Mining and Analyzing Social Media: Part 1

Dave King

HICSS 47 - January 2014

RedPrairie i2 manugistics E3 intactix Arthur

Agenda: Part 1

- Introduction
  - Bio
  - Some definitions
  - Growing Interest
  - Resources
- Social Homophily as an Example
  - Meaning and Import
  - Political Cyberbalkanization
    - Social Network Analysis
    - Text Mining

Agenda: Part 2

- Introduction to Social Network Analysis Metrics
  - Degrees of Separation
  - Hooray for Bollywood
- Standard Measures
  - Centrality
  - Cohesion
- Levels of Social Network Analysis
  - Facebook & Egocentric Analysis
  - Searching for Cohesive Subgroups

Dave King

Biography

University Professor

Dir R&D Execucom

SVP & CTO Comshare

EVP R&D JDA Software

‘70 ‘80 ‘90 ‘00 ‘10

App Stats
Math Modeling

DSS
Optimization
App Stats

Nat Lang for DB Query
DSS Interface
HICSS (83)

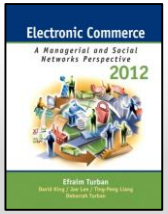
DSS
Multi-Dim DB
EIS & BI

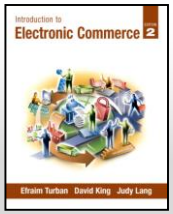
ECom
HICSS
Track Intern
& Dig Econ

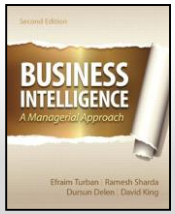
Merch Planning
SCP Planning
Optimization

HICSS
Tutorial
Web Mining &
Analysis

Sentiment Analysis



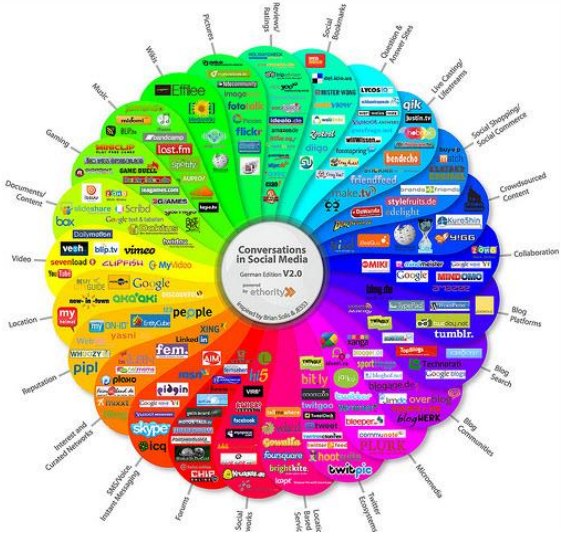


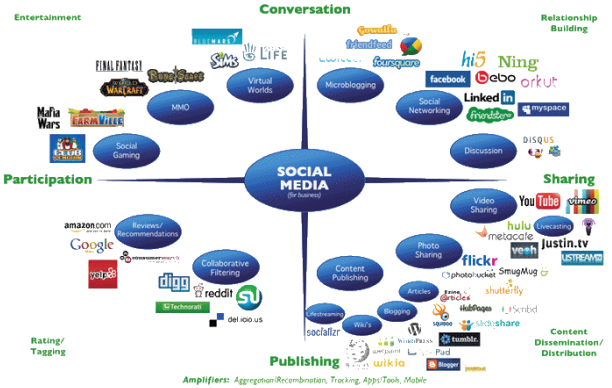


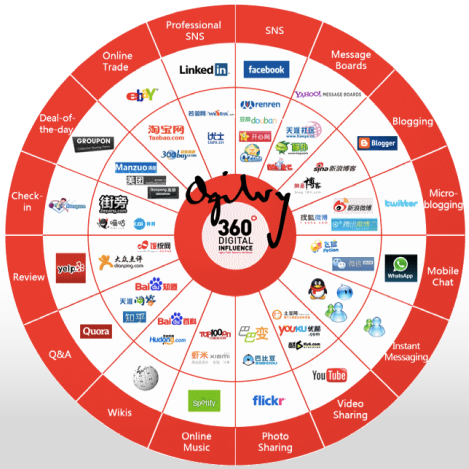
Social Media Defined

Social Media

View from the Pinwheel

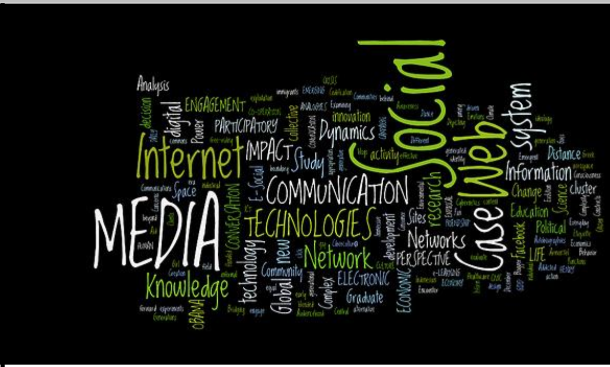


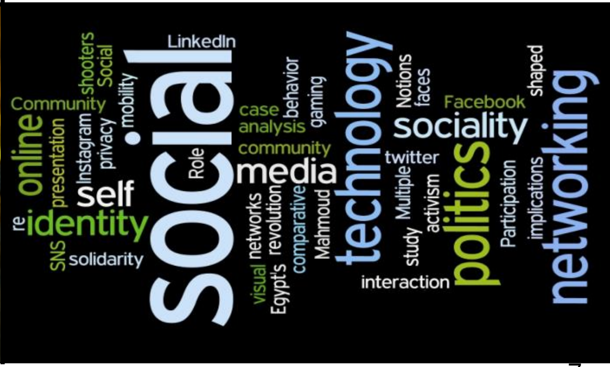




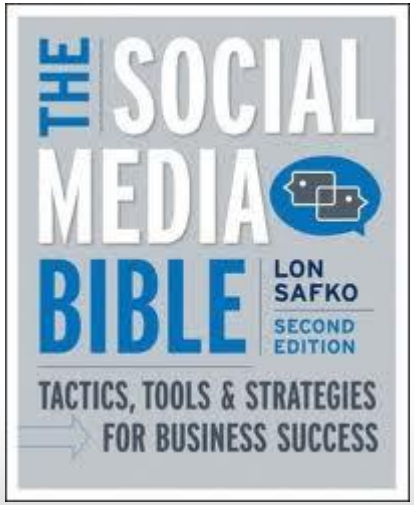
Social Media

View from the Word Clouds





Defined



is the media we use to be social. That's it.

Mining and Analyzing Social Media

Two Approaches

Content

Connections

Social Media

Content

Emails
Tweets
SMS
Tags
IM Chat
Bookmarks
Photos
FAQs
Video
Discussion
Music
Presentations
Conversations
Ratings
Wiki Entries
Reviews
Lifestreams
Articles
News Stories
Recommendations
Location
Profiles
RSS Feeds
Blogs

Social Media

Social Connections

Family Friends Groups

Followers

Contacts

Authors

Readers

Senders

Recipients

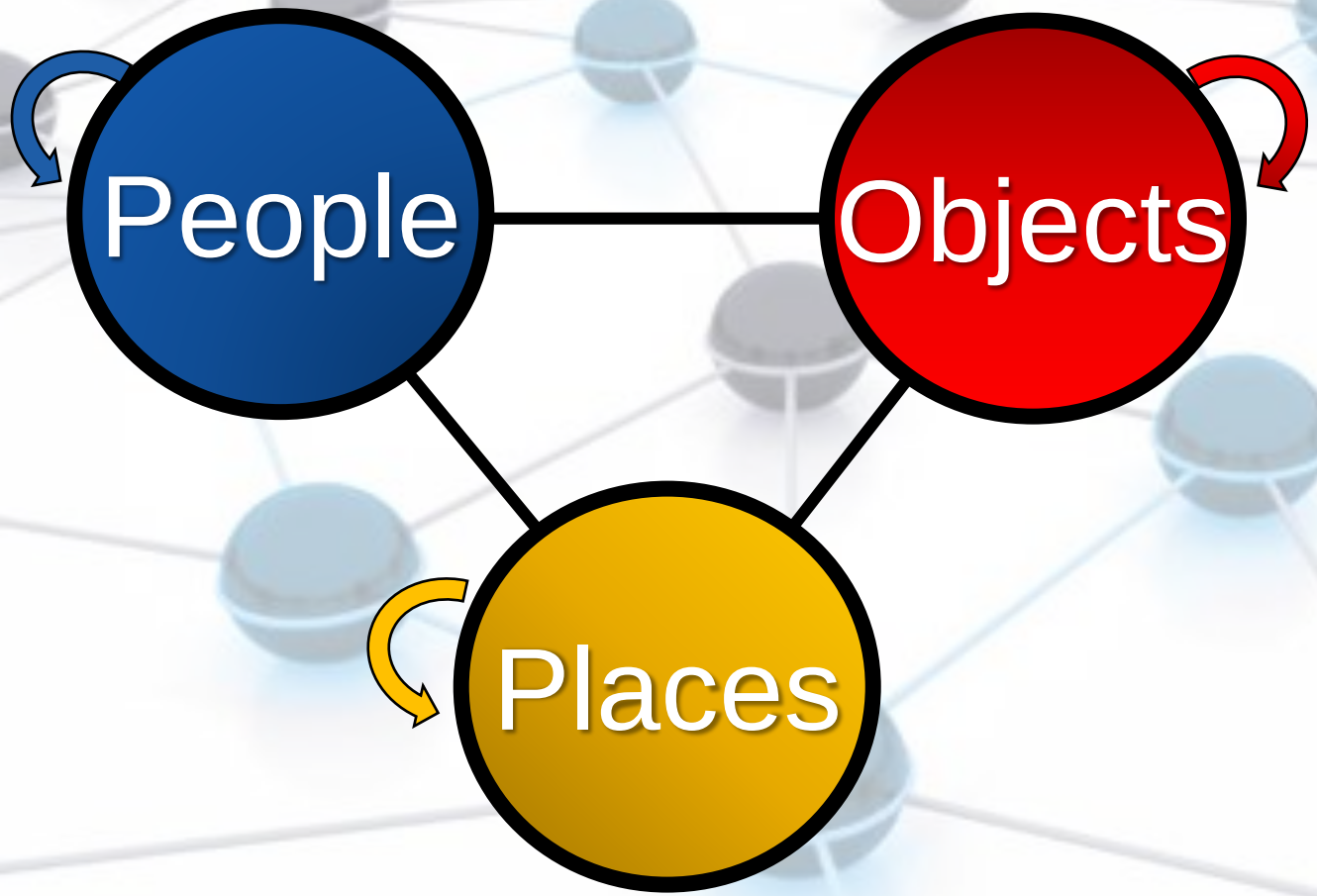
Participants

Answerers

Collaborators Subscribers Collaborators

Social Media

Connections



Types of Mining and Analysis

Mining and Analyzing Social Media

The Content

Data Mining
Discovering meaningful patterns from large data sets using pattern recognition technologies

Web Mining
Data Mining focused on the analysis of Web Usage, Structure & Content

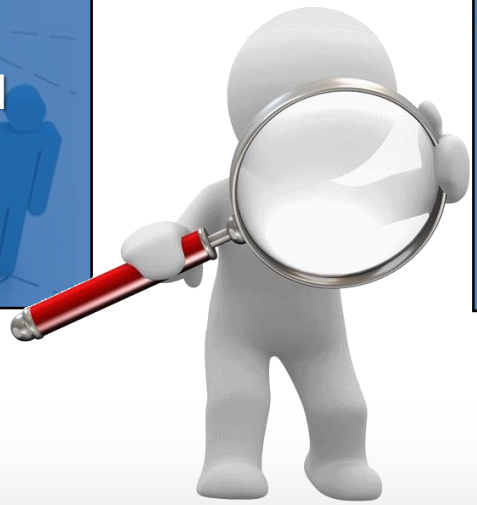
Text Mining
Using natural language processing & data mining to discover patterns in a collection of "documents

Mining and Analyzing Social Media

The Connections

Social Network Analysis [SNA]

The mapping and measuring of relationships and flows between people, groups, organizations, computers, URLs, and other connected information/knowledge entities.



Network science

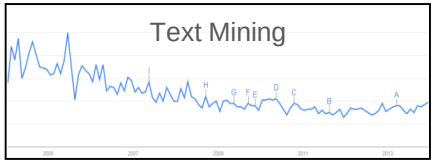
Study of the theoretical foundations of network structure and behavior and the application of networks to many subfields including SNA, collaboration networks, emergent systems, and physical and life science systems

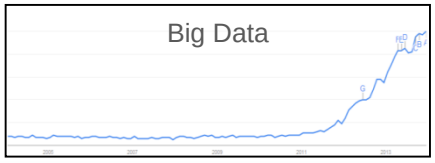
Interest

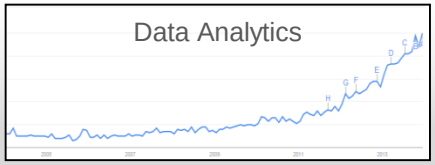
You are not alone or maybe you are?

