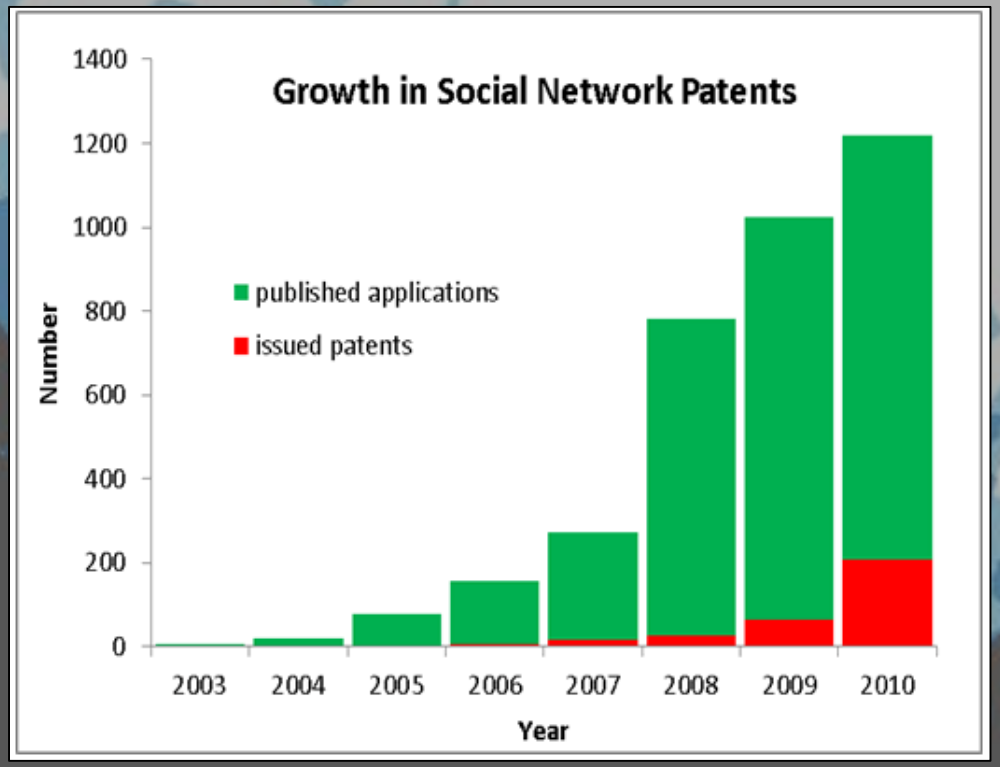


"Ten years ago, the field of Social Network Analysis was a scientific backwater. We were the misfits, rejected from both mainstream sociology and mainstream computer science… The advent of the Social Internet changed everything."

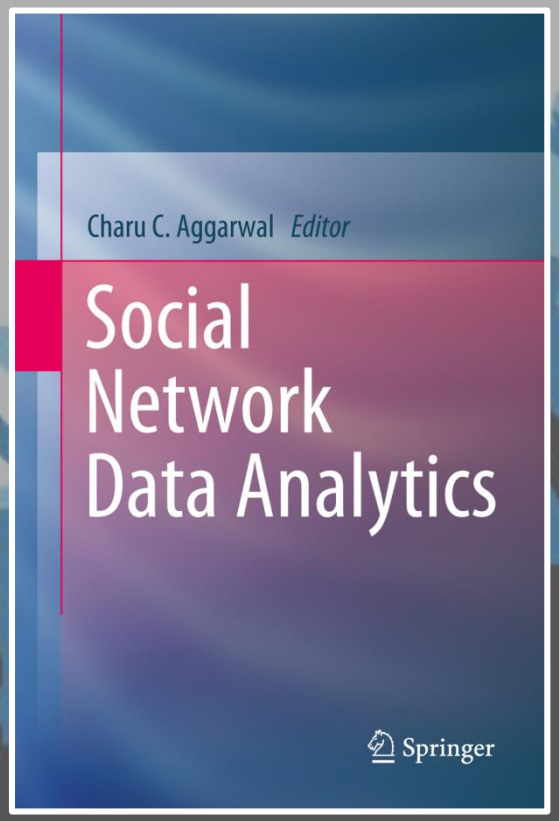
Social Network Analysis: Growing Interest







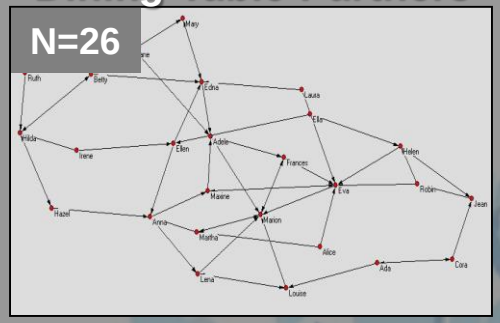
Social Network Analysis: Growing Interest



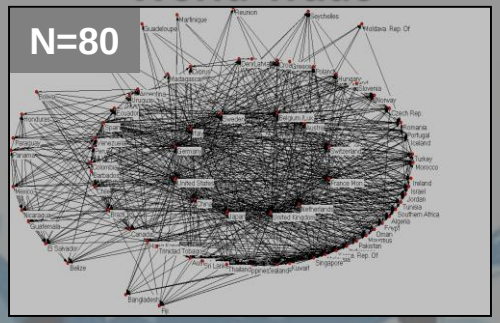
… the availability of massive amounts of data in an online setting has given a new impetus towards a scientifically and statistically robust study of the field of social networks

Social Network Analysis: Growing Interest

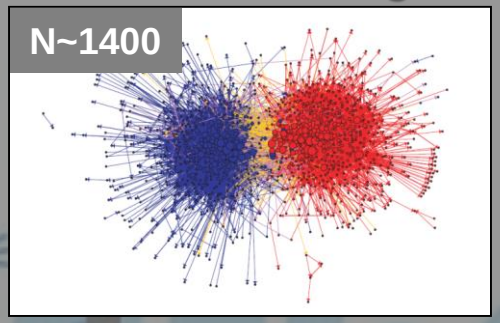
Dining Table Partners



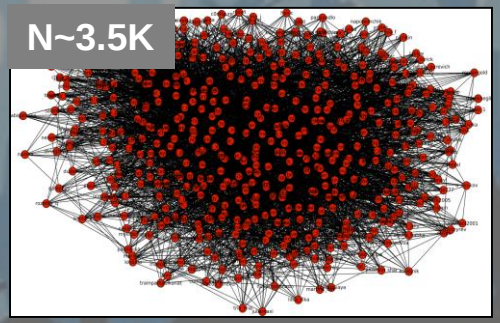
World Trade



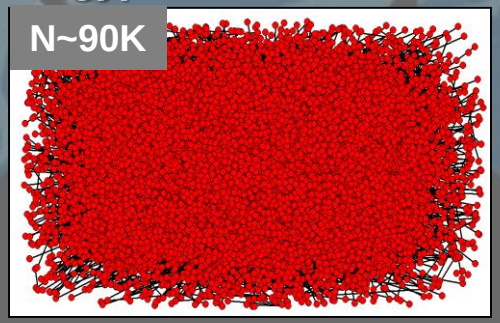
US Political Blogs



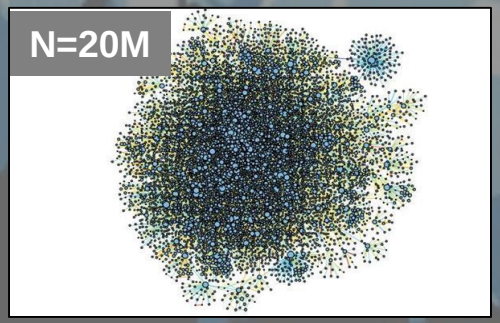
Russian LiveJournal



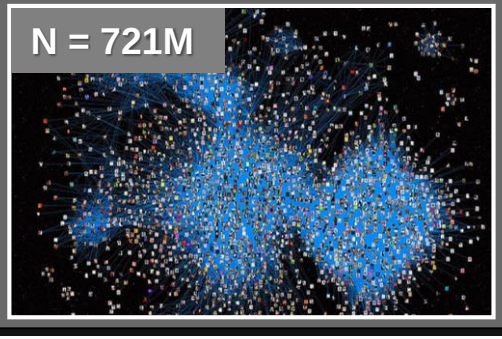
Egyptian Revolution



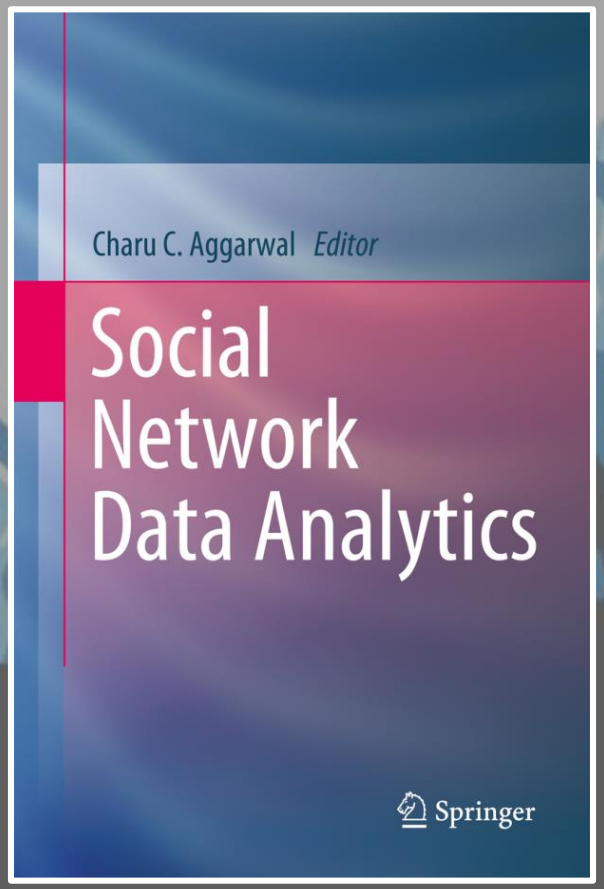
Mobile Phones



Facebook Friends

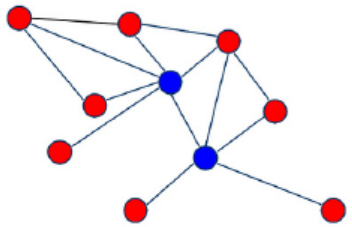


Social Network Analysis: Types of Structural Analysis



- Social Influence Analysis
- Expert Discovery
- Node Classification
- Link Prediction
- Community, Subgroup & Clique Detection in Social Networks
- Evolution in dynamic Social Networks
- Statistical Analysis and Comparison – Small Worlds, Weak Ties, and Random Models
- Visualization

Social Network Analysis: Introduction



Social Network Analysis (SNA)

including a tutorial on concepts and methods

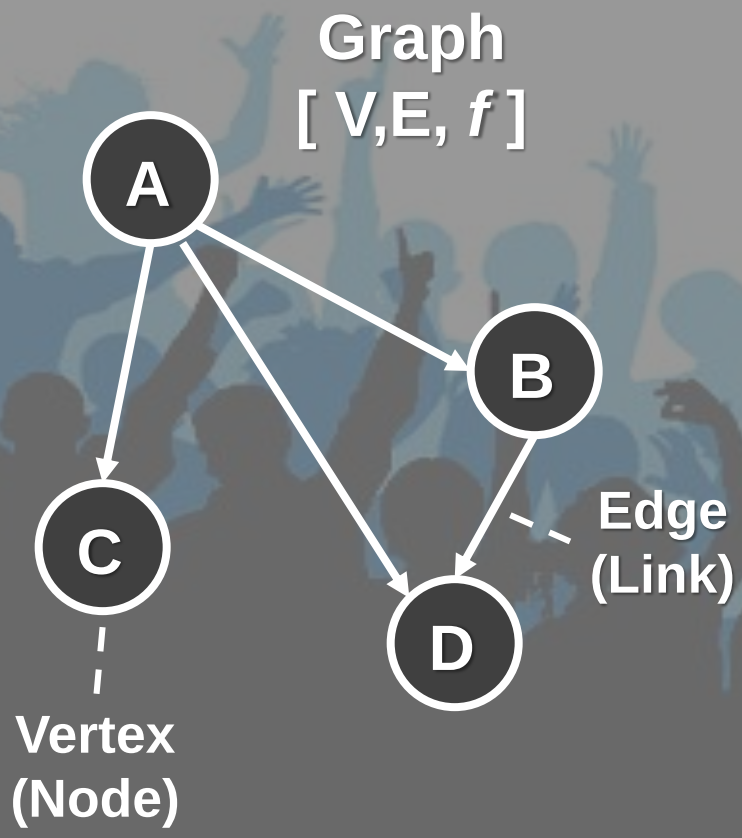
Social Media – Dr. Giorgos Cheliotis (gcheliotis@nus.edu.sg)
Communications and New Media, National University of Singapore

Social Network Analysis: Key Elements

Graph or Network
The set of vertices/nodes, edges/links and the relationship/function connecting them.