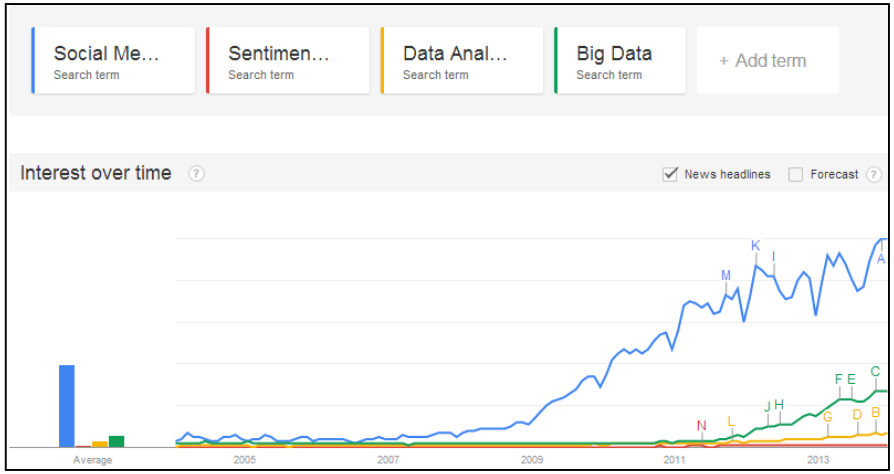
Interest

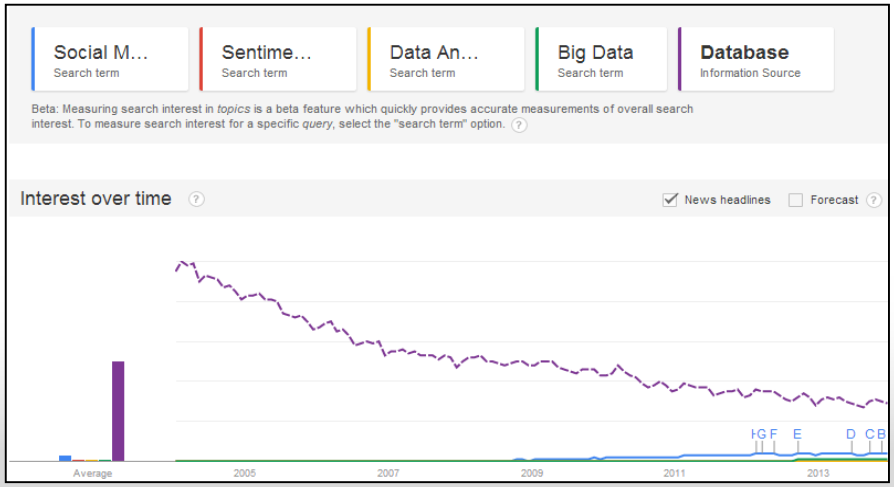
Search - Google Trends

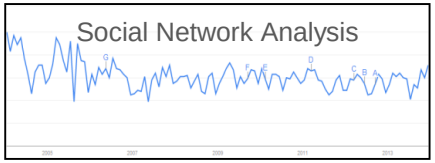


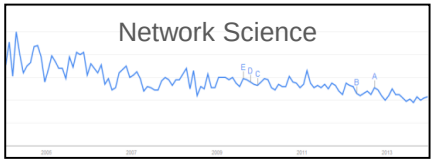


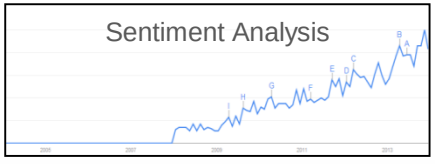


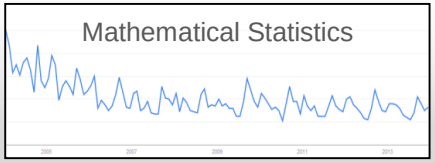






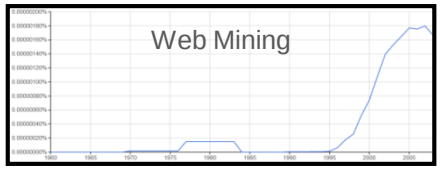


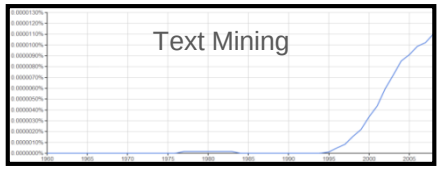


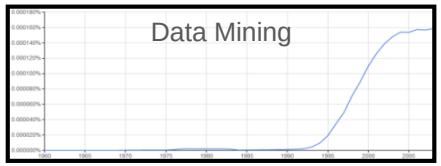


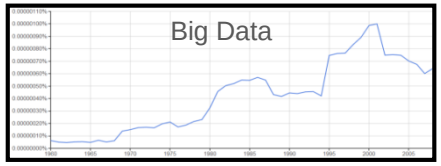
Interest

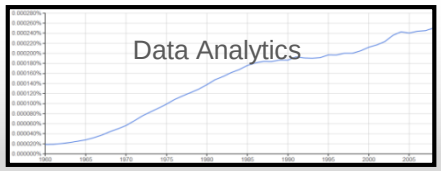
Book Titles – Google Ngram Viewer

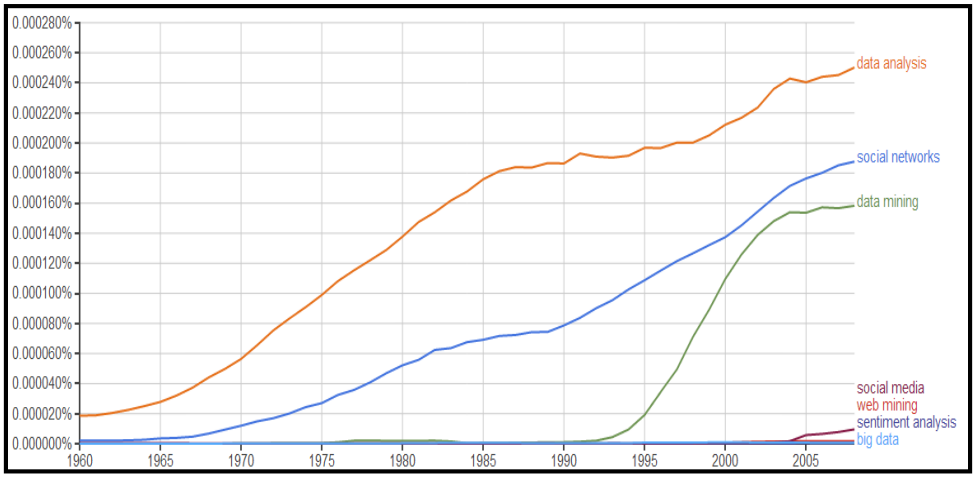


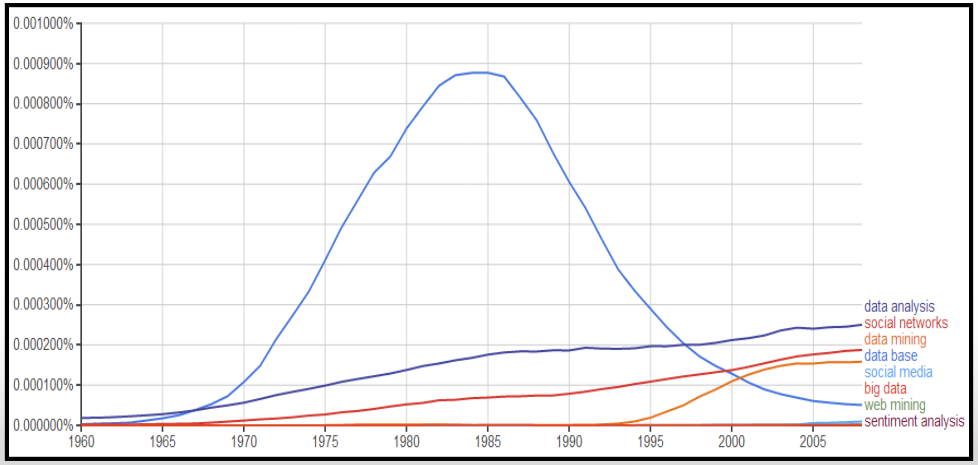


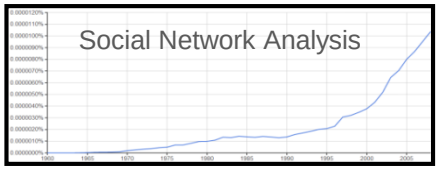


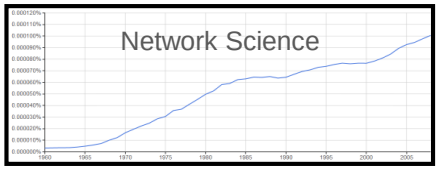


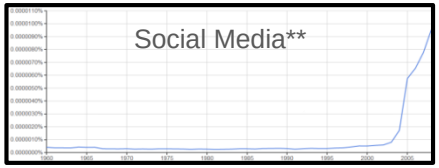


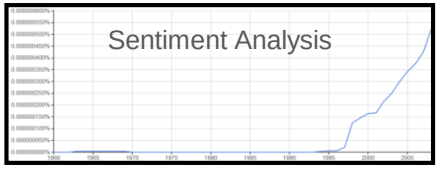


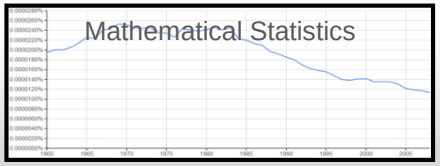










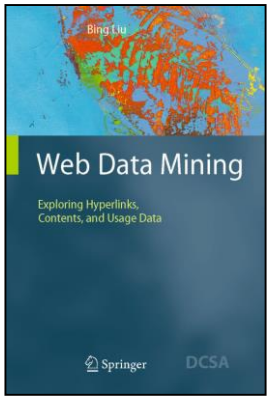


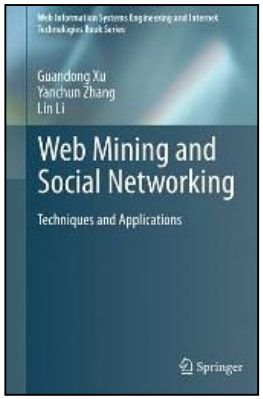
Resources

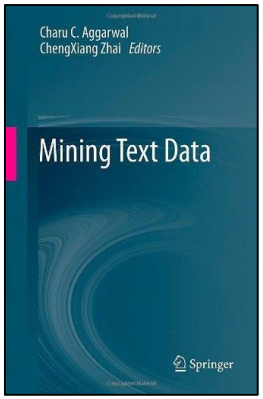
Customers who spent a lot of money on these items also…

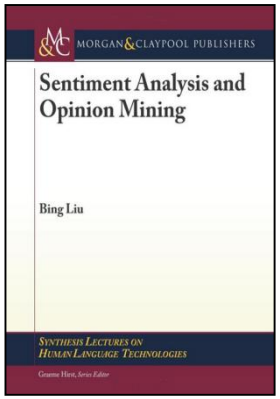
Web & Text Mining Resources

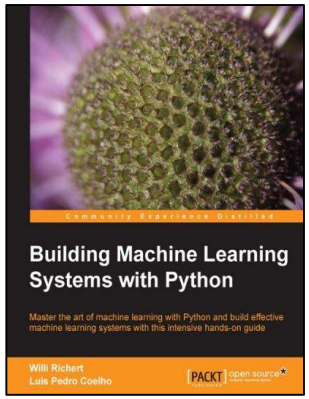
General Coverage and Programming

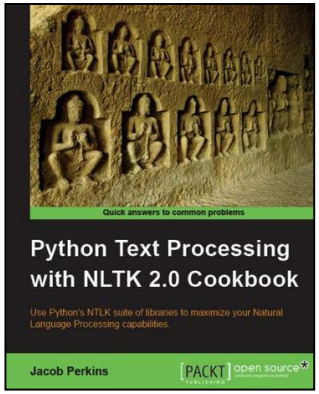


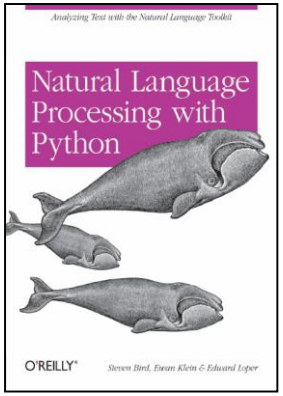


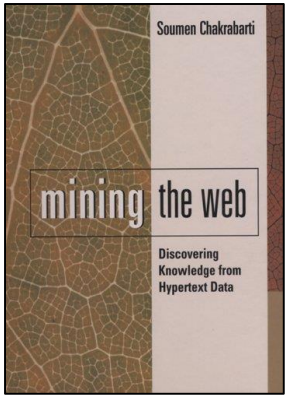


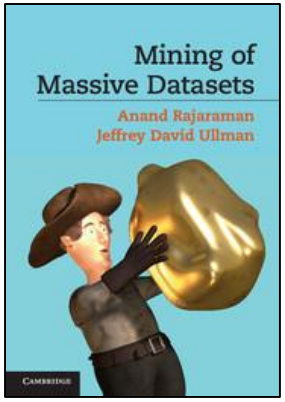


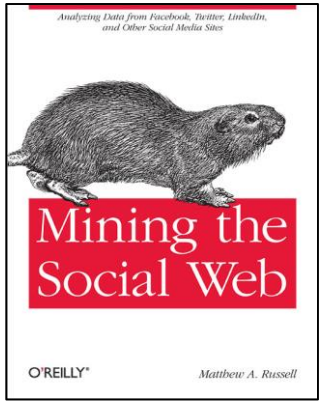






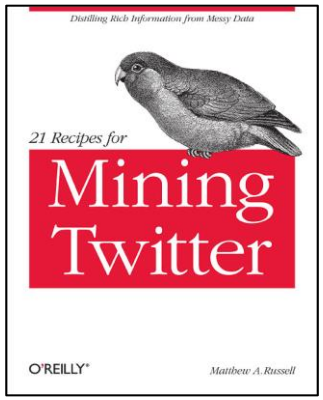


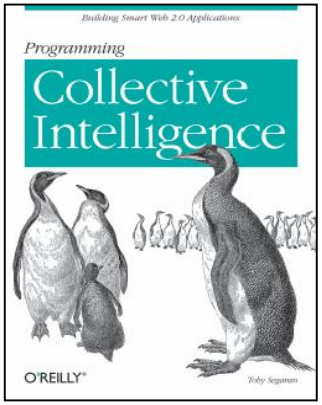


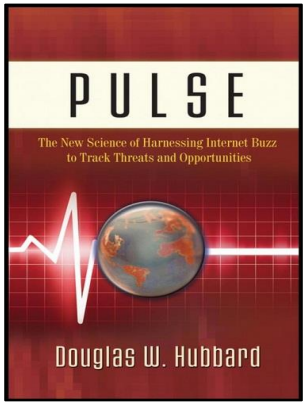


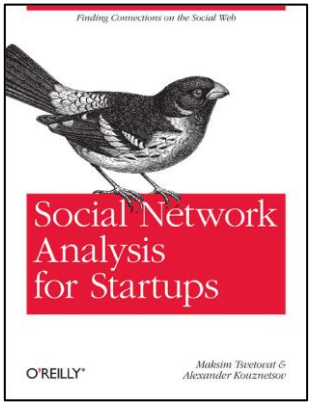
Text Mining & Social Network Analysis

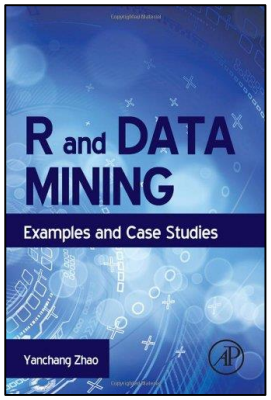
Computational Models and Programming

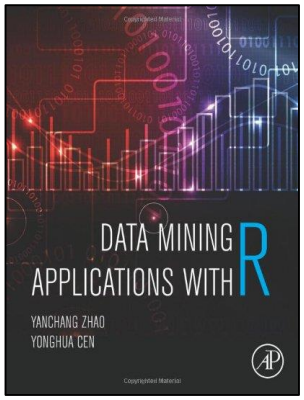


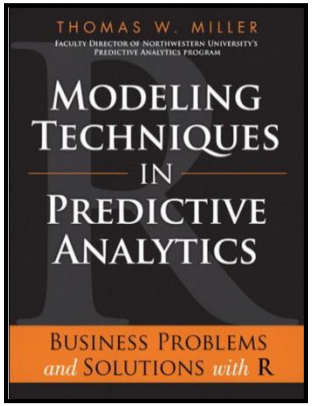


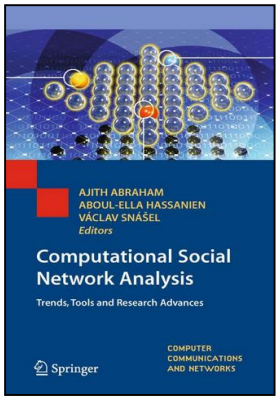


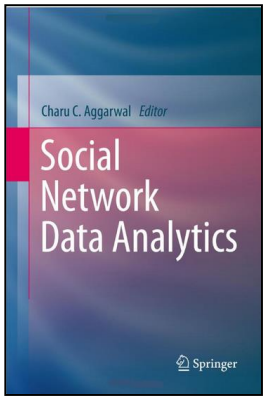






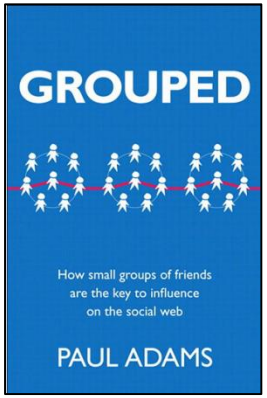


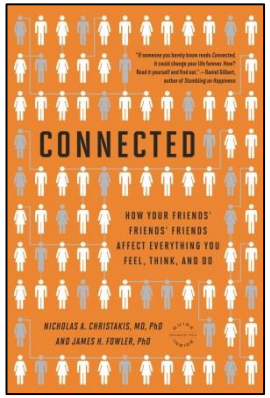


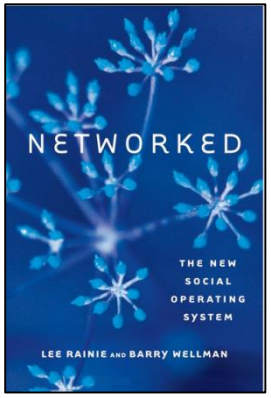


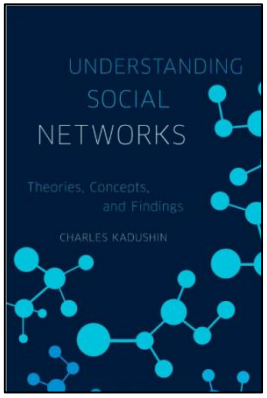
Social Network Analysis

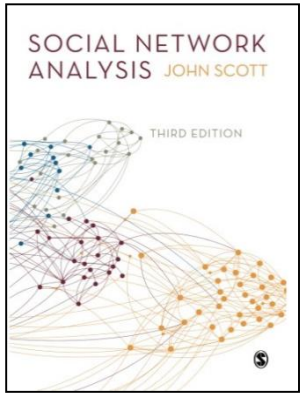
General Coverage of SNA

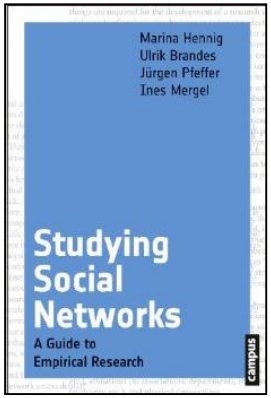


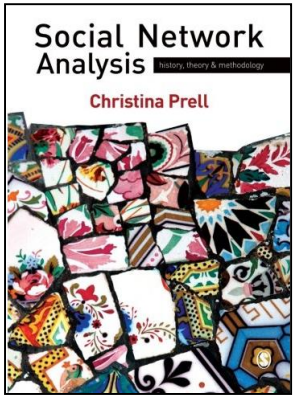


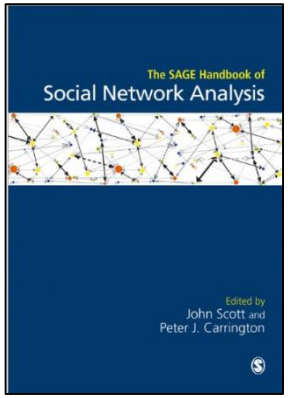


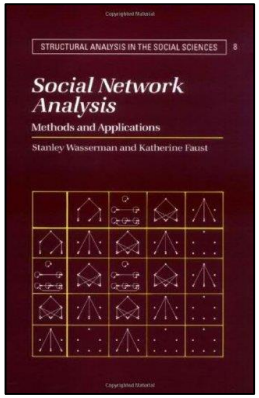


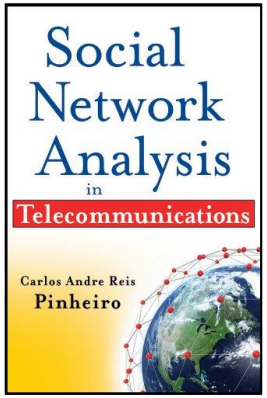


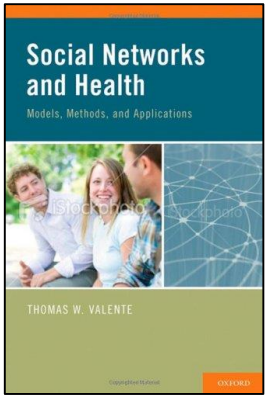


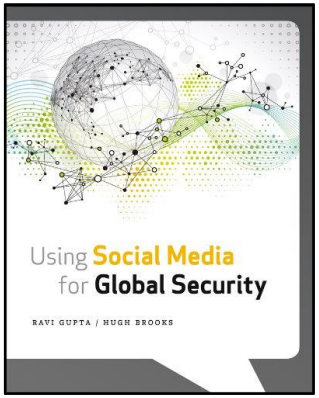






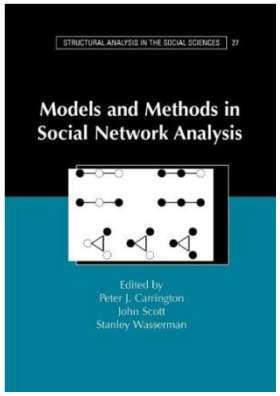


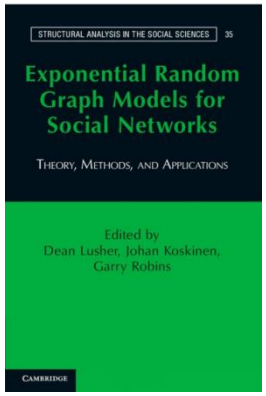


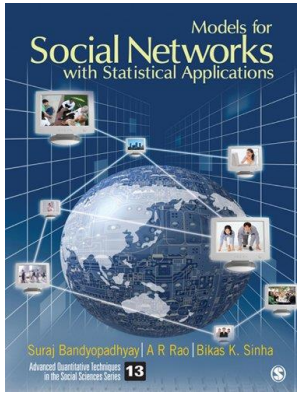


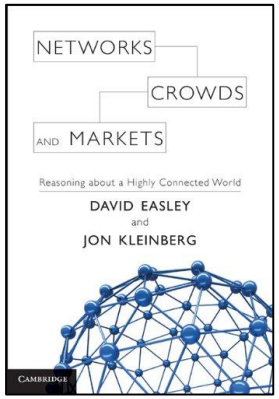
Social Network Analysis

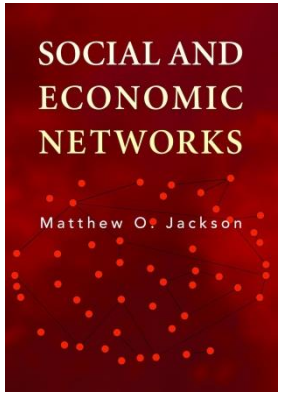
Mathematical Models and SNA Tools

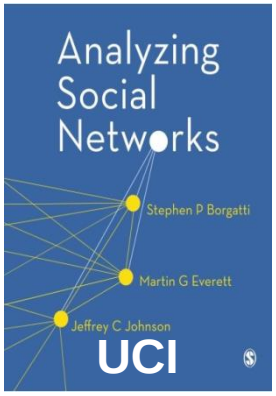


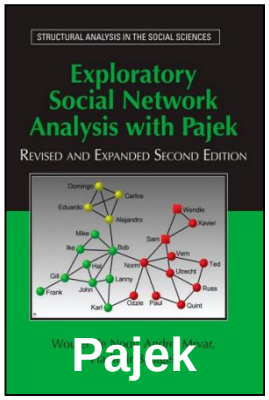


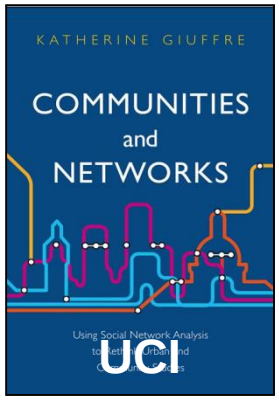


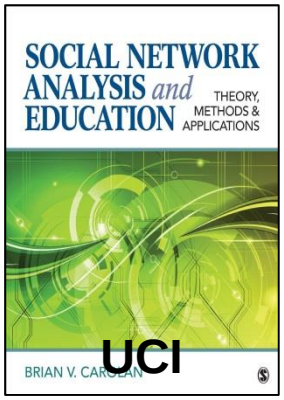


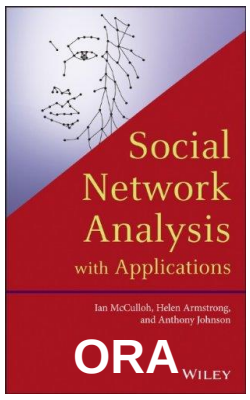


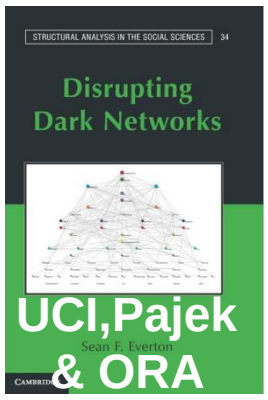


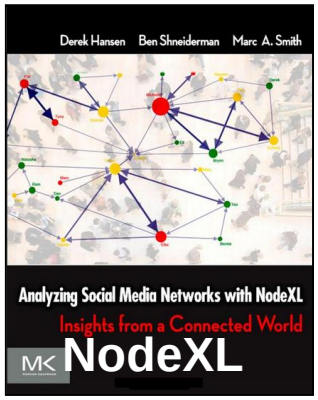






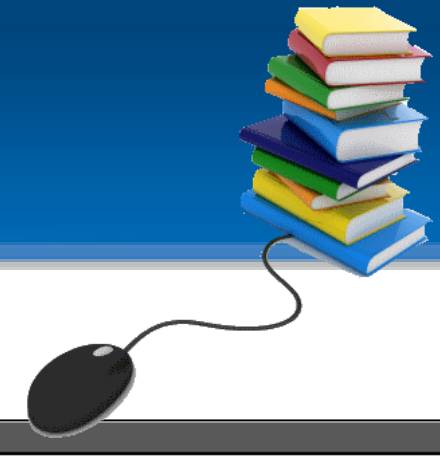






Social Network Analysis

Online Tutorials and Classes



Author	Date	Type	SubType	Software	Link
Adamic, L.	2013	Class	Video and slides	Gephi, NetLogo	https://www.coursera.org/course/sna
Agrawal, D. et al.	2011	Tutorial	Slides		https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=10&ved=0CH0QFjAJ&url=http%3A%2F%2Fwww.cs.ucsb.edu%2F~cbudak%2Fvldb_tutorial.pdf&ei=RqiPUpzyJYjwoATLl4CAAw&usg=AFQjCNGFhrHAIBRCPcFuZgT67fbfhd-Iww&bvm=bv.56988011,d.cGU
Clauset, A.	2013	Class	Slides	Gephi	http://tuvalu.santafe.edu/~aaronc/courses/5352/
Cook, J.	2013	Class	Video		http://www.umasocialmedia.com/socialnetworks/syllabus-spring13/
Hanneman R. and Mark Riddle	2005	Tutorial	Online Book	UCINET, NetDraw	http://faculty.ucr.edu/~hanneman/
Hearst, M. and Gilad Mishne	2012	Tutorial	Video		http://www.youtube.com/playlist?list=PLE8C1256A28C1487F
Katz,D. and Michael Bommarito	2010	Tutorial	Slides		http://computationallegalstudies.com/network-analysis-and-law-tutorial/
Leskovec, J.	2012	Tutorial	Slides		https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=5&ved=0CE0QFjAE&url=http%3A%2F%2Fkdd2012.sigkdd.org%2Fsites%2Fimages%2Fsummerschool%2FJure-Leskovec-part1.pdf&ei=DzqiUpunEIf3oAS804KgBQ&usg=AFQjCNGBK9MNkgMDn2KTeadjvybC7LOGSg&bvm=bv.57752919,d.cGU
Leskovec, J.	2011	Tutorial	Slides & Video		http://snap.stanford.edu/proj/socmedia-kdd/
Leskovec, J.	2011	Tutorial	Slides & Video		http://videolectures.net/single_leskovec_social/
Leskovec, J.	2009	Tutorial	Slides & Video		http://videolectures.net/icml09_leskovec_msain/
McFarland, D. et al.	2010	Tutorial	Lab Exercises	R, SoNIA	http://sna.stanford.edu/rlabs.php
Sharma, P.	2010	Tutorial	Article	Gephi, R	http://www.r-bloggers.com/social-network-analysis-using-r-and-gephis/
Srivastava, J.	2008	Tutorial	Slides		http://www.siam.org/meetings/sdm08/tutorials.php