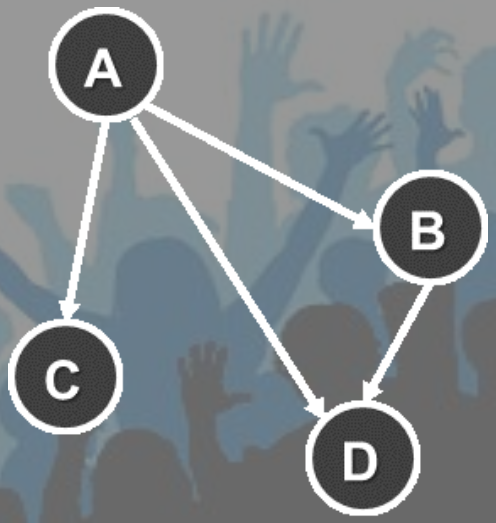
Vertices or Nodes
The "things"

Edges or Links
The "relationships"



Social Network Analysis: Alternative Representations

Graph



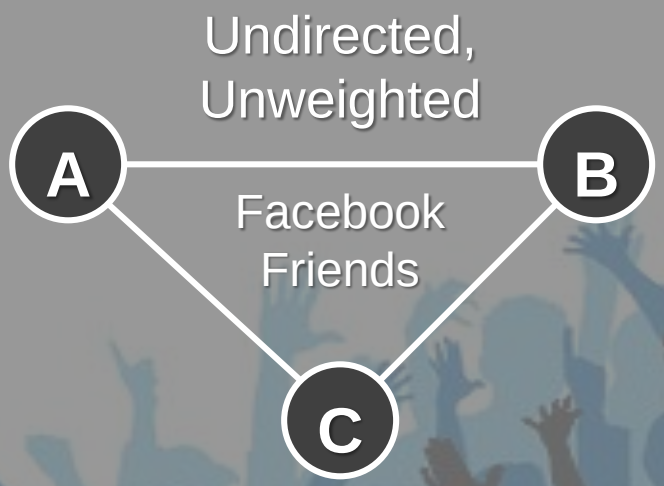
Edge List

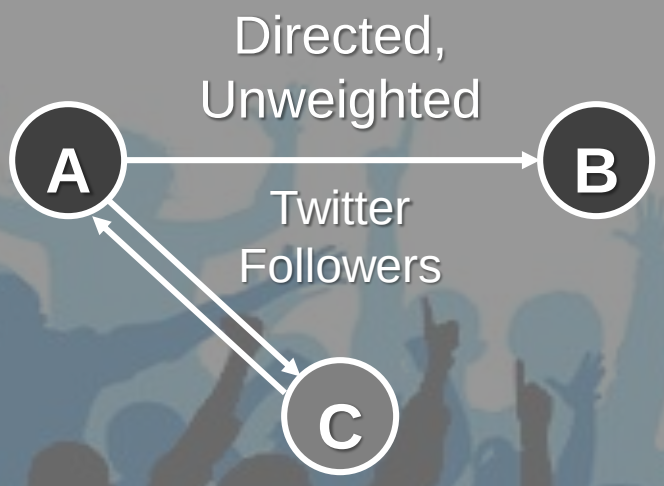
A	B
A	C
A	D
B	D

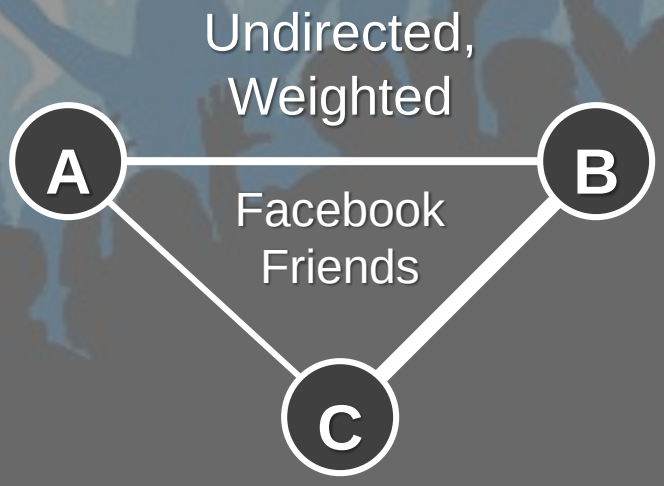
Adjacency Matrix

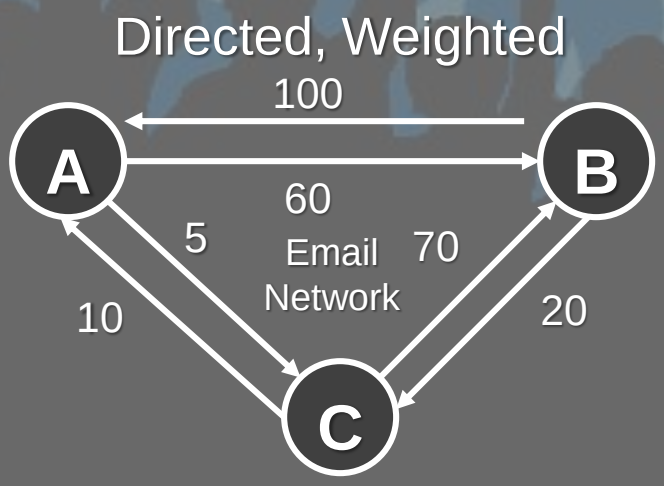
	A	B	C	D
A	-	1	1	1
B	0	-	0	1
C	0	0	-	0
D	0	0	0	-

Social Network Analysis: Types of Edges or Links



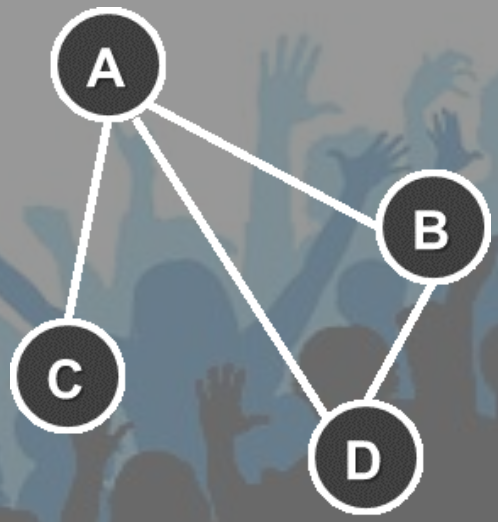






Social Network Analysis: Alternative Representations

Graph



Edge List

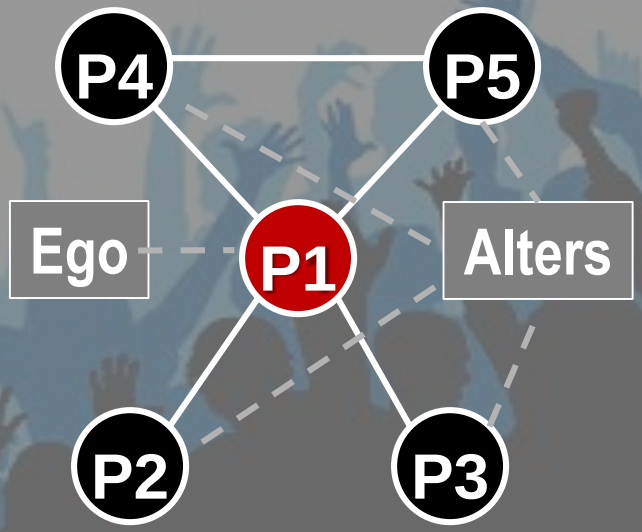
A	B
A	C
A	D
B	D

Adjacency Matrix

	A	B	C	D
A	-	1	1	1
B	1	-	0	1
C	1	0	-	0
D	1	1	0	-

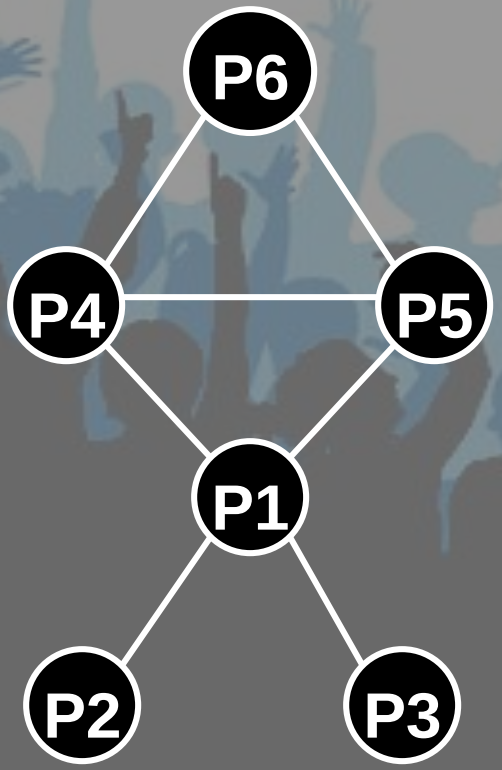
Social Network Analysis: Types of Networks & Approaches

Ego-Centered Approach "Ego-Network"



Vertex (Ego), Neighbors (Alters) & all lines among the Neighbors

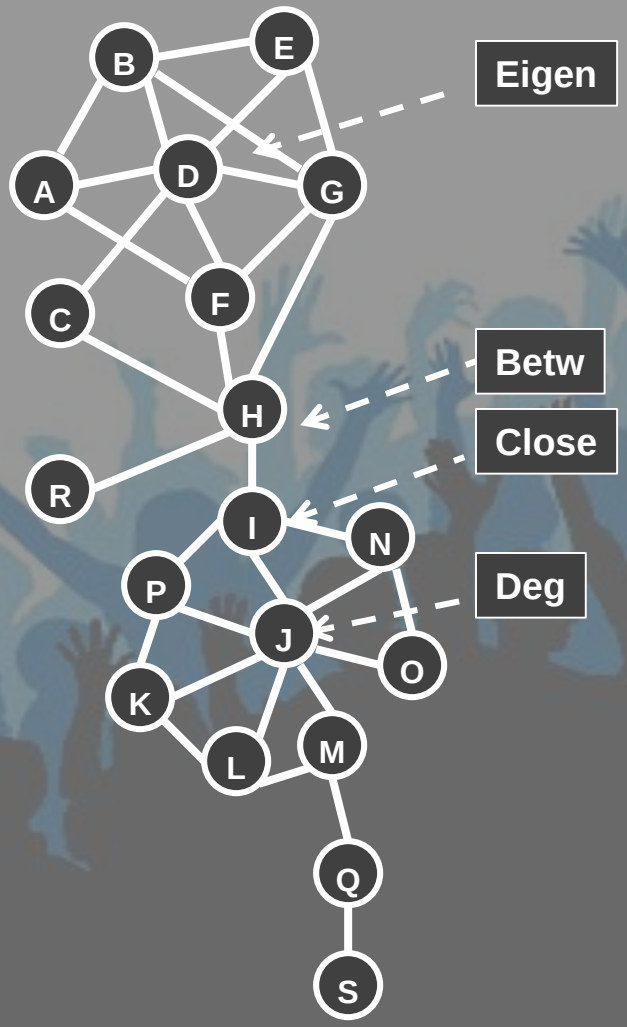
Socio-centered Approach "Whole" Network



Social Network Analysis: Centrality – Who is key?

Measure	Definition	Interpretation	Reasoning
Degree	Number of edges or links. In degree- links in, Out-degree - links out	How connected is a node? How many people can this person reach directly?	Higher probability of receiving and transmitting information flows in the network. Nodes considered to have influence over larger number of nodes and or are capable of communicating quickly with the nodes in their neighborhood.
Betweenness	Number of times node or vertex lies on shortest path between 2 nodes divided by number of all the shortest paths	How important is a node in terms of connecting other nodes? How likely is this person to be the most direct route between two people in the network?	Degree to which node controls flow of information in the network. Those with high betweenness function as brokers. Useful where a network is vulnerable.
Closeness	1 over the average distance between a node and every other node in the network	How easily can a node reach other nodes? How fast can this person reach everyone in the network?	Measure of reach. Importance based on how close a node is located with respect to every other node in the network. Nodes able to reach most or be reached by most all other nodes in the network through geodesic paths.
Eigenvector	Proporational to the sum of the eigenvector centralities of all the nodes directly connected to it.	How important, central, or influential are a node's neighbors? How well is this person connected to other well-connected people?	Evaluates a player's popularity. Identifies centers of large cliques. Node with more connections to higher scoring nodes is more important.

Social Network Analysis: Centrality – Who is most important?