An Example

Social Homophily from Two Perspectives

Social Homophily

aka Assortative Mixing

Sound bites

- **Love of the same**
- **Birds of a feather flock together**
- **Likes associate with likes**
- **Likes copy likes**
- **Likes think alike**

Social Homophily

Some things that are homogenous

Type	Sub Type	Examples
Status	Socio-Demographic	Age, Education, Occupation, Income, Race/Ethnicity, Gender, Religion, Location…
	Behavioral	Interactions, Consumption, Exchange, Adoption, Deviance, Mating/Marriage…
Value	Intrapersonal	Beliefs, Attitudes, Values, Aspirations, Ideas/Knowledge, Emotions…

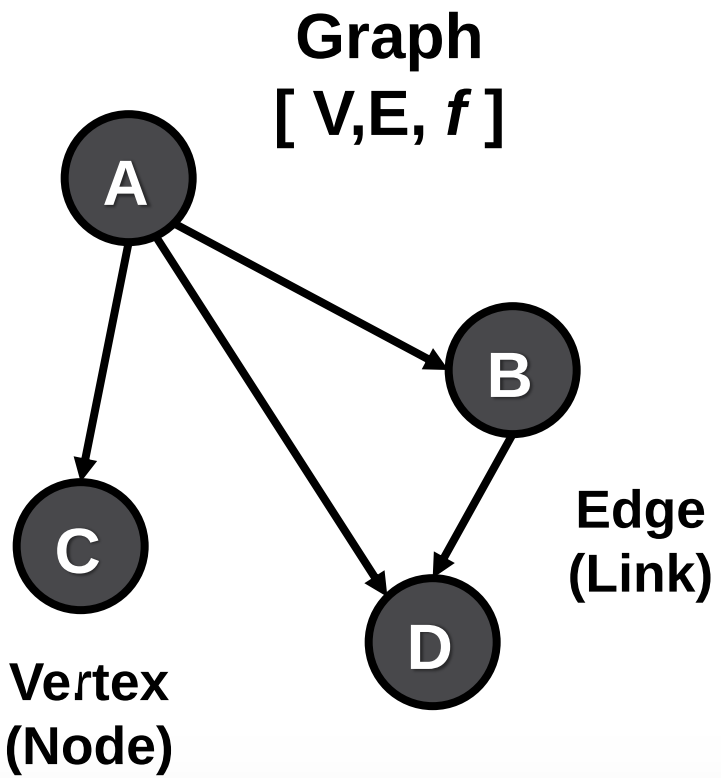
Social Network Analysis

Key Elements

Vertices or Nodes
The "things"

Edges or Links
The "relationships"

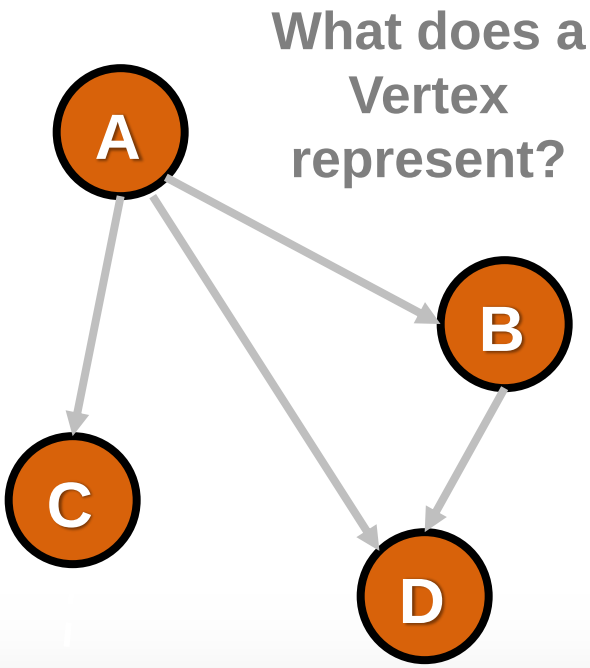
Graph or Network
The set of vertices/nodes, edges/links and the relationship/function connecting them.



Social Network Analysis

Types of Vertices

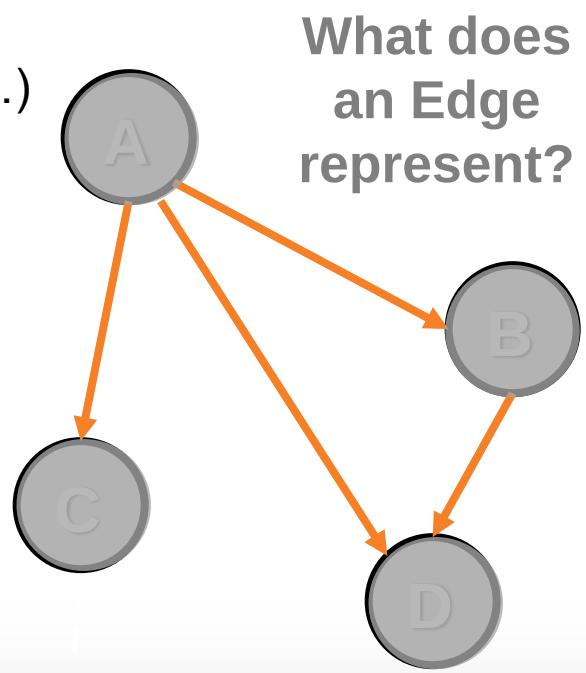
- Social Entities
  - People or social structures such as workgroups, teams, organizations, institutions, states, or even countries.
- Content
  - Web pages, keyword tags, or videos.
- Locations
  - Physical or virtual locations or events.
- Primary building blocks of social media
  - Friends in social networking sites, posts or authors in blogs, or pages in wikis.



Social Network Analysis

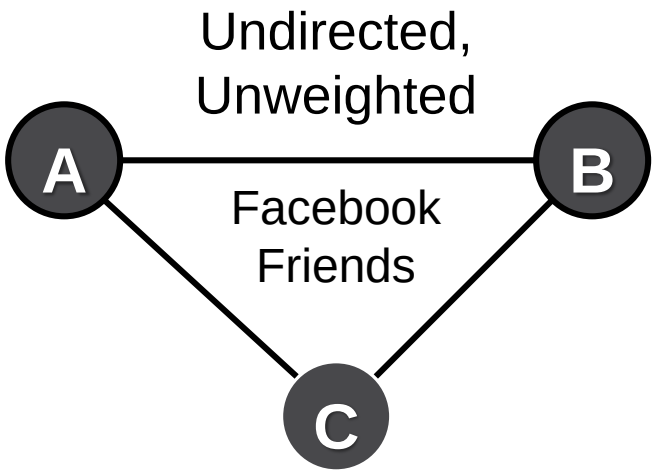
Types of Relations

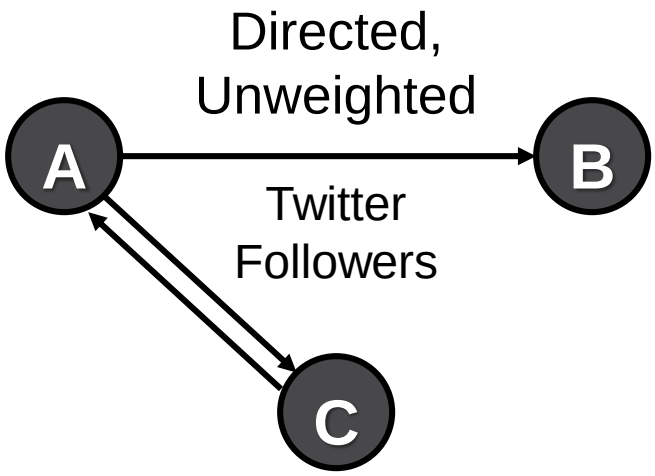
- Similarities
  - Location (same spatial and temporal space)
  - Participation (same club, same event, …)
  - Attributes (Age, gender, same attitudes, …)
- Relational Roles
  - Kinship (mother of, sibling of, …)
  - Other Roles (friend of, boss of, …)
- Relational Cognition
  - Affective (Likes, Hates…)
  - Perceptual (Knows, Knows of, …)
- Relational Events
  - Interactions (Sold to, talked to, helped, …)
  - Flows (Information, beliefs, money, …)

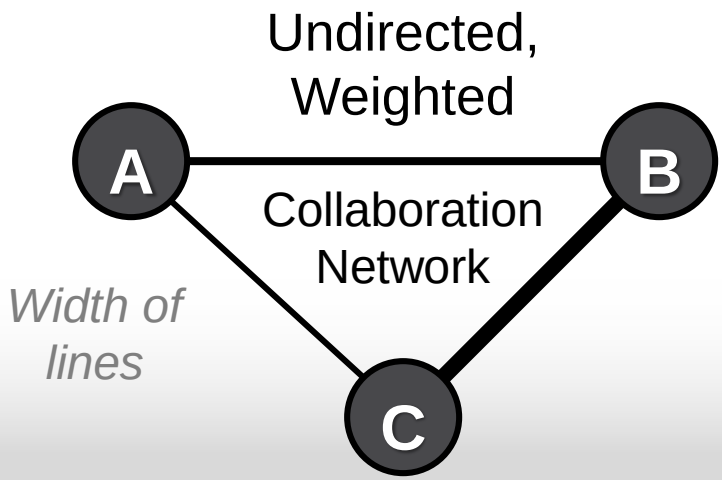


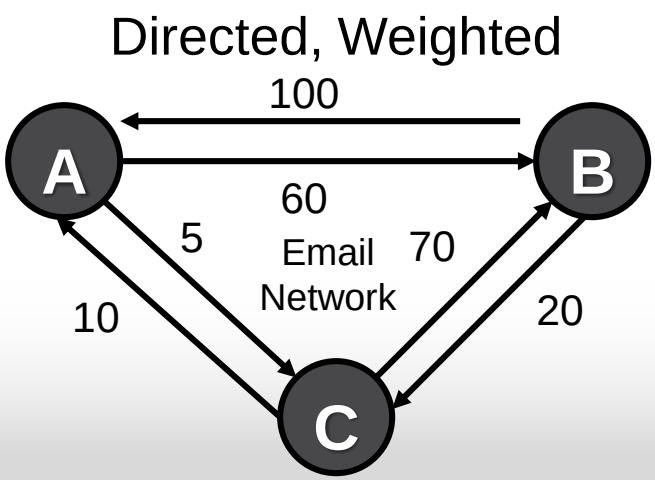
Social Network Analysis

Types of Edges or Links





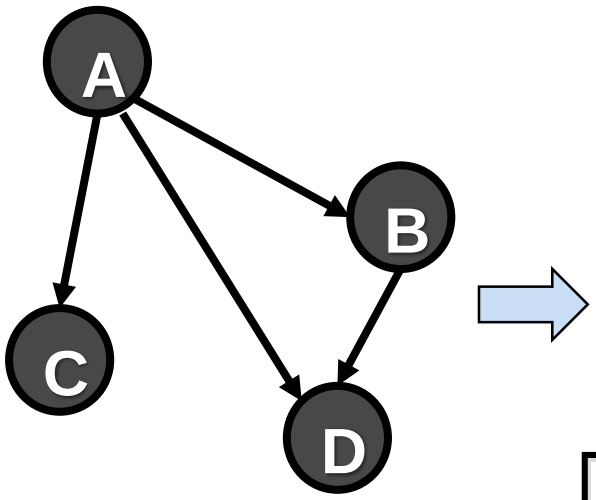




Social Network Analysis

Alternative Representations

Graph



Edge List

A	B
A	C
A	D
B	D

Adjacency Matrix

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

Adjacency List

A	B, C, D
C	D

XML

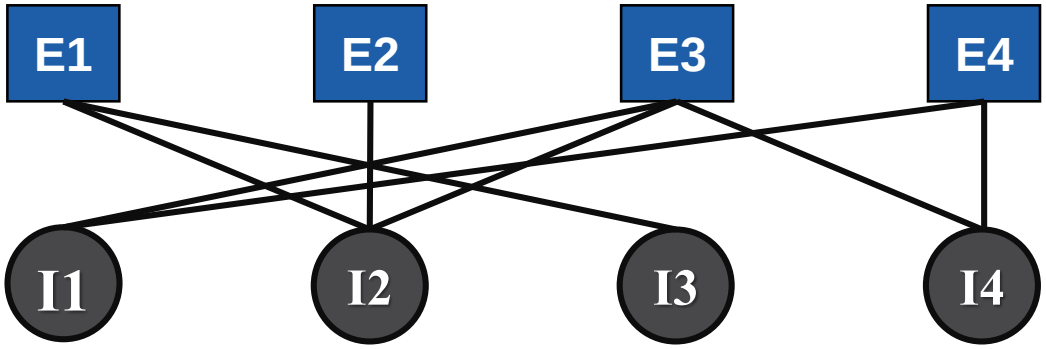
```
<Node>
    <Label> A </Label>
    <Connection> B </Connection>
    <Connection> C </Connection>
    <Connection> D </Connection>
</Node>
<Node>
    <Label> C </Label>
    <Connection> D </Connection>
</Node>
```