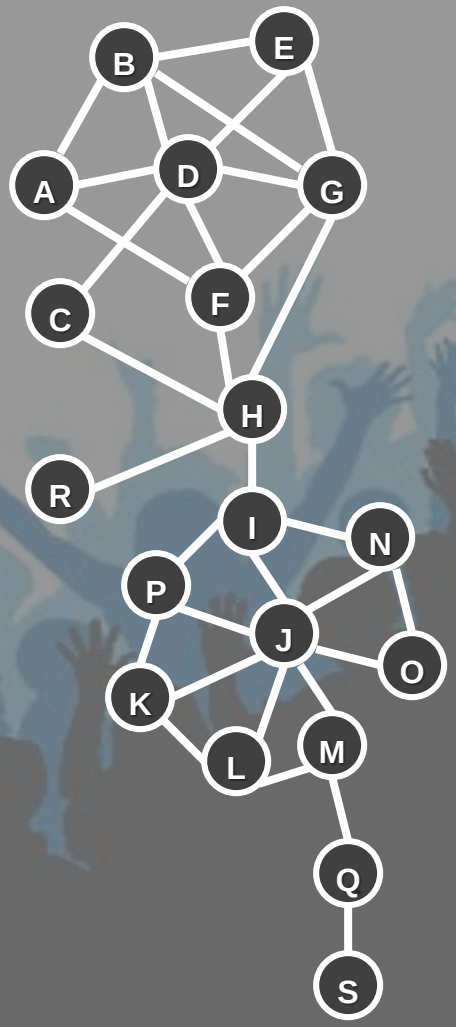


Node	Degree	Normed Degree	Betweenness	Closeness	Eigen Vector
A	3	0.17	0.00	0.29	0.29
B	4	0.22	0.01	0.30	0.36
C	2	0.11	0.03	0.35	0.18
D	6	0.33	0.04	0.31	0.46
E	3	0.17	0.00	0.29	0.30
F	4	0.22	0.11	0.36	0.35
G	5	0.28	0.19	0.37	0.43
H	5	0.28	0.58	0.45	0.28
I	4	0.22	0.53	0.46	0.13
J	7	0.39	0.43	0.43	0.12
K	3	0.17	0.00	0.32	0.06
L	3	0.17	0.01	0.33	0.05
M	3	0.17	0.21	0.33	0.04
N	3	0.17	0.03	0.38	0.07
O	2	0.11	0.00	0.31	0.05
P	3	0.17	0.03	0.38	0.08
Q	2	0.11	0.11	0.26	0.01
R	1	0.06	0.00	0.32	0.07
S	1	0.06	0.00	0.21	0.00

Social Network Analysis: Cohesion – Overall structure

Cohesion	Definition	Interpretation	Reasoning
Density	Ratio of the number of edges in the network over the total number of possible edges between all pairs of nodes	How well connected is the overall network?	Perfectly connected network is called a "clique" and has a density of 1.
Average Path Length (Distance)	Average number of edges or links between any two nodes (along the shortest path)	On average, how far apart are any two nodes?	This is synonymous with the "degrees of separation" in a network.
Diameter	Longest (shortest path) between any two nodes	At most, how long will it take to reach any node in the network? Sparse networks usually have greater diameters.	Measure of the reach of the network
Clustering	A node's *clustering coefficient* is the density of it's 1.5 degree egocentric network (ratio of connecting among ego's alters). For entire network it is the average of all the coefficients for the individual nodes.	What proportion of ego's alters are connected? More technically, how many nodes form triangular subgraphs with their adjacent nodes?	Measures certain aspects of "cliquishness." Proportion of your friends that are also friends with each other. Another way to measure is to determine (in a undirected) graph the ratio of the number of times that two links eminating from the same node are also linked.
Centralization	Normalize ratio of the sum of the variances of the centrality of each node from the most central node to the maximum sum possible	Indicates how unequal the distribution of centrality is in a network.	Measures how much variance there is in the distribution of centrality in a network. The measure applies to all forms of centrality.

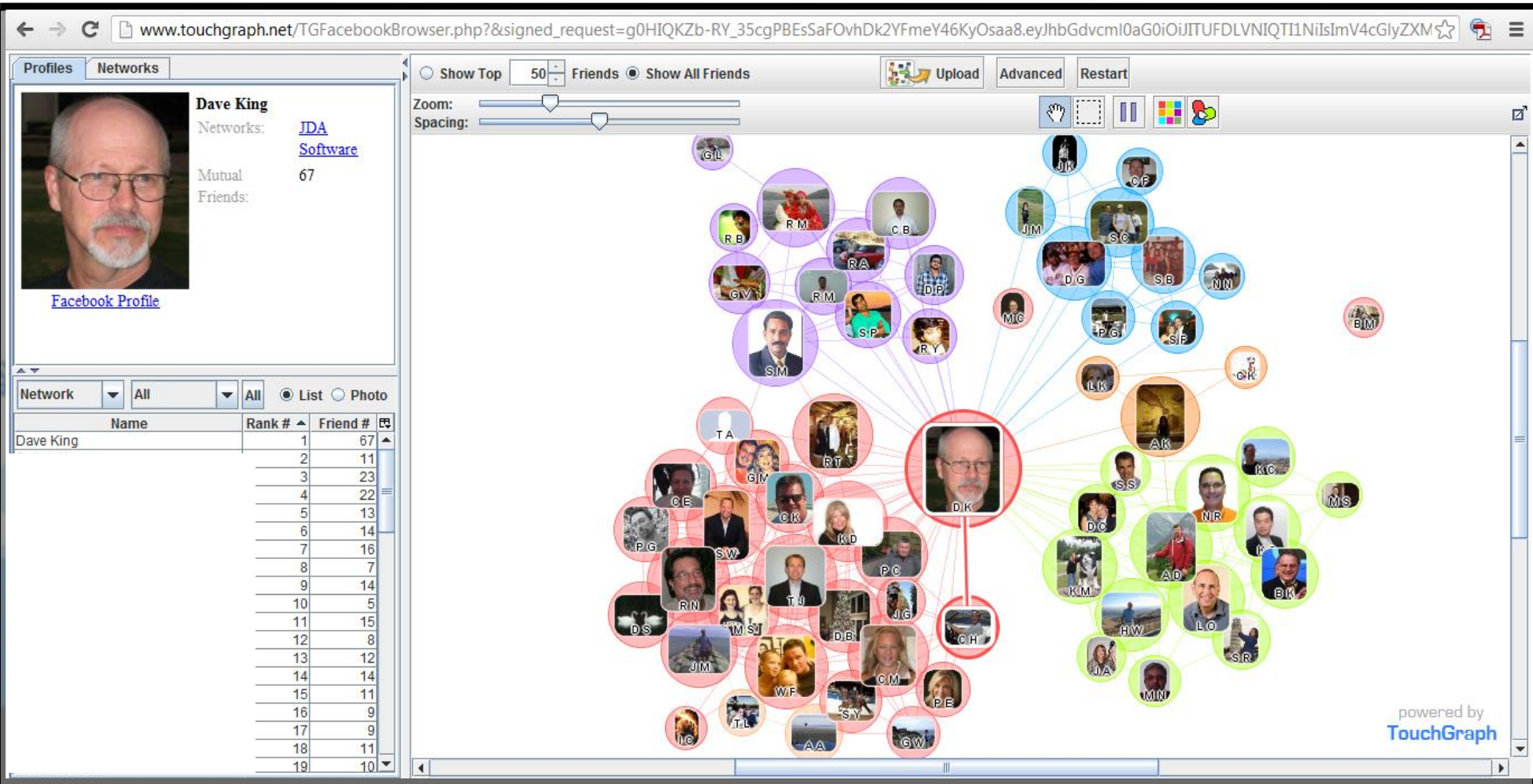
Social Network Analysis: Cohesion – How well connected?



Measure	Value
Density	0.19
Average Degree	3.37
Average Distance	3.06
Diameter	8
Clustering Coefficient	0.43
Degree Centralization	0.22
Betweenness Centralization	0.48
Closeness Centralization	0.27
Eigenvector Centralization	0.56

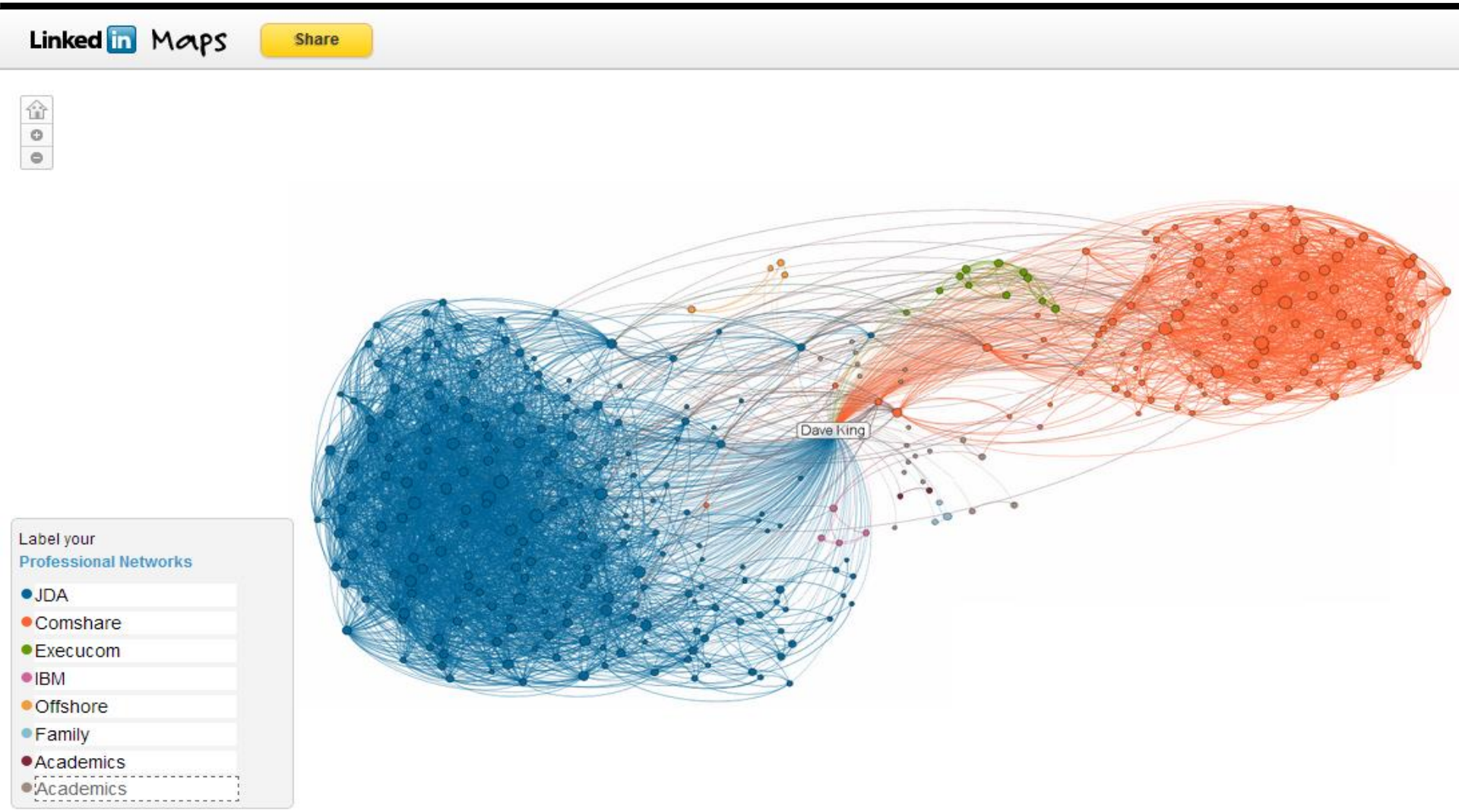
Node	Clustering
A	0.67
B	0.67
C	0.00
D	0.40
E	1.00
F	0.50
G	0.50
H	0.10
I	0.33
J	0.29
K	0.67
L	0.67
M	0.33
N	0.67
O	1.00
P	0.67
Q	0.00
R	NA
S	NA