Social Network Analysis

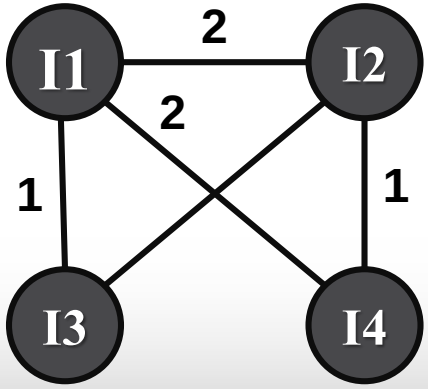
Multimodal networks

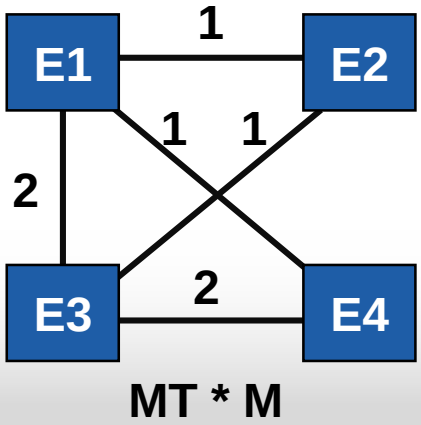
Bipartite or Bimodal Network
Linking individuals to events (participation, membership, topic, tag, post, …)

Examples:
Authors-to-papers (they authored)
Actors-to-Movies (they appeared in)
Users-to-Movies (they rated)
"Folded" networks:
Author collaboration networks
Movie co-rating networks



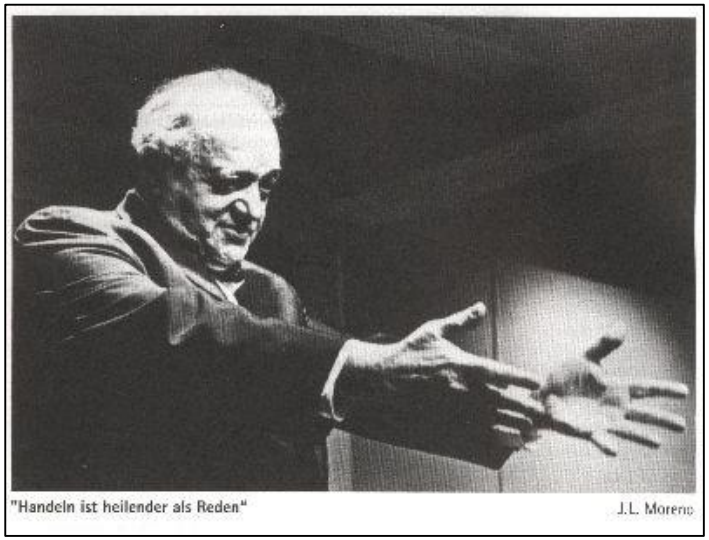
	E1	E2	E3	E4
I1	1	0	1	1
I2	1	1	1	0
I3	1	0	0	0
I4	0	0	1	1



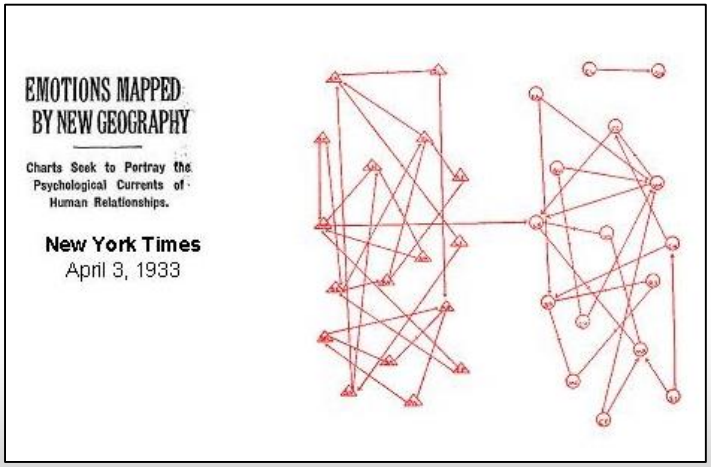


Social Network Analysis

One of the founding fathers



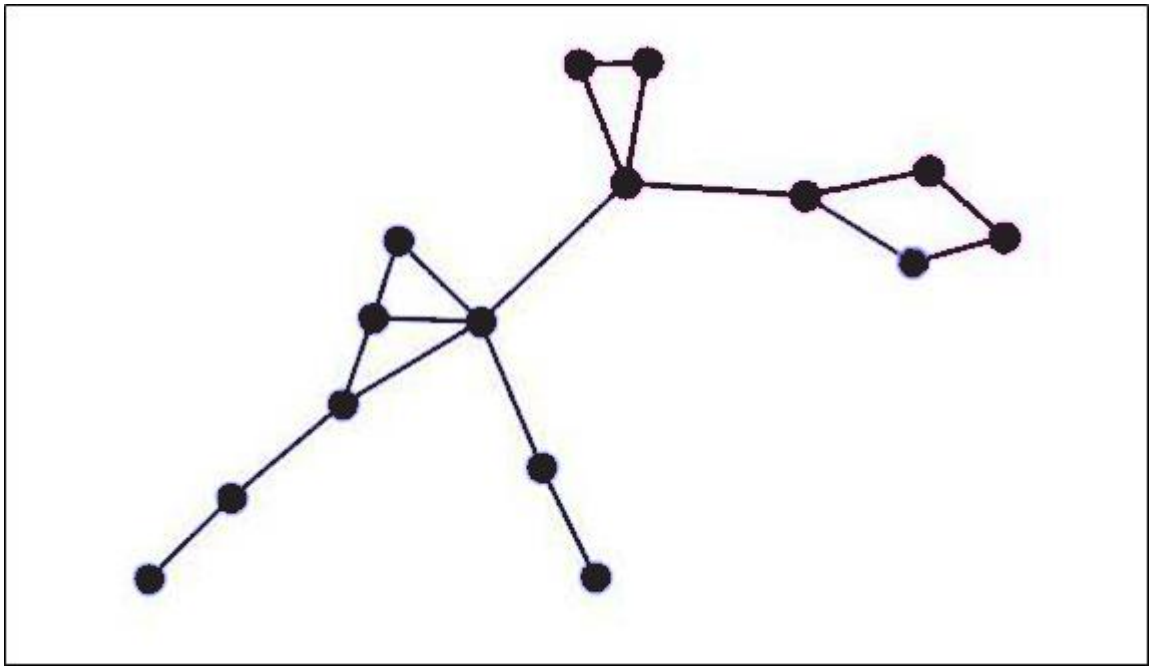
"Handeln ist heilender als Reden" J.L. Moreno



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

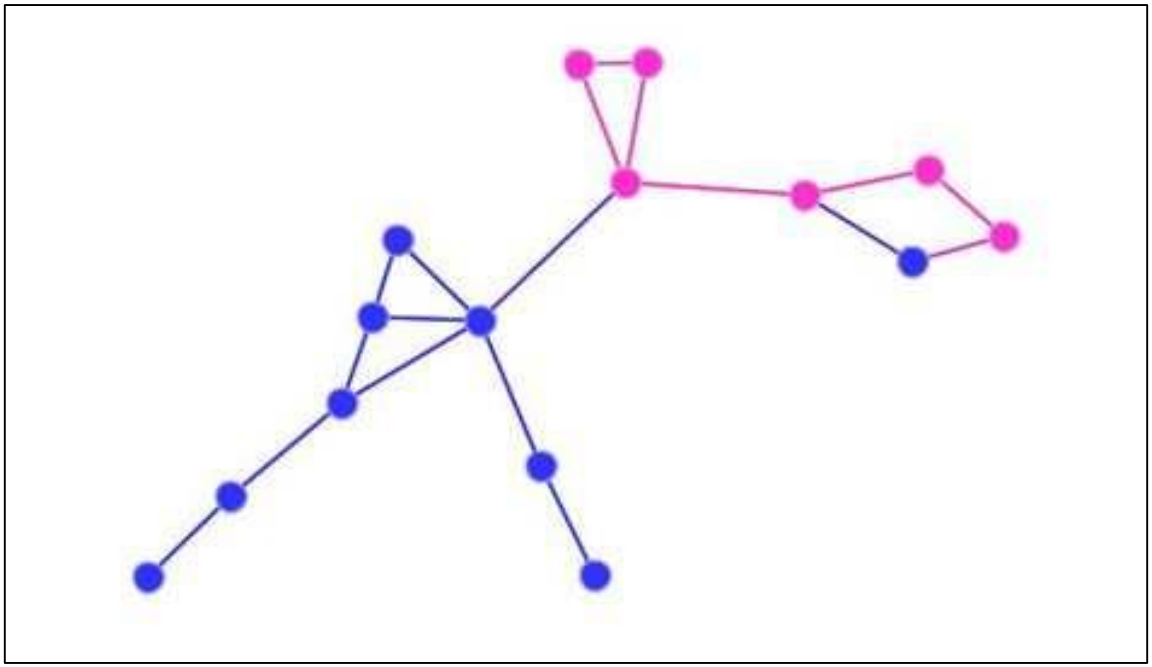
Social Homophily

Any guesses what this is and what it shows?



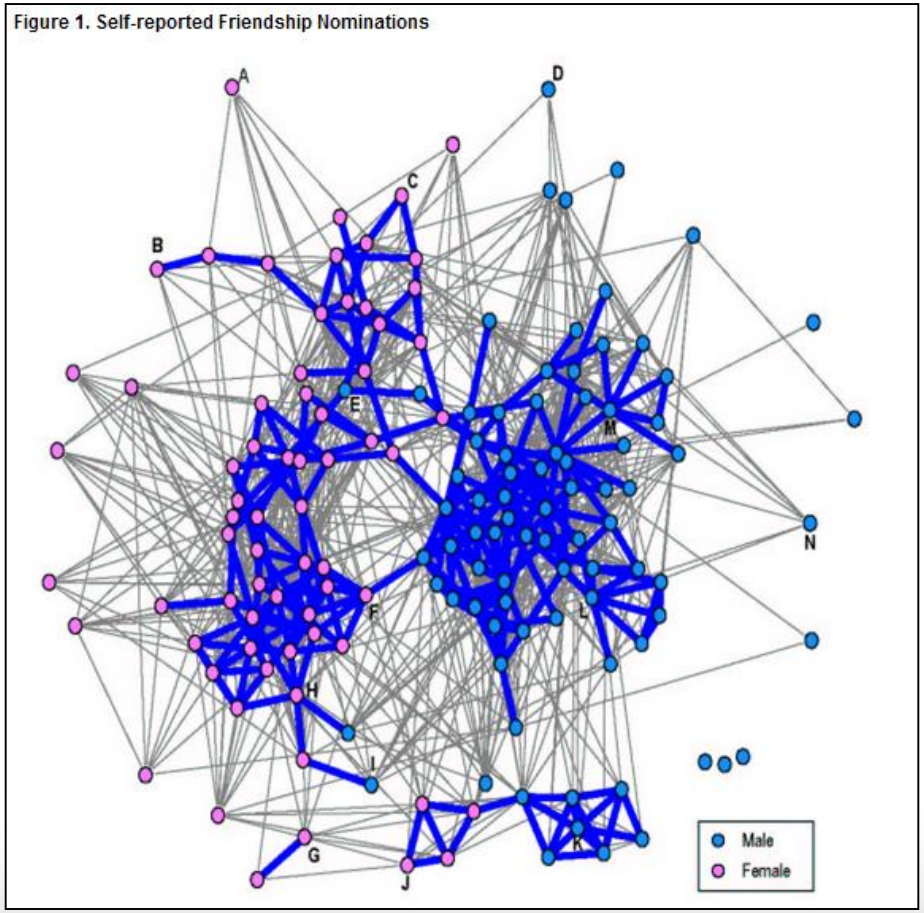
Social Homophily

Any guesses what this is and what it shows?