Social Network Analysis: Ego Centered – Simple Example

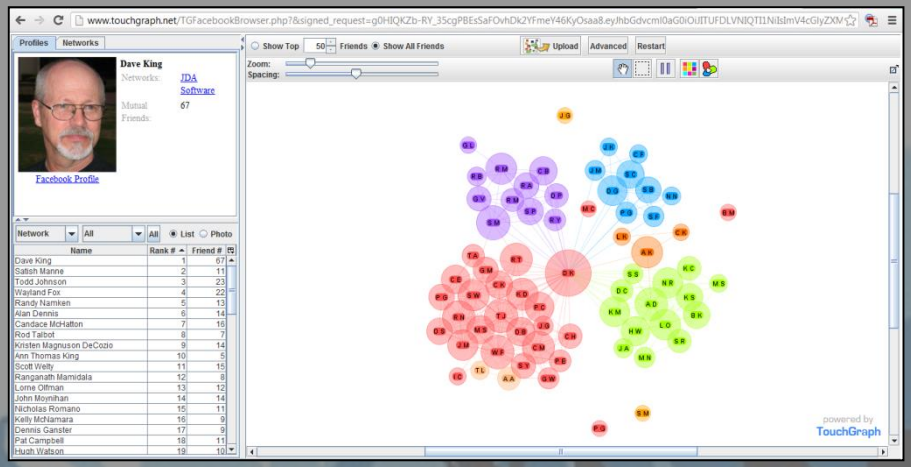


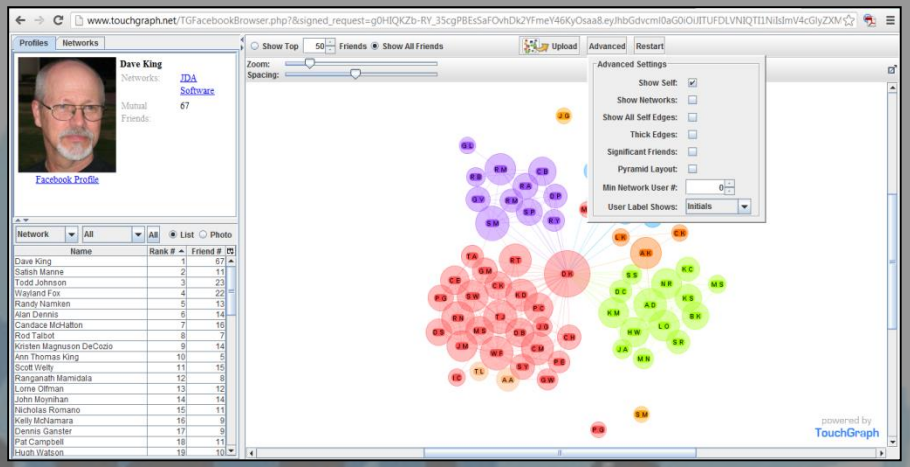
http://apps.facebook.com/touchgraph/

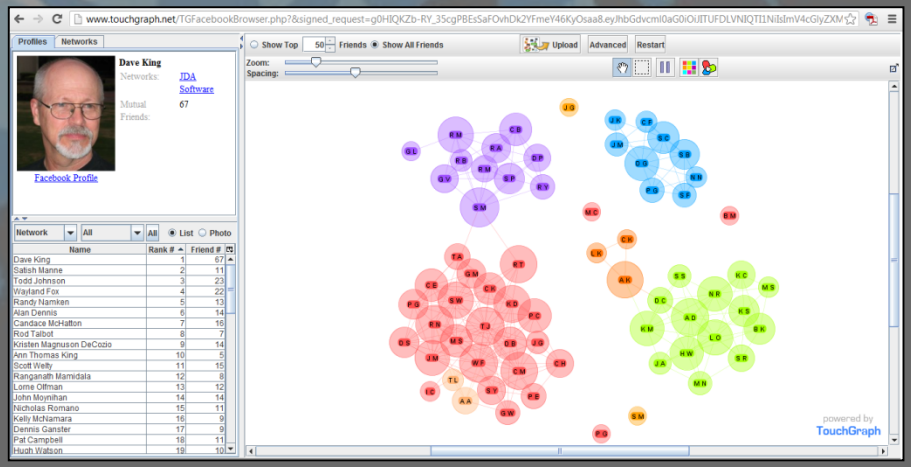
Social Network Analysis: Ego Centered – Another Example

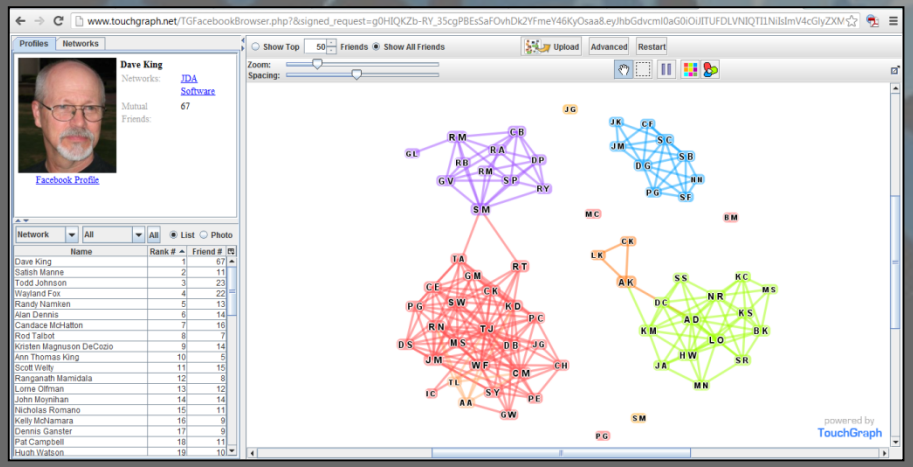


Social Network Analysis: Ego-Centered – Simple Example



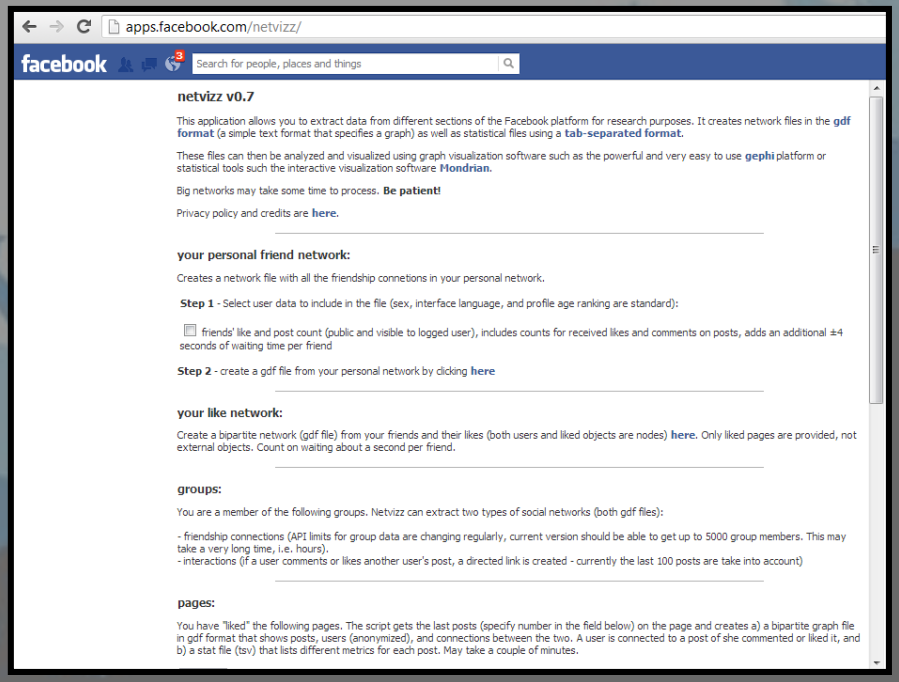






Social Network Analysis: Ego Analysis – Simple Example

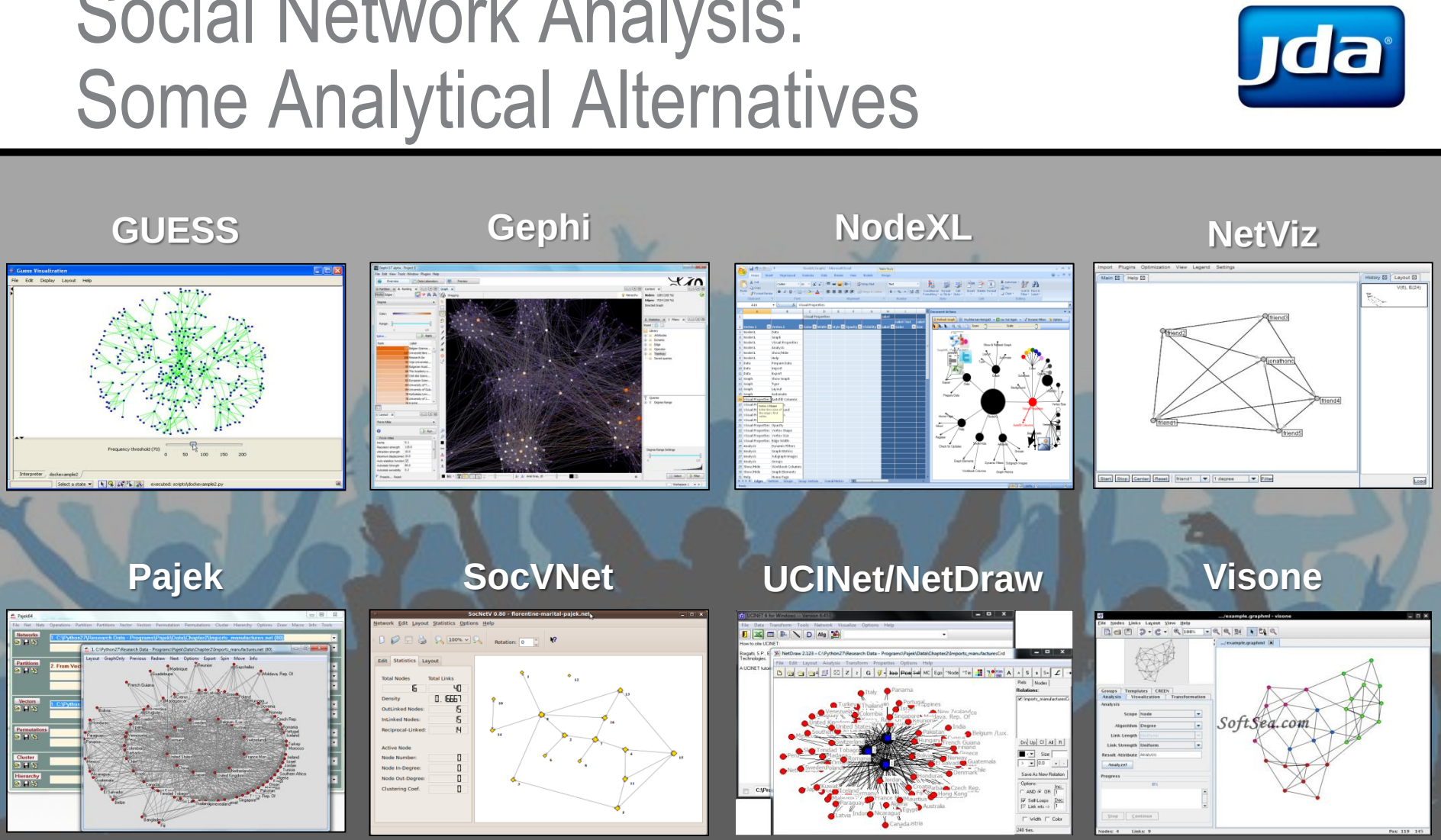
Netvizz7.0



.gdf (GUESS) file format

```
nodedef>name VARCHAR,label VARCHAR,sex VARCHAR,locale VARCHAR,agerank INT,
like_count INT,post_count INT,post_like_count INT,
post_comment_count INT,post_engagement_count INT
4945386,HW,male,en_US,67,6,3,11,,1
13307576,LO,male,en_US,66,19,21,54,,0
512057631,AA,male,en_US,65,20,37,32,,1
534718919,CB,male,en_US,64,38,239,1384,,2
539959548,RA,male,en_US,63,69,448,1737,,1
543650513,SR,female,en_US,62,0,1,2,,3
558159260,NR,male,en_US,61,15,262,148,,5

...

100001597623987,SM,male,en_US,4,3,22,99,,1
100003001526236,SP,male,en_US,3,0,3,2,,1
100003603794852,CF,male,en_US,2,5,36,24,,0
100004022119721,JG,female,en_US,1,0,0,,,0
edgedef>node1 VARCHAR,node2 VARCHAR
4945386,13307576
4945386,558159260
4945386,601254884
4945386,626874213
4945386,662011575

...

100000599719815,100001132688907
100000634883450,100001132688907
100001132688907,100001378371279
100001305975715,100001597623987
100001305975715,100003001526236
100001393609640,100003603794852
100001597623987,100003001526236
```

Social Network Analysis: Some Analytical Alternatives



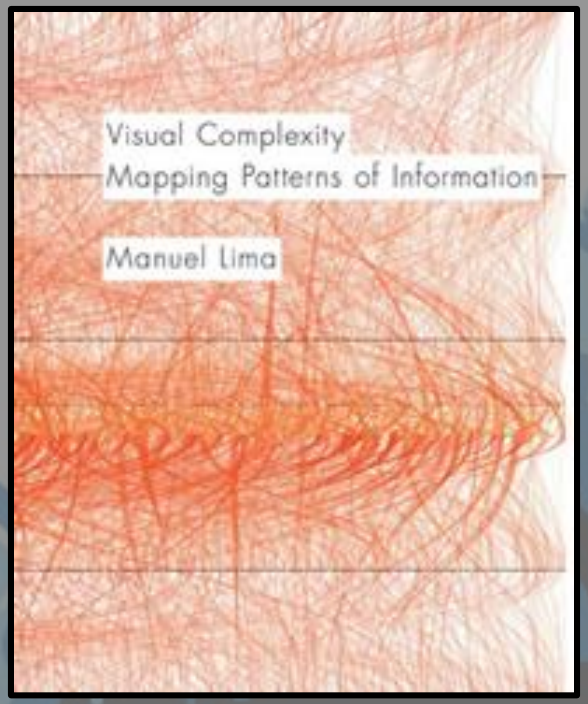
Visual/Analytical Packages

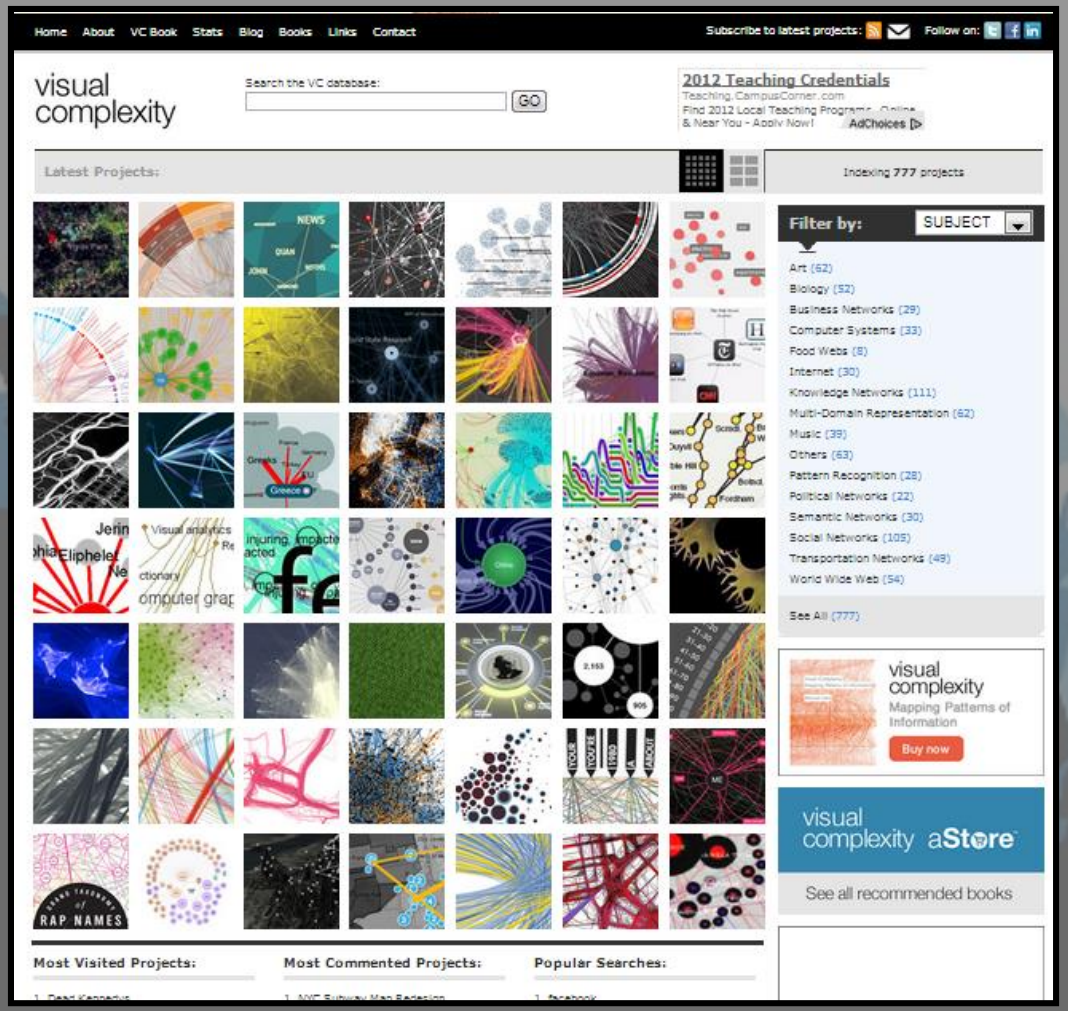
Social Network Analysis: Some Analytical Alternatives

- igraph (R, Python, C): Creating and manipulating graphs
- libSNA (Python): Open-source library for social network analysis (2008 last update)
- NetworkX (Python): Package for complex networks
- SNA (R): Social Network Analysis tools

Visual/Analytical Libraries/Modules

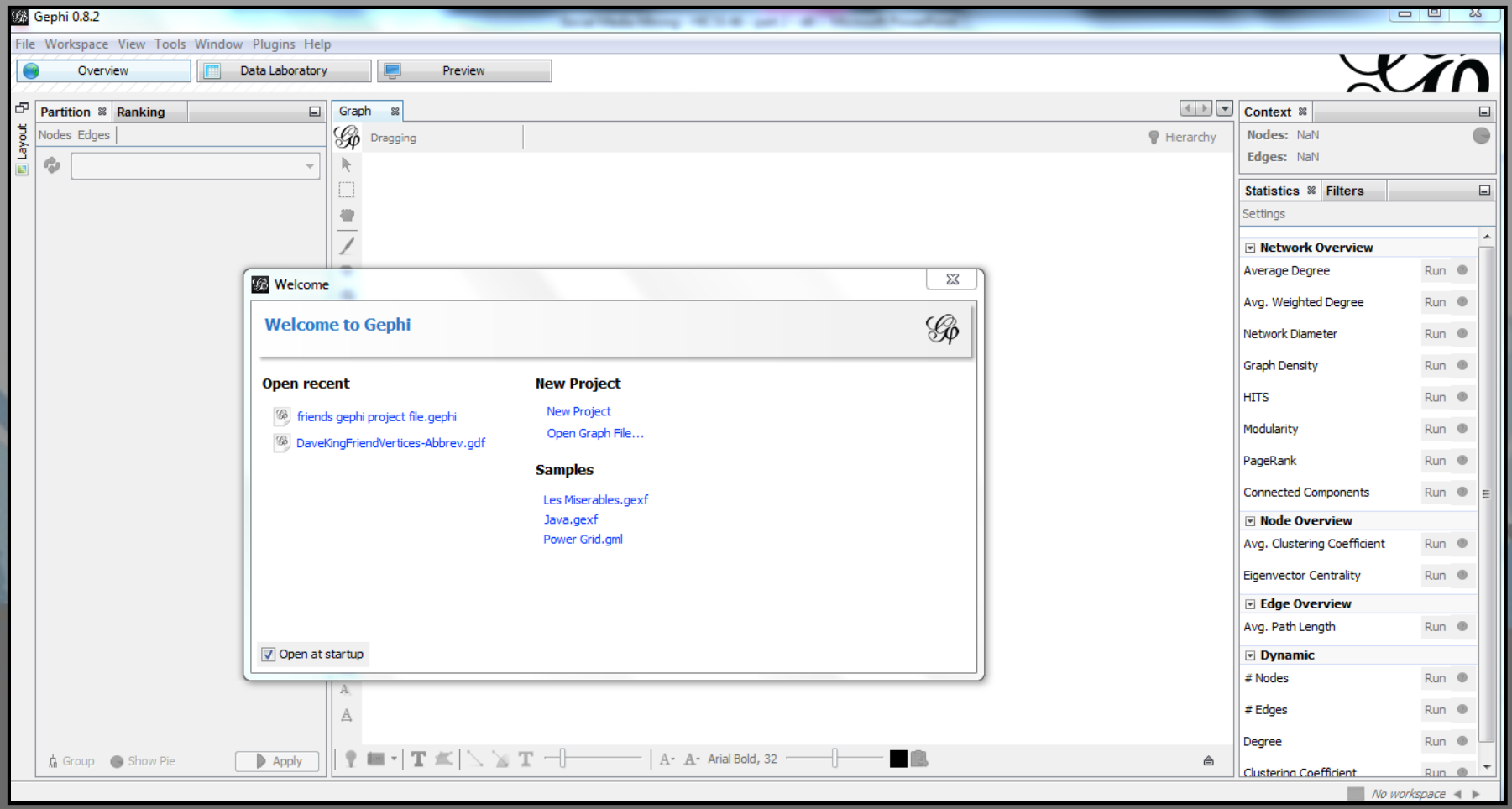
Social Network Analysis: Some Analytical Alternatives



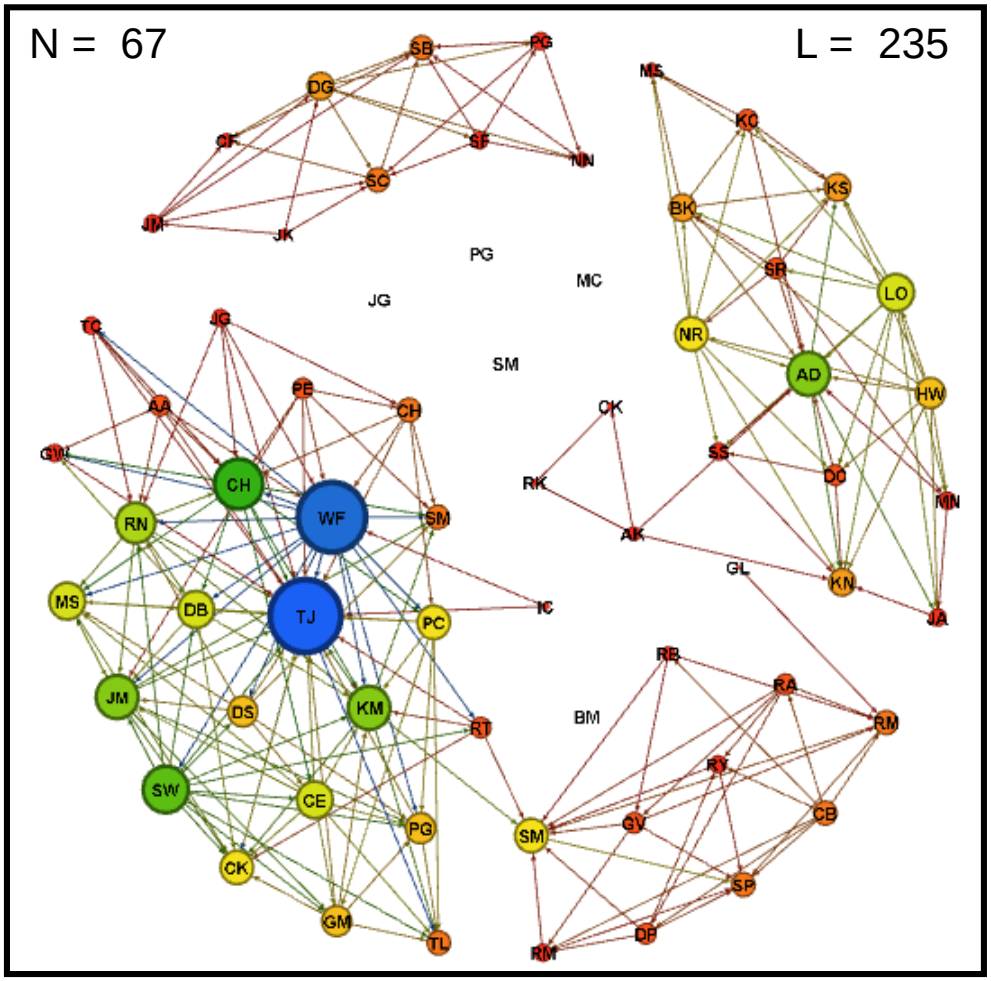


Social Network Analysis: Ego Analysis – Simple Example

Social Network Analysis: Ego Analysis – Simple Example



Social Network Analysis: Ego Analysis – Simple Example



Label	Degree
TJ	22
WF	21
CH	15
SW	14
AD	13

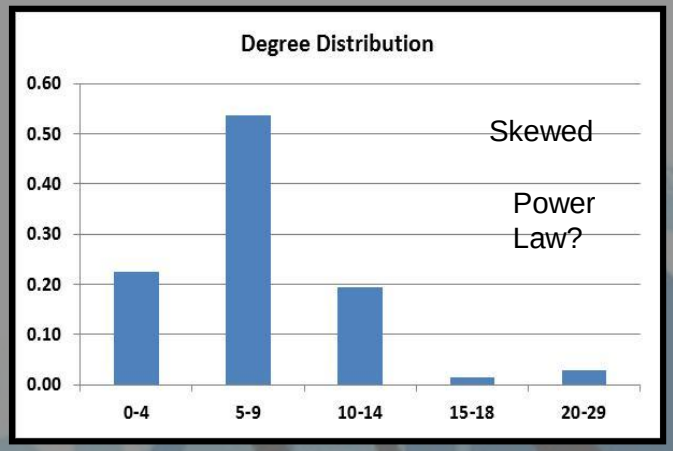
Label	Closeness
GL	3.71
CB	3.46
RY	2.83
RM	2.83
RB	2.83

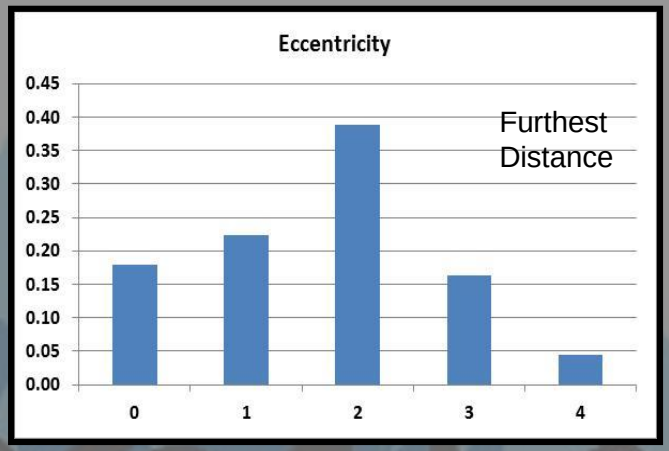
Label	Betweenness
SM	254.83
RN	175.54
RT	83.68
TJ	83.67
WF	70.93