Social Homophily

Gender homophily in adolescent school group

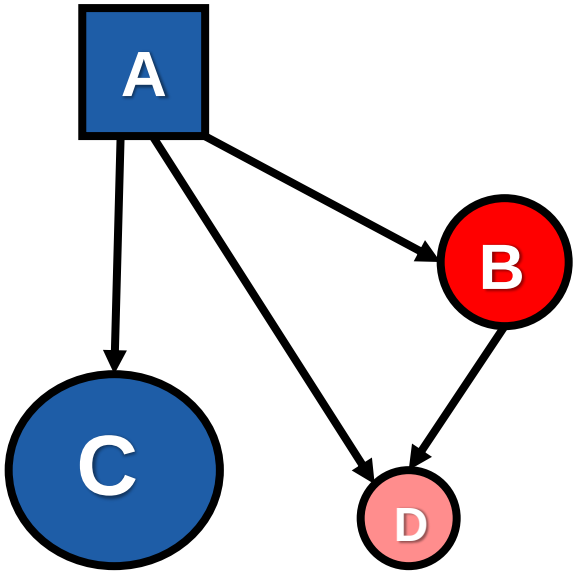


Density or Distinction? The Roles of Data Structure and Group Detection Methods in Describing Adolescent Peer Groups
Gest, S. et al. Journal of Social Structure: Vol. 6, http://www.cmu.edu/joss/content/articles/volume8/GestMoody/

Social Networks

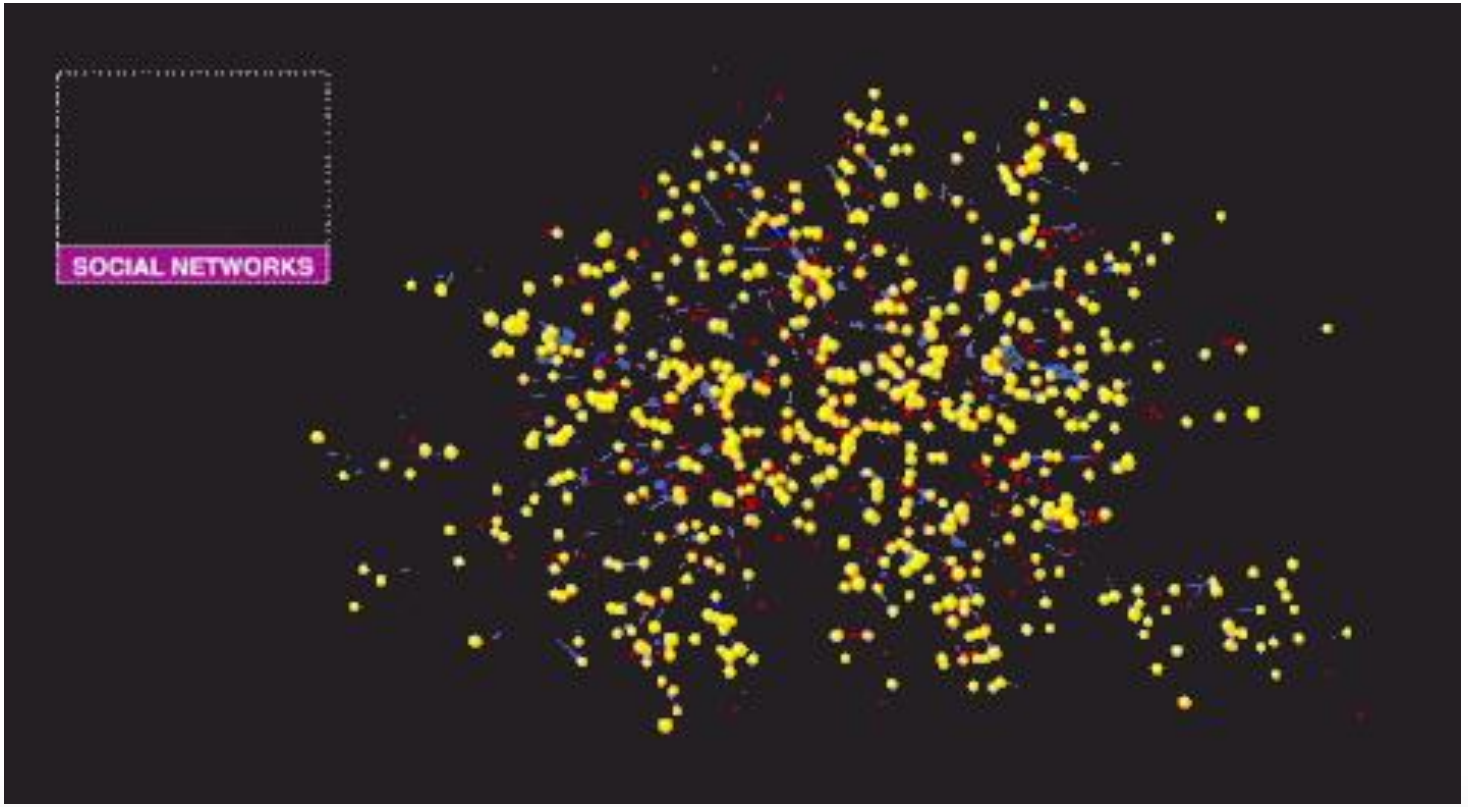
Role of Attributes

- Add insights to visualizations and analysis
- May include:
  - Demographic characteristics (age, gender, race)
  - Other characteristics (income, education, location)
  - Use of the system (number of logins, blogs posted, edits made)
  - Network Metrics (centralization)
- Often designated by color, shape, size and/or opacity of the vertices



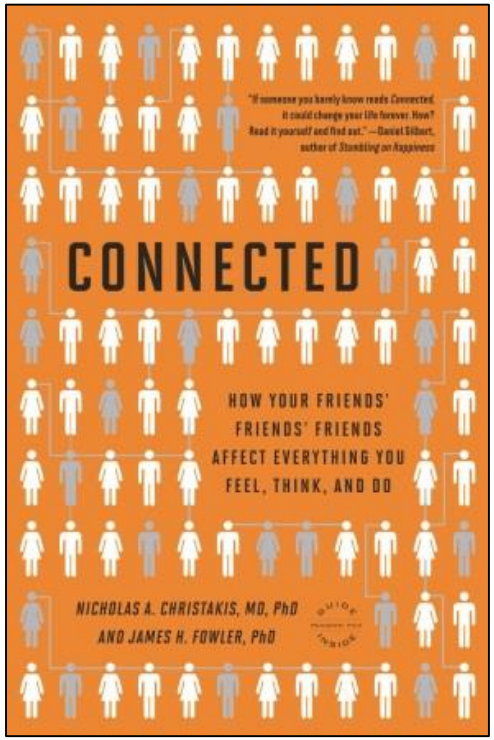
Social Homophily

Any guesses what this is and what it shows?



Social Homophily

The beauty of social networks



ted.com/talks/nicholas_christakis_the_hidden_influence_of_social_networks.html

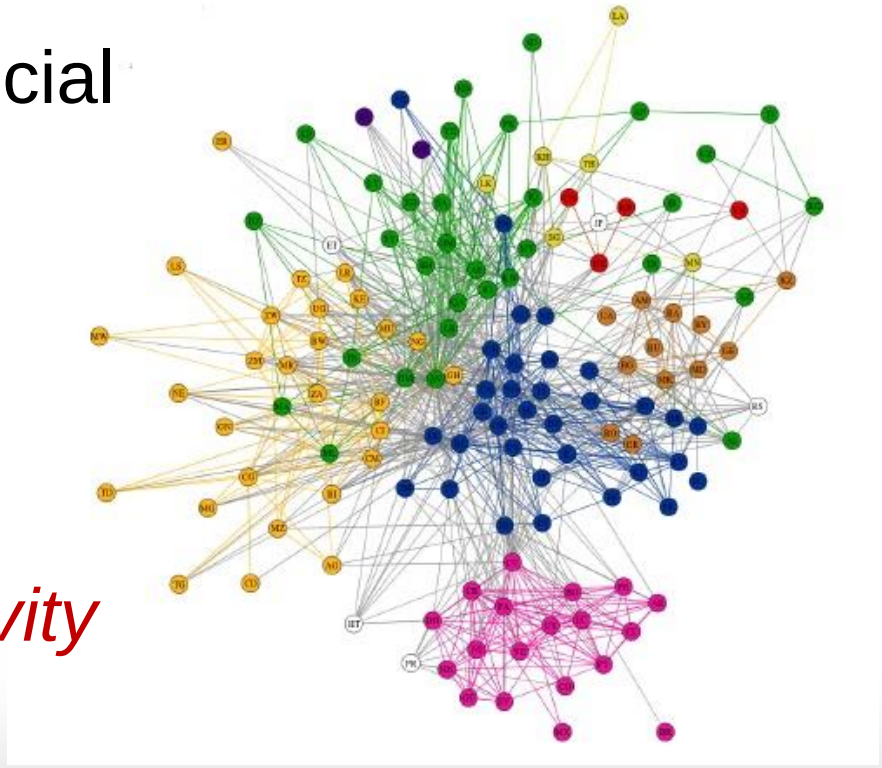
Social Homophily

Regardless of cause the resulting pattern is

Clusters of connections emerge throughout the social space such that there is

dense intra-connectivity within clusters

and *sparse inter-connectivity* between clusters



Social Homophily

What produces a pattern of homophily?

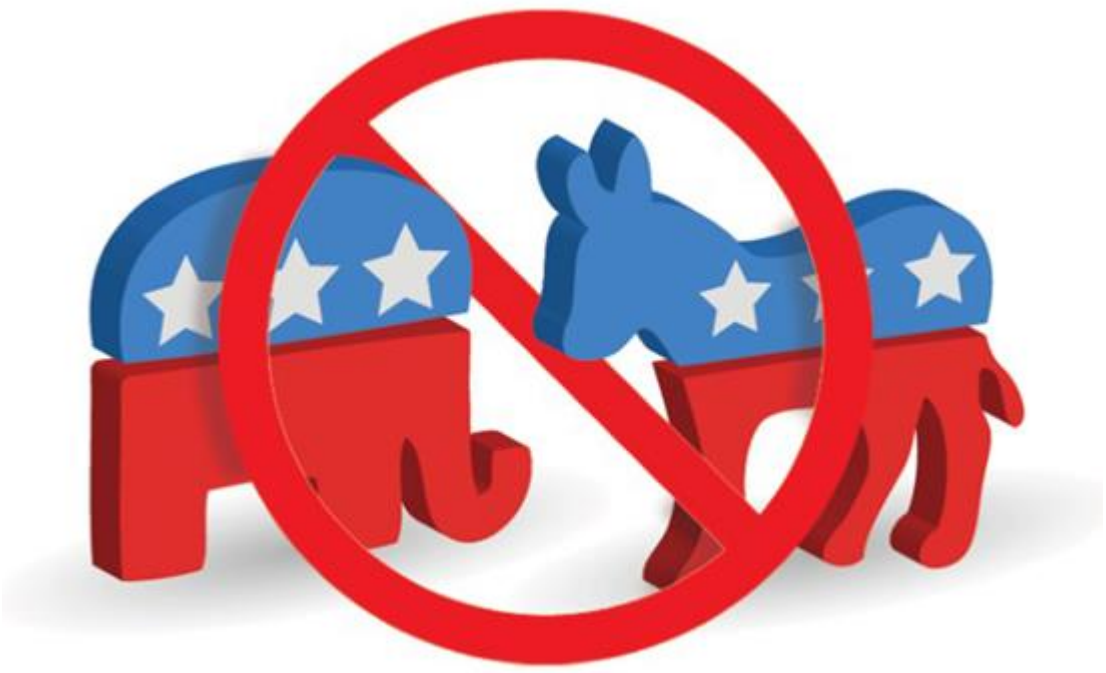
Cause: **Similarity breeds connection**

Effect: **Connection breeds similarity**

Spurious: **Some other factor is @ work**

Social Homophily

In the world of U.S. politics?

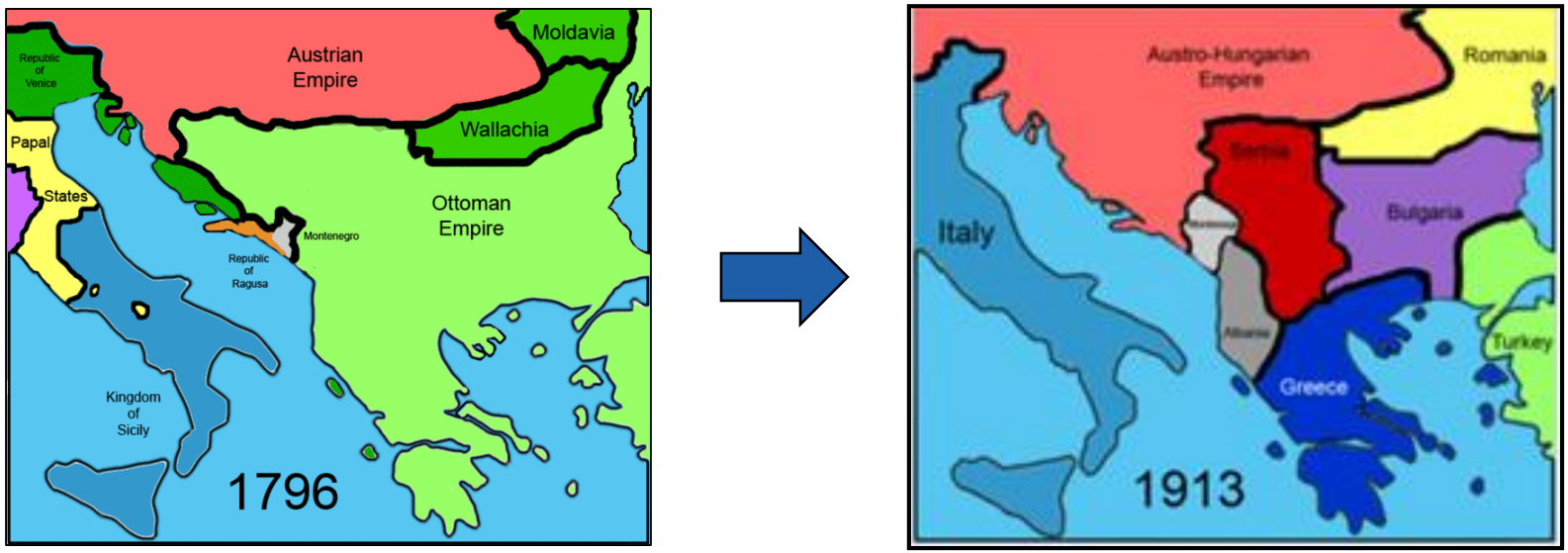


Discourse
Viewing
Voting
Purchasing
…

Antagonistic Polarization?