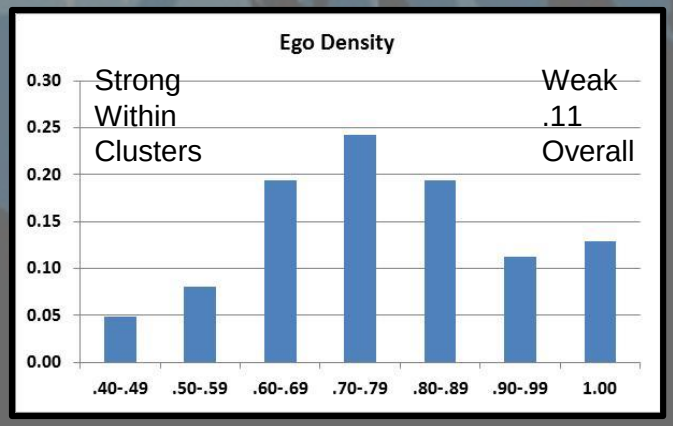
Label	Clustering
RK	1.00
CK	1.00
IC	1.00
MS	1.00
JK	1.00

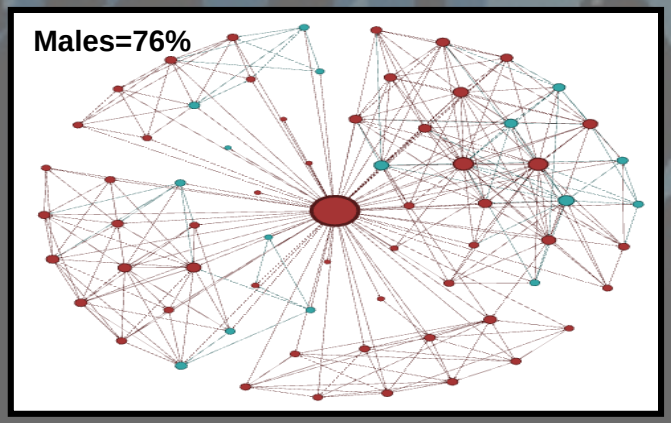
Label	Eigenvector
TJ	1.00
WF	0.94
SW	0.75
JM	0.70
CH	0.70

Social Network Analysis: Ego Analysis – Simple Example



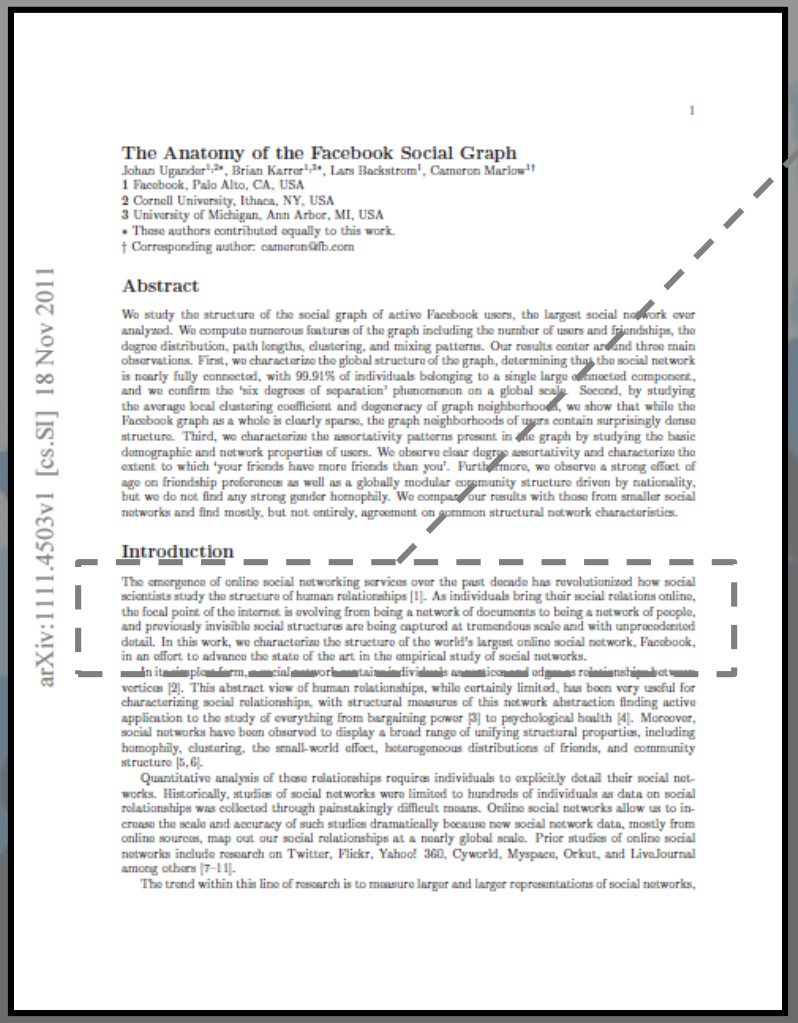






PL = N(N-1)/2 = 2211	Ego Density = L/PL = 235/2211 =.11

Social Network Analysis: Large Scale Networks



1

The Anatomy of the Facebook Social Graph

Johan Ugander[1,2*], Brian Karrer[1,3*], Lars Backstrom[1], Cameron Marlow[1†]
1 Facebook, Palo Alto, CA, USA
2 Cornell University, Ithaca, NY, USA
3 University of Michigan, Ann Arbor, MI, USA
* These authors contributed equally to this work.
† Corresponding author: cameron@fb.com

Abstract

We study the structure of the social graph of active Facebook users, the largest social network ever analyzed. We compute numerous features of the graph including the number of users and friendships, the degree distribution, path lengths, clustering, and mixing patterns. Our results center around three main observations. First, we characterize the global structure of the graph, determining that the social network is nearly fully connected, with 99.91% of individuals belonging to a single large connected component, and we confirm the 'six degrees of separation' phenomenon on a global scale. Second, by studying the average local clustering coefficient and degeneracy of graph neighborhoods, we show that while the Facebook graph as a whole is clearly sparse, the graph neighborhoods of users contain surprisingly dense structure. Third, we characterize the assortativity patterns present in the graph by studying the basic demographic and network properties of users. We observe clear degree assortativity and characterize the extent to which 'your friends have more friends than you'. Furthermore, we observe a strong effect of age on friendship preferences as well as a globally modular community structure driven by nationality, but we do not find any strong gender homophily. We compare our results with those from smaller social networks and find mostly, but not entirely, agreement on common structural network characteristics.

Introduction

The emergence of online social networking services over the past decade has revolutionized how social scientists study the structure of human relationships [1]. As individuals bring their social relations online, the focal point of the internet is evolving from being a network of documents to being a network of people, and previously invisible social structures are being captured at tremendous scale and with unprecedented detail. In this work, we characterize the structure of the world's largest online social network, Facebook, in an effort to advance the state of the art in the empirical study of social networks.

In its simplest form, a social network entails individuals and edges as relationships between vertices [2]. This abstract view of human relationships, while certainly limited, has been very useful for characterizing social relationships, with structural measures of this network abstraction finding active application to the study of everything from bargaining power [3] to psychological health [4]. Moreover, social networks have been observed to display a broad range of unifying structural properties, including homophily, clustering, the small-world effect, heterogeneous distributions of friends, and community structure [5, 6].

Quantitative analysis of these relationships requires individuals to explicitly detail their social networks. Historically, studies of social networks were limited to hundreds of individuals as data on social relationships was collected through painstakingly difficult means. Online social networks allow us to increase the scale and accuracy of such studies dramatically because new social network data, mostly from online sources, map out our social relationships at a nearly global scale. Prior studies of online social networks include research on Twitter, Flickr, Yahoo! 360, Cyworld, Myspace, Orkut, and LiveJournal among others [7–11].

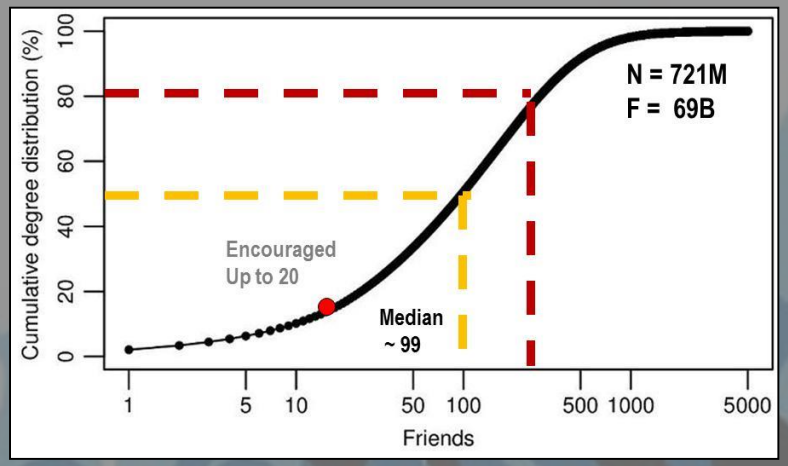
The trend within this line of research is to measure larger and larger representations of social networks,

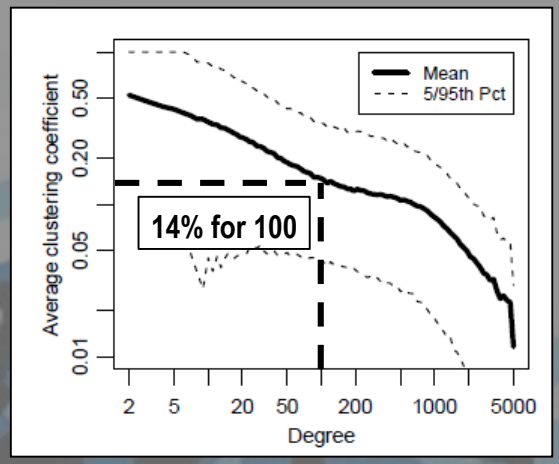
The emergence of online social networking services over the past decade has revolutionized how social scientists study the structure of human … previously invisible social structures are being captured at tremendous scale and with unprecedented detail.

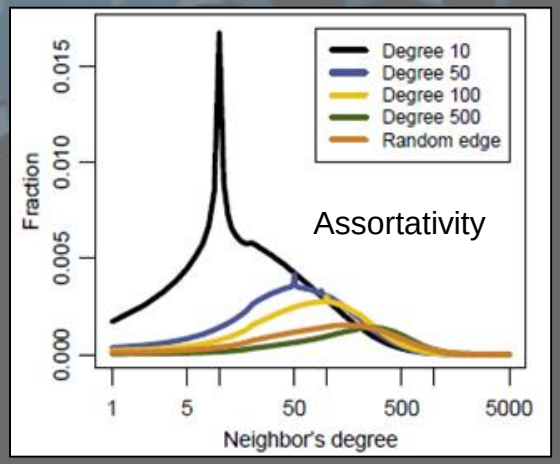
Active*	Global	US
Members	721M	149M
Friends	68.7B	15.9B
Aver. Friends	190	214
Total Pop	6.9B	260M

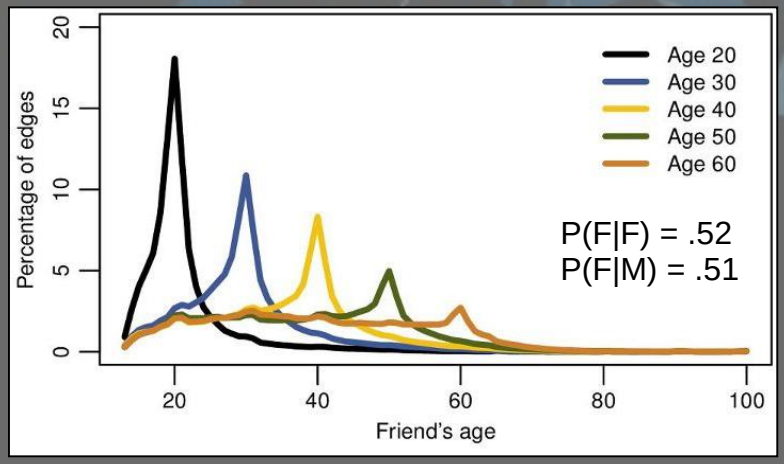
Accessed within 28 days of May '11
At least one friend
Over 13 years of age

Social Network Analysis: Large Scale Networks









Social Network Analysis: Is it a Small World after all?

arXiv:1111.4570v2 [cs.SI] 7 Dec 2011

Four Degrees of Separation

Lars Backstrom* Paolo Boldi† Marco Rosa† Johan Ugander* Sebastiano Vigna†

December 8, 2011

Abstract

Frigyes Karinthy, in his 1929 short story "Láncszemek" ("Chains") suggested that any two persons are distanced by at most six friendship links.[1] Stanley Milgram in his famous experiment [20, 23] challenged people to route postcards to a fixed recipient by passing them only through direct acquaintances. The average number of intermediaries on the path of the postcards lay between 4.4 and 5.7, depending on the sample of people chosen.

We report the results of the first world-scale social-network graph-distance computations, using the entire Facebook network of active users (≈ 721 million users, ≈ 69 billion friendship links). The average distance we observe is 4.74, corresponding to 3.74 intermediaries or "degrees of separation", showing that the world is even smaller than we expected, and prompting the title of this paper. More generally, we study the distance distribution of Facebook and of some interesting geographic subgraphs, looking also at their evolution over time.

The networks we are able to explore are almost two orders of magnitude larger than those analysed in the previous literature. We report detailed statistical metadata showing that our measurements (which rely on probabilistic algorithms) are very accurate.

1 Introduction

At the 20th World-Wide Web Conference, in Hyderabad, India, one of the authors (Sebastiano) presented a new tool for studying the distance distribution of very large graphs: HyperANF [3]. Building on previous graph compression [4] work and on the idea of diffusive computation pioneered in [21], the new tool made it possible to accurately study the distance distribution of graphs orders of magnitude larger than it was previously possible.

One of the goals in studying the distance distribution is the identification of interesting statistical parameters that can be used to tell proper social networks from other complex networks, such as web graphs. More generally, the distance distribution is one interesting *global* feature that makes it possible to reject probabilistic models even when they match local features such as the in-degree distribution.

In particular, earlier work had shown that the *spid*[2], which measures the *dispersion* of the distance distribution, appeared to be smaller than 1 (underdispersion) for social networks, but larger than one (overdispersion) for web graphs [3]. Hence, during the talk, one of the main open questions was "What is the spid of Facebook?".

Lars Backstrom happened to listen to the talk, and suggested a collaboration studying the Facebook graph. This was of course an extremely intriguing possibility: beside testing the "spid hypothesis", computing the distance distribution of the Facebook graph would have been the largest Milgram-like [20] experiment ever performed, orders of magnitudes larger than previous attempts (during our experiments Facebook has ≈ 721 million active users and ≈ 69 billion friendship links).

This paper reports our findings in studying the distance distribution of the largest electronic social network ever created. That world is smaller than we thought: the average distance of the current Facebook graph is 4.74. Moreover, the spid of the graph is just 0.08, corroborating the conjecture [3] that proper social networks have a spid well below one. We also observe, contrary to previous literature analysing graphs orders of magnitude smaller, both a stabilisation of the average distance over time, and that the density of the Facebook graph over time does not neatly fit previous models.

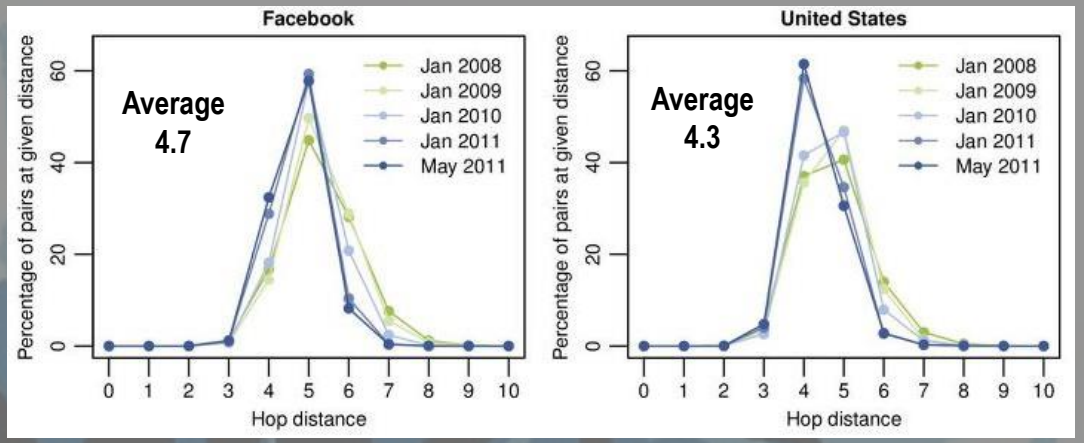
Towards a deeper understanding of the structure of the Facebook graph, we also apply recent compression techniques

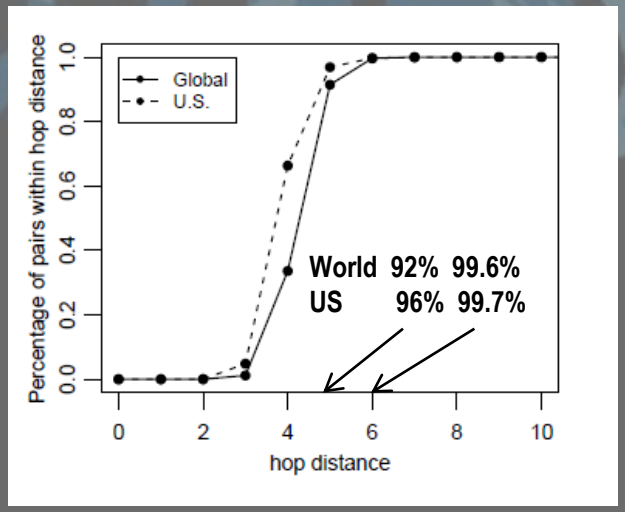
*Facebook.

†DSI, Università degli Studi di Milano, Italy. Paolo Boldi, Marco Rosa and Sebastiano Vigna have been partially supported by a Yahoo! faculty grant and by MIUR PRIN "Query log e web crawling".

[1]The exact wording of the story is slightly ambiguous: "He bet us that, using no more than five individuals, one of whom is a personal acquaintance, he could contact the selected individual [...]". It is not completely clear whether the selected individual is part of the five, so this could actually allude to distance five or six in the language of graph theory, but the "six degrees of separation" phrase stuck after John Guare's 1990 eponymous play. Following Milgram's definition and Guare's interpretation (see further on), we will assume that "degrees of separation" is the same as "distance minus one", where "distance" is the usual path length (the number of arcs in the path).

[2]The spid (shortest-paths index of dispersion) is the variance-to-mean ratio of the distance distribution.





Social Network Analysis: Second Example

The Political Blogosphere and the 2004 U.S. Election: Divided They Blog

Lada Adamic
HP Labs
1501 Page Mill Road
Palo Alto, CA 94304
lada.adamic@hp.com

Natalie Glance
Intelliseek Applied Research Center
5001 Baum Blvd.
Pittsburgh, PA 15217
nglance@intelliseek.com

4 March 2005

Abstract

In this paper, we study the linking patterns and discussion topics of political bloggers. Our aim is to measure the degree of interaction between liberal and conservative blogs, and to uncover any differences in the structure of the two communities. Specifically, we analyze the posts of 40 "A-list" blogs over the period of two months preceding the U.S. Presidential Election of 2004, to study how often they referred to one another and to quantify the overlap in the topics they discussed, both within the liberal and conservative communities, and also across communities. We also study a single day snapshot of over 1,000 political blogs. This snapshot captures blogrolls (the list of links to other blogs frequently found in sidebars), and presents a more static picture of a broader blogosphere. Most significantly, we find differences in the behavior of liberal and conservative blogs, with conservative blogs linking to each other more frequently and in a denser pattern.

1 Introduction

The 2004 U.S. Presidential Election was the first Presidential Election in the United States in which blogging played an important role. Although the term weblog was coined in 1997, it was not until after 9/11 that blogs gained readership and influence in the U.S. The next major trend in political blogging was "warblogging": blogs centered around discussion of the invasion of Iraq by the U.S.[1]

The year 2004 saw a rapid rise in the popularity and proliferation of blogs. According to a report from the Pew Internet & American Life Project published in January 2005, 32 million U.S. citizens now read weblogs. However, 62% of online Americans still do not know what a weblog is.[2] Another report from the same project showed that Americans are turning to the Internet in increasing numbers to stay informed about politics: 63 million in mid-2004 vs. 30 million in March 2000.[3]

A significant fraction of that traffic was directed specifically to blogs, with 9% of Internet users saying they read political blogs "frequently" or "sometimes" during the campaign.[4] Indeed, political blogs showed a large growth in traffic in the months preceding the election.[5]

Recognizing the importance of blogs, several candidates and political parties set up weblogs during the 2004 U.S. Presidential campaign. Notably, Howard Dean's campaign was particularly

[1] http://en.wikipedia.org/wiki/Weblog
[2] http://www.pewinternet.org/PPF/r/144/report_display.asp
[3] http://www.pewinternet.org/PPF/r/141/report_display.asp
[4] http://www.pewinternet.org/pdfs/PIP_blogging_data.pdf
[5] http://techcentralstation.com/011058B.html

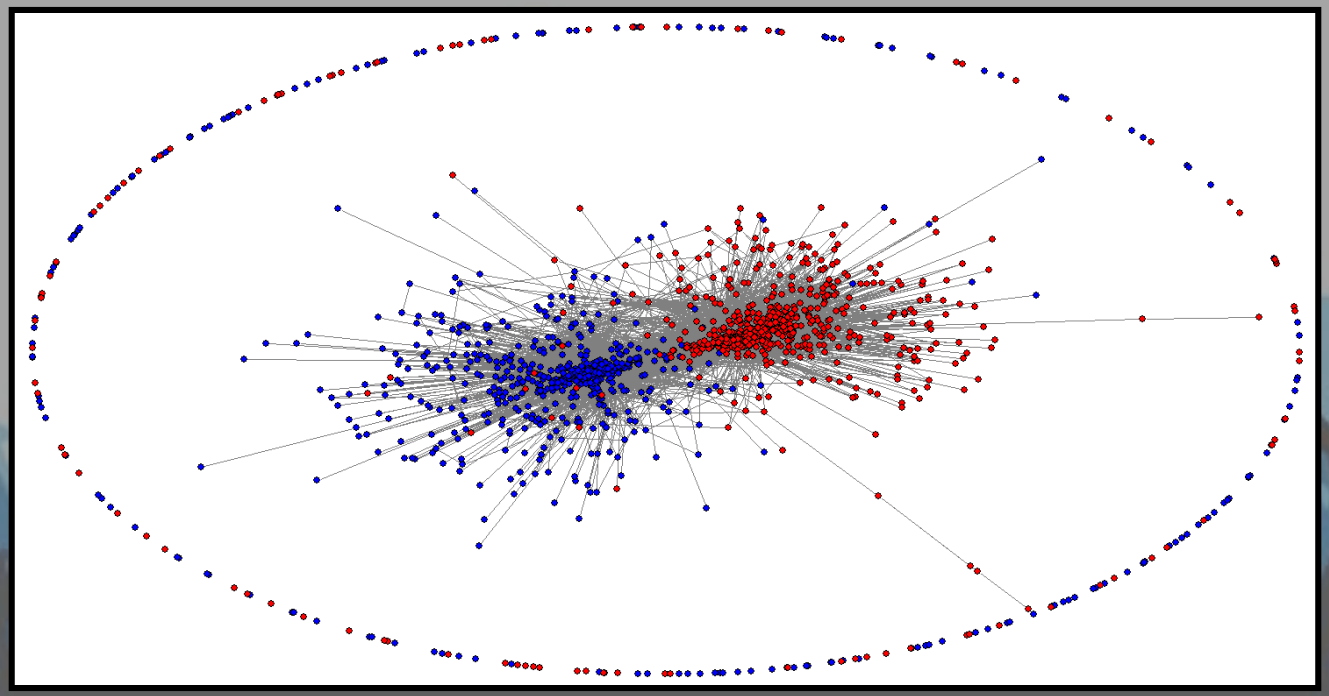
1

- Single day snapshot of a Snowball Sample of Political Blogs (N=1490)
- Manually assigned as Liberal or Conservative
- Focus on Blogrolls and front page citations
- Primary question: Cyber-balkanization?

Social Network Analysis: Balkanization (regular kind)

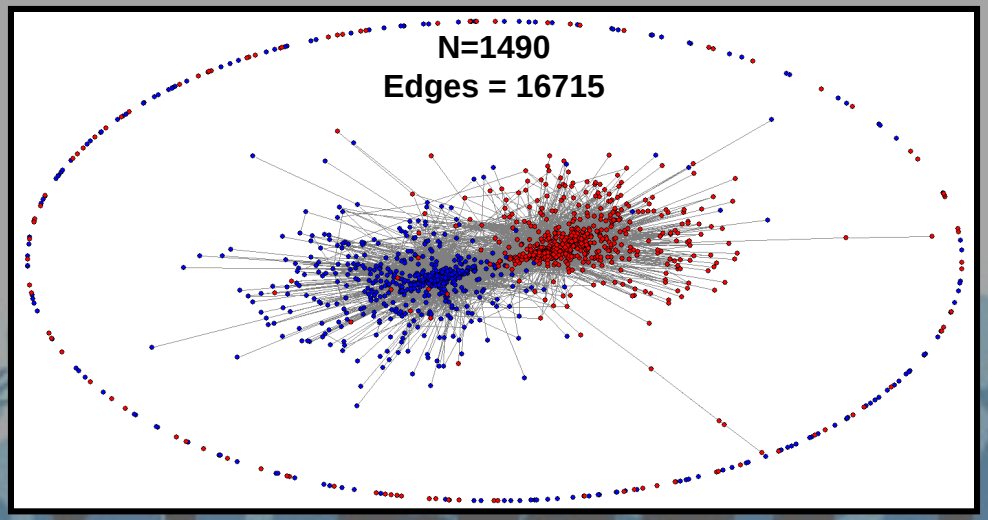
Process of fragmentation or division of a region or state into smaller regions or states that are often hostile or non-cooperative with each other.