Social Homophily

Balkanization

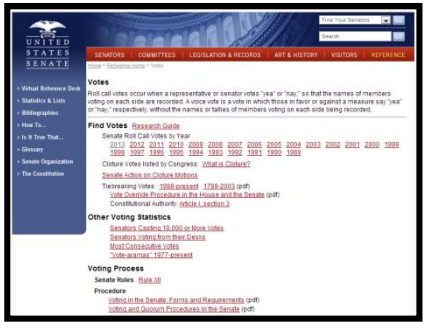


Process of fragmentation or division of a region or state into smaller regions or states that are often hostile or non-cooperative with each other

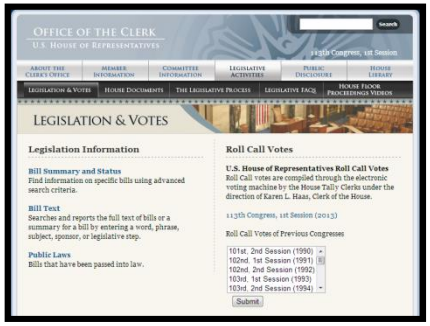
Social Homophily

Balkanization of the US Congress (Data Sets)

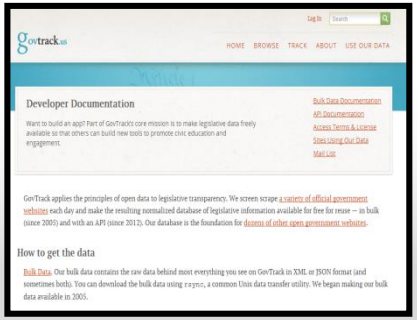
senate.gov/
reference/
Index/Votes.htm

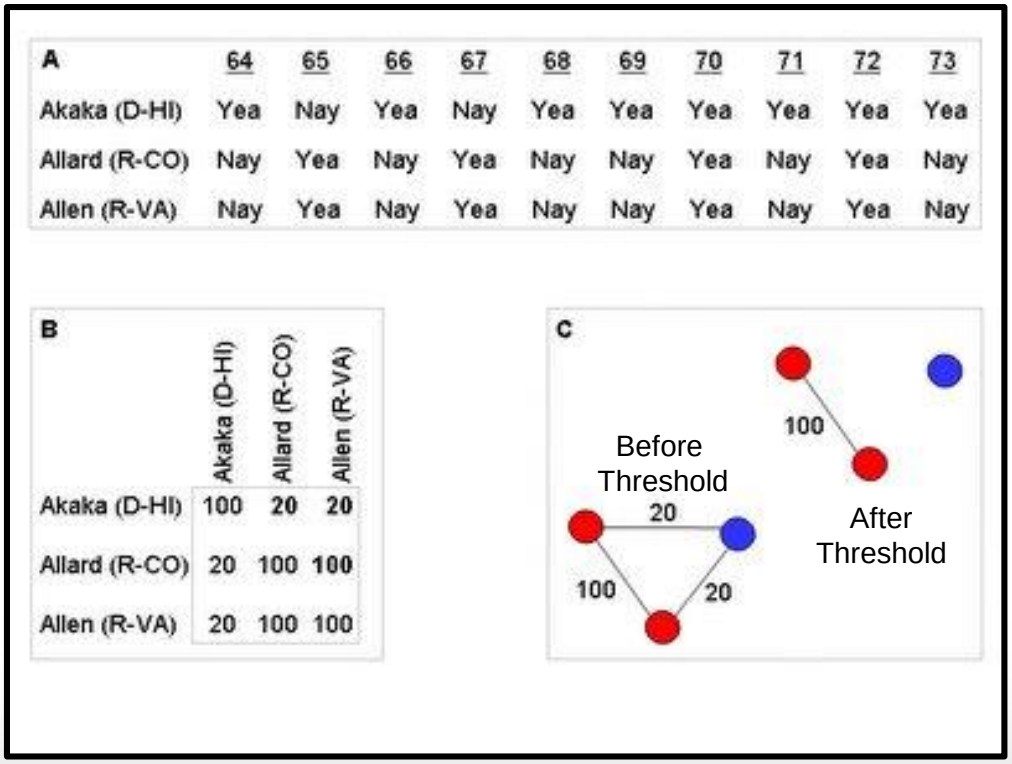


clerk.house.gov/
legislative/
legvotes.aspx



govtrack.us/
developers





A	64	65	66	67	68	69	70	71	72	73
Akaka (D-HI)	Yea	Nay	Yea	Nay	Yea	Yea	Yea	Yea	Yea	Yea
Allard (R-CO)	Nay	Yea	Nay	Yea	Nay	Nay	Yea	Nay	Yea	Nay
Allen (R-VA)	Nay	Yea	Nay	Yea	Nay	Nay	Yea	Nay	Yea	Nay

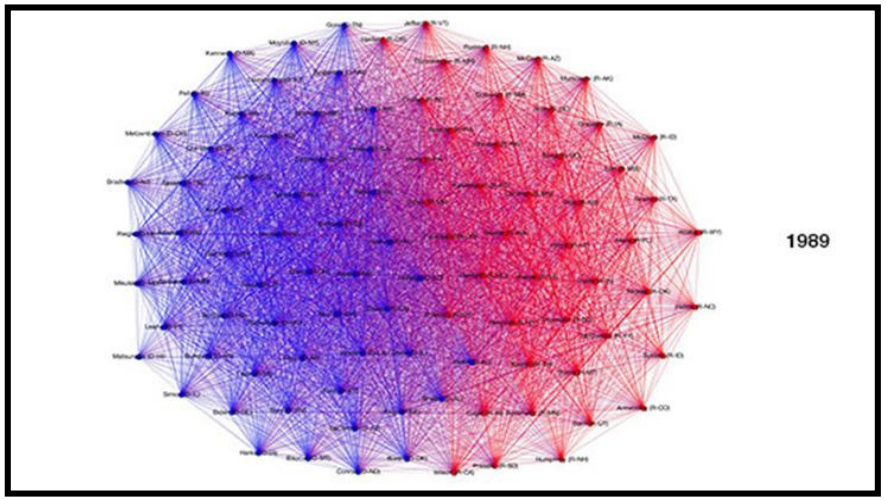
B

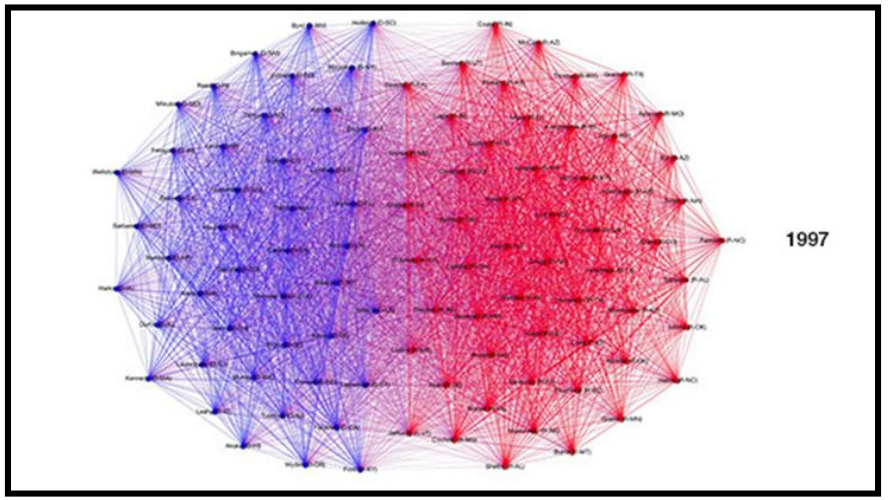
	Akaka (D-HI)	Allard (R-CO)	Allen (R-VA)
Akaka (D-HI)	100	20	20
Allard (R-CO)	20	100	100
Allen (R-VA)	20	100	100

"Senate Voting Patterns," Morris, F. analyticalvisions.blogspot.com/2006/04/senate-voting-patterns-part-2.html

Social Homophily

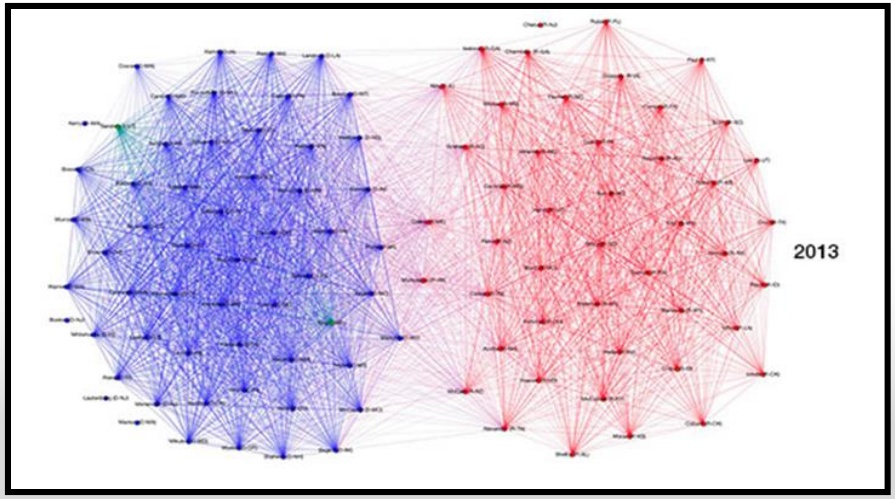
Balkanization of US Senate (1989-2013) - Votes





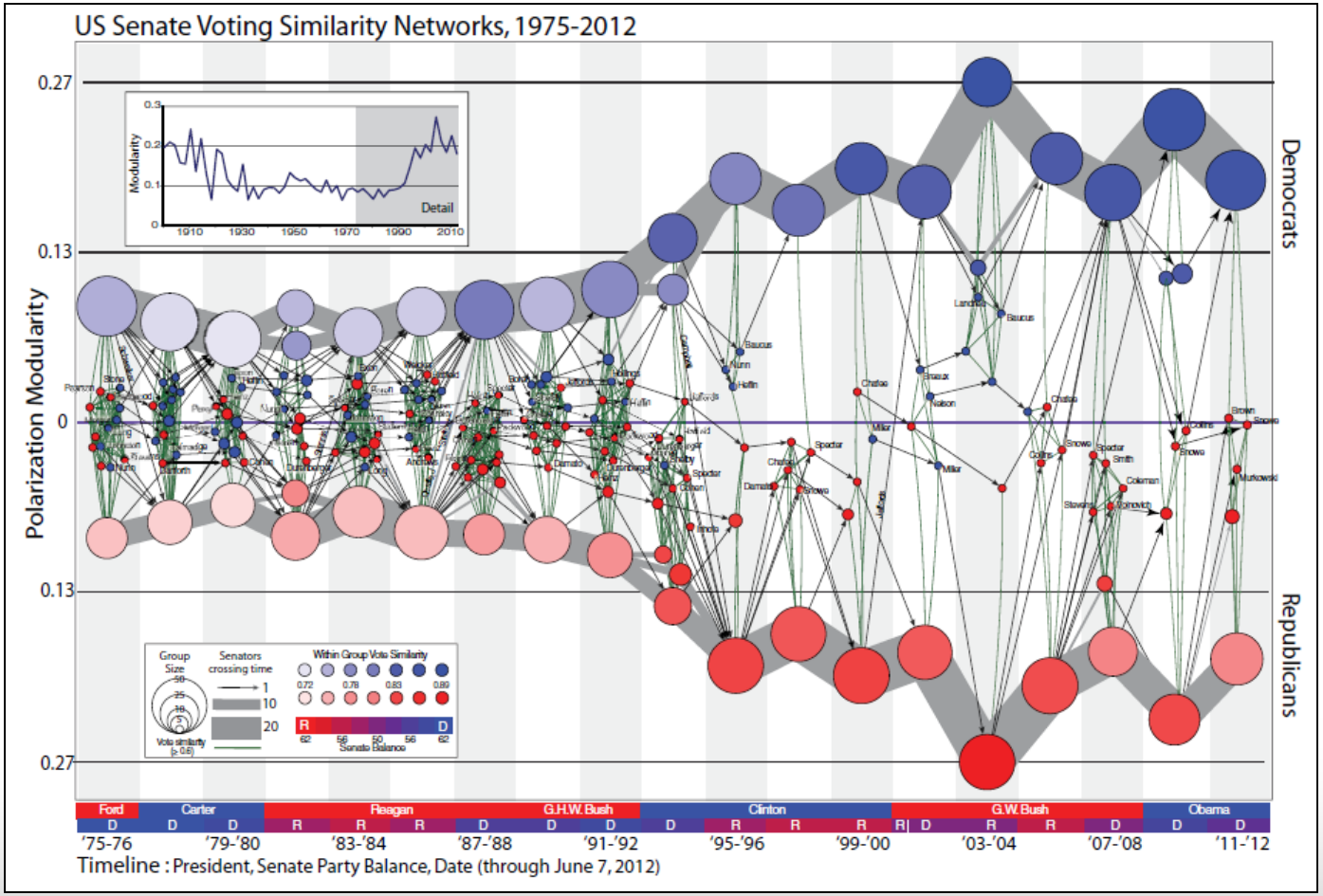
U.S. Senate More Divided Than Ever Data Shows, Swallow,E. ,Forbes, Oct. 17, 2013, forbes.com/sites/ericaswallow/2013/11/17/senate-voting-relationships-data/

Partisan voting patterns – 1989 to present
Interactive data visualization
static.davidchouinard.com/congress/



Social Homophily

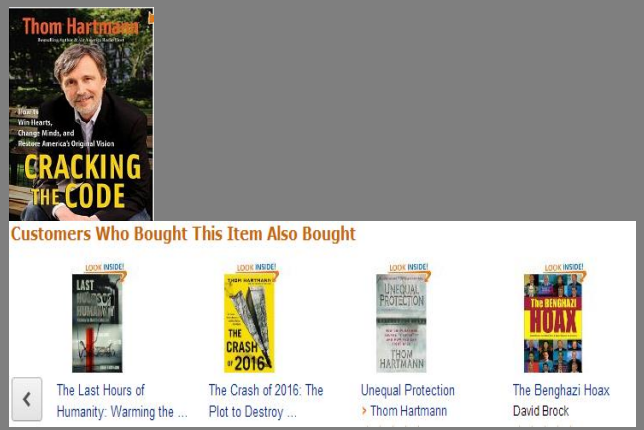
Balkanization of US Senate (1975-2012) - Votes