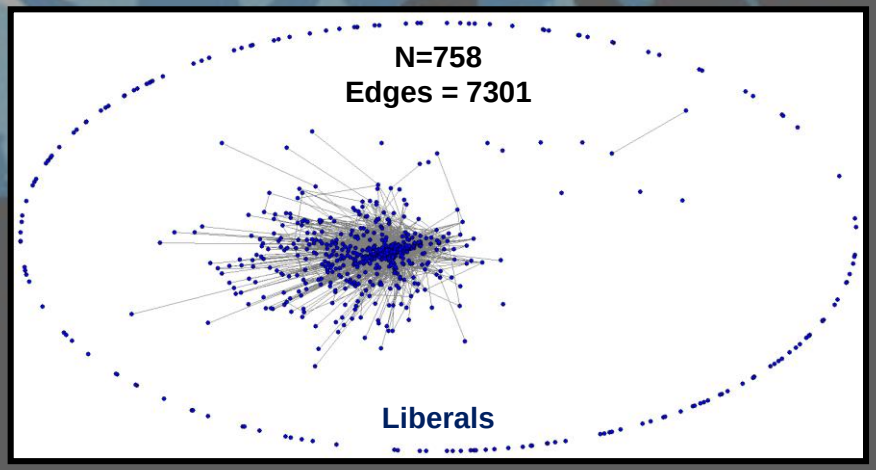
Social Network Analysis: Cyber-balkanization?

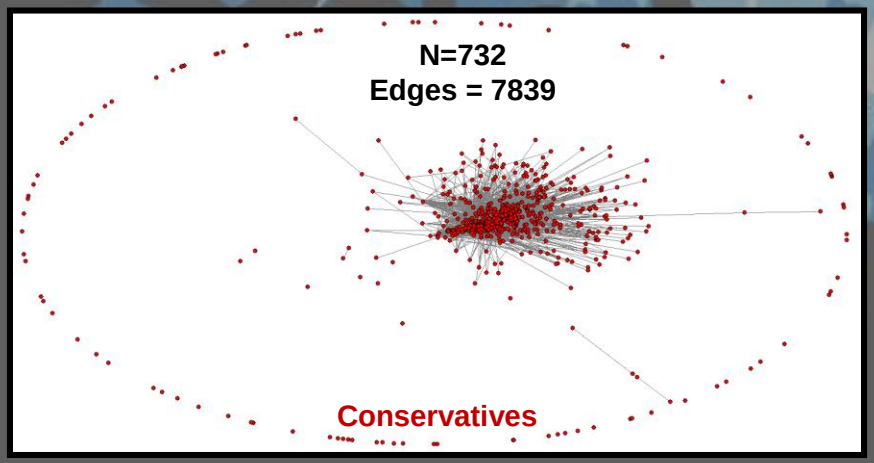


Proliferation of specialized online news sources allows people with different political leanings to be exposed only to information in agreement with their previously held views.

Social Network Analysis: Cyber-balkanization?







Social Network Analysis: Political Blogs – Cyberbalkanization?

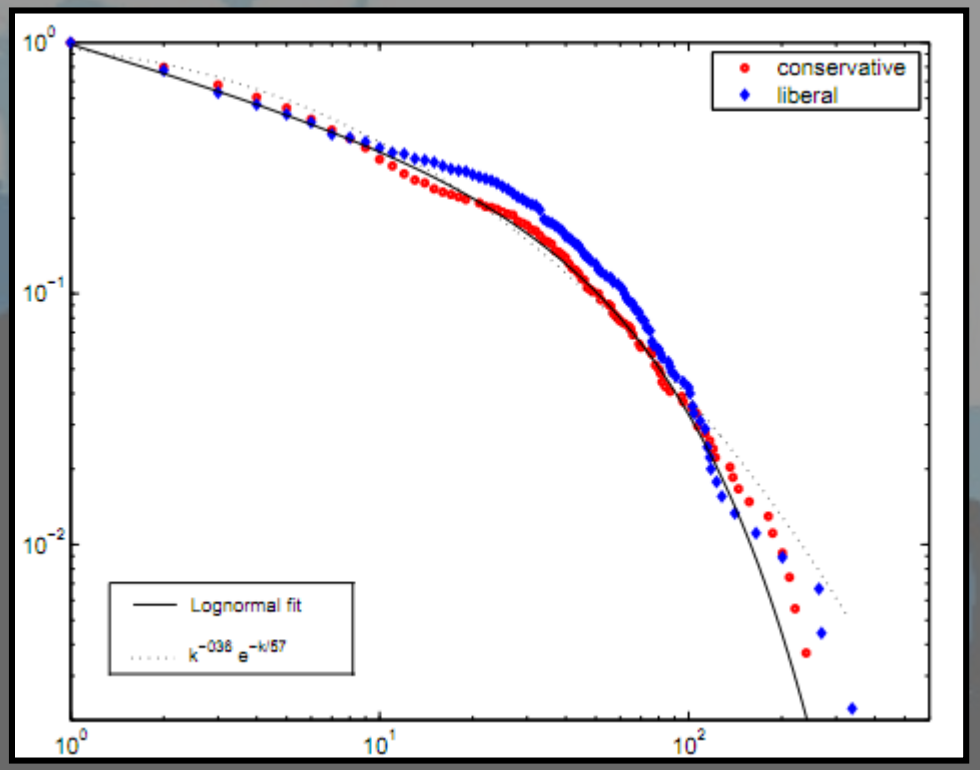
Viewpoint	Lib In Links	Cons In Links	Total In Links	%Lib	%Cons
dailykos.com	292	46	338	86%	14%
www.talkingpointsmemo.com	242	22	264	92%	8%
atrios.blogspot.com	230	39	269	86%	14%
www.washingtonmonthly.com	165	36	201	82%	18%
www.wonkette.com	83	30	113	73%	27%
www.juancole.com	149	16	165	90%	10%
yglesias.typepad.com/matthew	104	24	128	81%	19%
www.crookedtimber.org	81	19	100	81%	19%
www.mydd.com	107	8	115	93%	7%
www.oliverwillis.com	97	20	117	83%	17%
blog.johnkerry.com	21	2	23	91%	9%
www.pandagon.net	118	5	123	96%	4%
www.talkleft.com	126	15	141	89%	11%
digbysblog.blogspot.com	115	3	118	97%	3%
www.politicalwire.com	87	16	103	84%	16%
www.j-bradford-delong.net	98	11	109	90%	10%
www.prospect.org/weblog	102	11	113	90%	10%
americablog.blogspot.com	64	5	69	93%	7%
www.theleftcoaster.com	78	4	82	95%	5%
www.jameswolcott.com	74	6	80	93%	8%
Total Liberal	2433	338	2771	88%	12%

www.powerlineblog.com	26	195	221	12%	88%
instapundit.com	43	234	277	16%	84%
www.littlegreenfootballs.com/weblog	10	171	181	6%	94%
www.hughhewitt.com	11	146	157	7%	93%
www.andrewsullivan.com/index.php	59	86	145	41%	59%
www.captainsquartersblog.com/mt	5	117	122	4%	96%
www.wizbangblog.com	14	125	139	10%	90%
www.indcjournal.com	6	60	66	9%	91%
www.michellemalkin.com	10	191	201	5%	95%
blogsforbush.com	4	208	212	2%	98%
www.allahpundit.com	2	37	39	5%	95%
belmontclub.blogspot.com	3	93	96	3%	97%
realclearpolitics.com	13	104	117	11%	89%
volokh.com	27	80	107	25%	75%
timblair.spleenville.com	7	80	87	8%	92%
windsofchange.net	16	65	81	20%	80%
www.vodkapundit.com	9	97	106	8%	92%
www.rogerlsimon.com	6	74	80	8%	93%
www.deanesmay.com	8	79	87	9%	91%
mypetjawa.mu.nu	0	51	51	0%	100%
Total Conservative	279	2293	2572	11%	89%

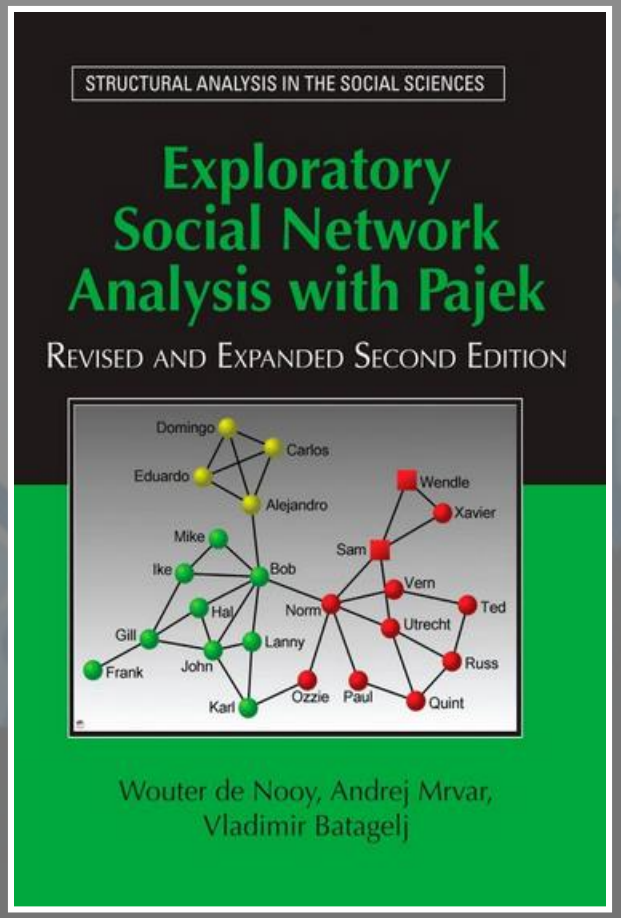
Measure	Liberal	Conservative
N	758	732
Out Links	74%	84%
In Links	67%	82%

Social Networks: Political Blogs - Metrics

Measure	Liberal	Conservative	Total
Density	0.02	0.03	0.01
No Components	188	107	268
Largest Comp	569	569	1222
Largest Comp%	75.10	84.97	82.01
Min Deg	0	0	0
Max Deg	305	296	351
Aver Deg	19.26	21.42	22.44
Deg Central	0.38	0.38	0.22
Diameter	6	7	8
Aver Dist	2.51	2.51	2.74
Betw Cent	0.10	0.16	0.06
EigVect Cent	0.23	0.26	0.22
Clust Coeff	0.31	0.20	0.22



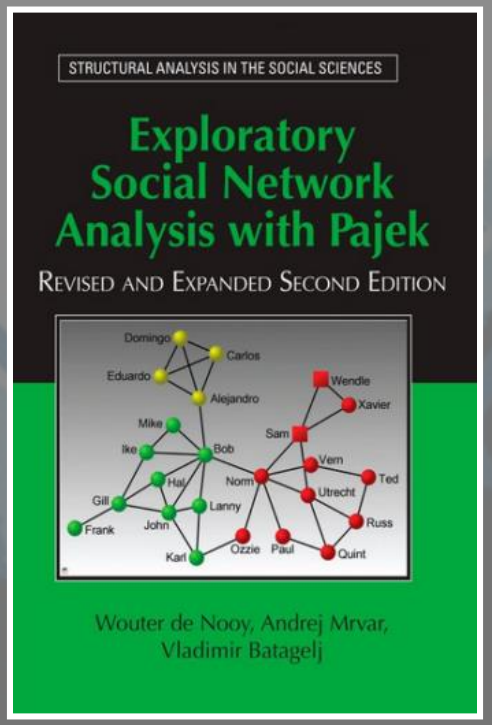
Social Network Analysis: Statistical Network Models



A Statistical Network Model

- Assumes that part of the structure of an observed network is random
- Mathematical description of a collection of possible networks and a probability distribution on this set
- Informs us about which network characteristics to expect if lines assigned to pairs of vertices at random

Social Network Analysis: Statistical Network Models



- Classic Bernouilli
- Conditional Uniform
- Small-World
- Preferential Attachment

Social Network Analysis: Political Blogs - Comparisons

Measure	Political Blogs	Bernoulli		Deg Conditional		Small World		Pref Attachment	
		2.5%	97.5%	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%
Number of Components	268	1	1	1	1	1	1	96	134
Fraction of Nodes in Largest Component	0.82	100	100	100	100	100	100	0.91	0.94
Diameter of the Largest Component	8	4	4	7	9	4	5	7	9
Average Path Length	2.74	2.61	2.63	3.29	3.36	2.98	3.01	3.07	3.14
Overall Clustering	0.226	0.017	0.018	0.029	0.031	0.355	0.372	0.095	0.109
Betweenness	0.065	0.002	0.003	0.010	0.021	0.003	0.004	0.038	0.064