Portrait of Political Party Polarization, Moody J. and P. Mucha, April 2013, Network Science, journals.cambridge.org/action/displayAbstract?fromPage=online&aid=8888768

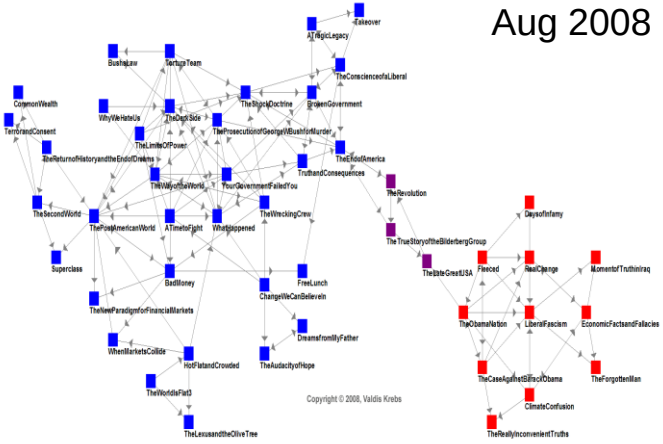
Social Homophily

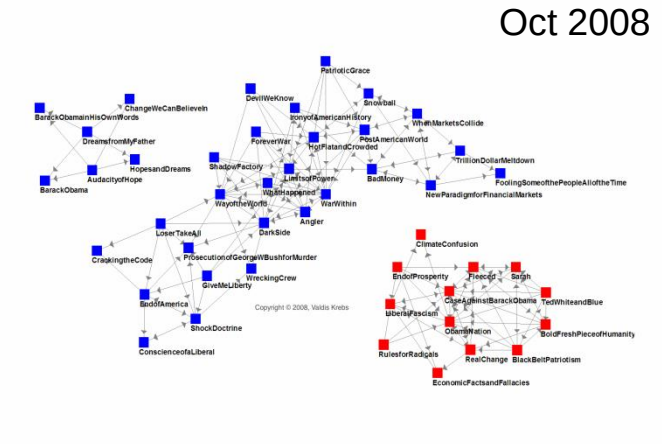
Are political books of "a feather" purchased together?



The Social Life of Books,Krebs V., orgnet.com/divided.html

		Book2	Book3	Book4	...	BookN
Book1	-	1	0	0	...	0
Book2	1	-	0	1	...	0
Book3	0	0	-	1	...	1
Book4	0	1	1	-	...	1
...	...	...	...	...	-	...
BookN	0	0	1	1	...	-





Social Homophily

Cyberbalkanization of the Political Blogsphere

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/011105B.html

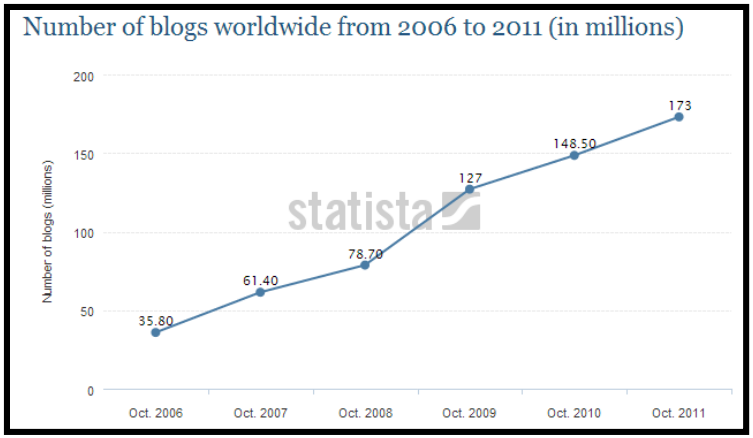
1

- 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
- Are we witnessing an era of “Cyberbalkinization?”

Social Homophily

An aside – what's a (We)blog?

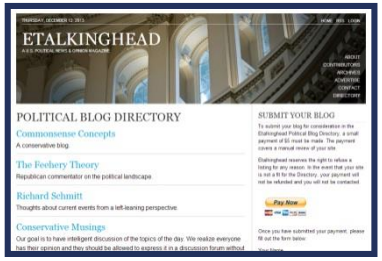
- Information or discussion site
  - Focused on a particular subject area
  - Discrete entries posted in reverse chronological order
  - Primarily text
  - Trend toward multiple author entries
- Some stats?
  - Wordpress ~ 60M blogs with 100M pageviews/day
  - Tumblr ~ 150M blogs with 90M posts/day
  - Blogger ~ 46M unique visitors/month



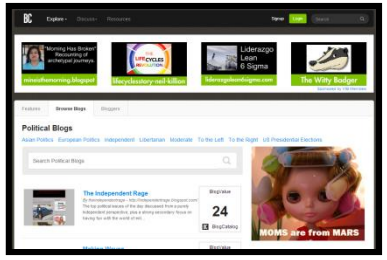
Social Homophily

Snowball Sample

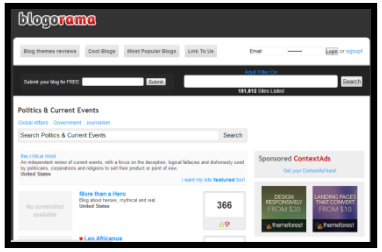
Political Blog Directories



eTalkingHead



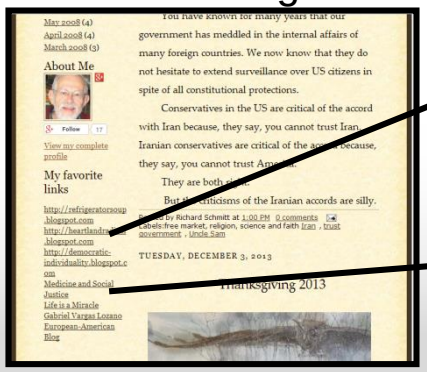
BlogCatalog

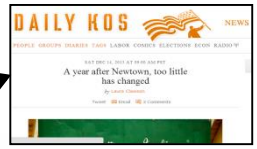


Blogarama

CampaignLine

Individual
Political Blogrolls

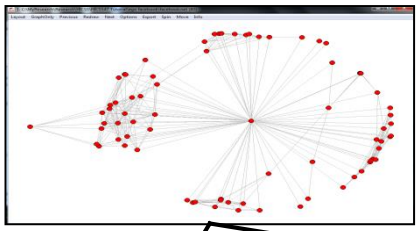


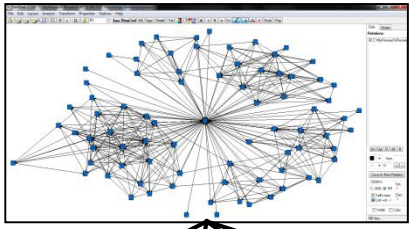


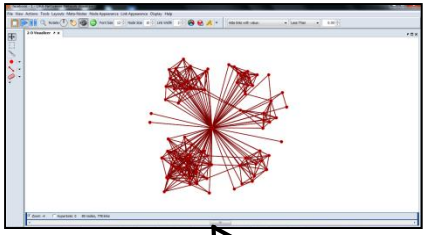
	Blog1	Blog2	Blog3	Blog4	...	BlogN
Blog1	-	1	0	0	...	0
Blog2	1	-	0	1	...	0
Blog3	0	0	-	1	...	1
Blog4	0	1	1	-	...	1
...	...	...	...	...	-	...
BlogN	0	0	1	1	...	-

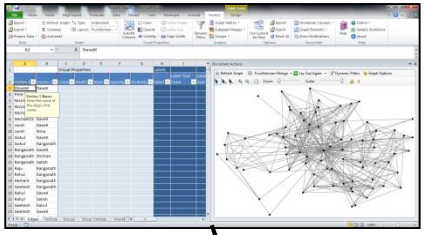
Social Homophily

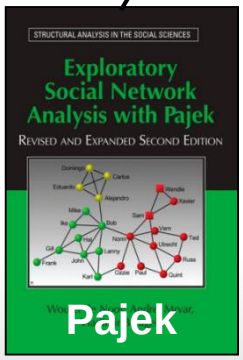
Social Network Packages

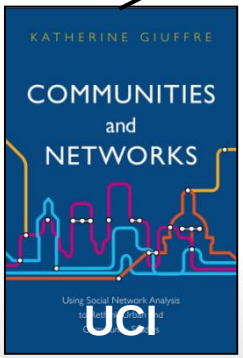


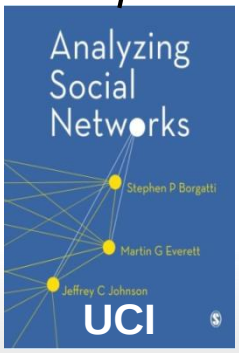


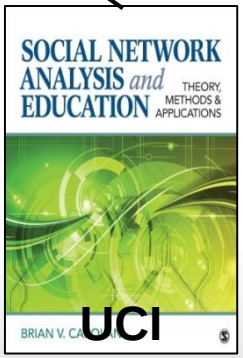


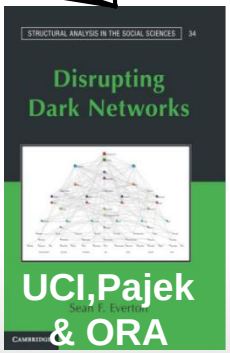


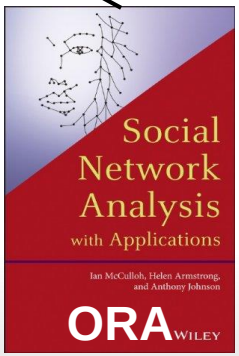


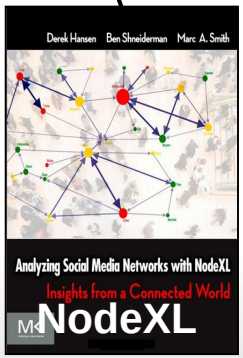






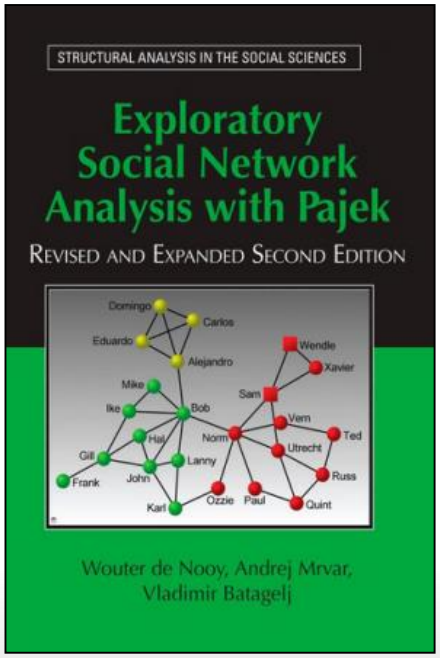


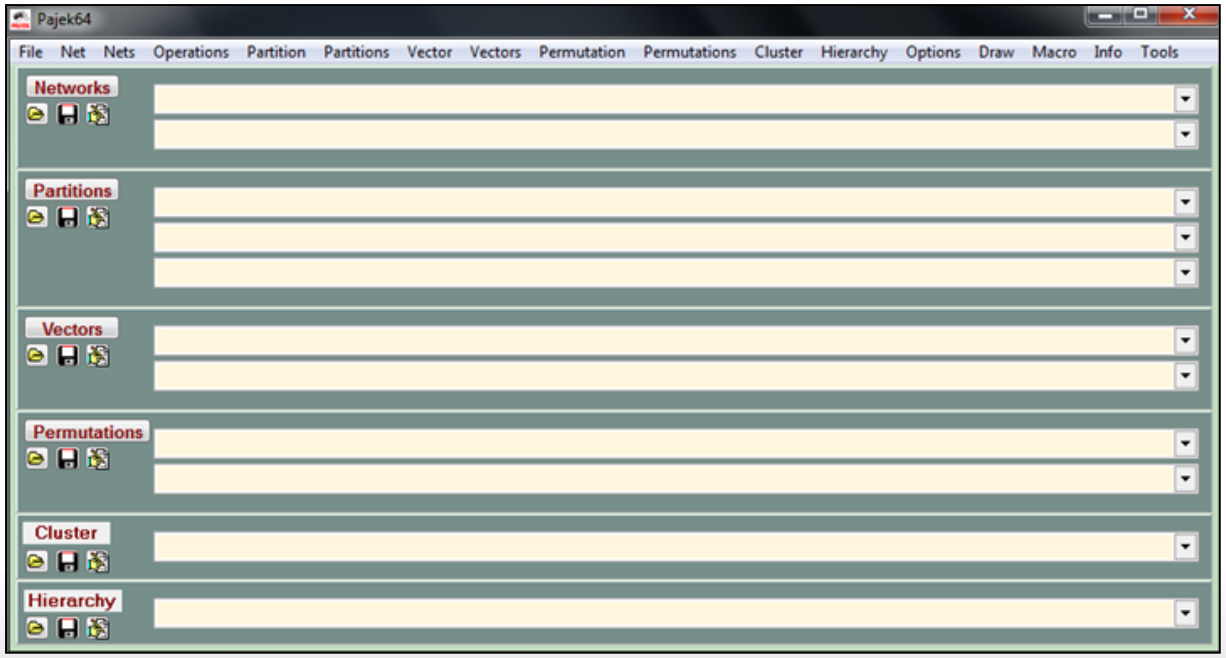




Social Homophily

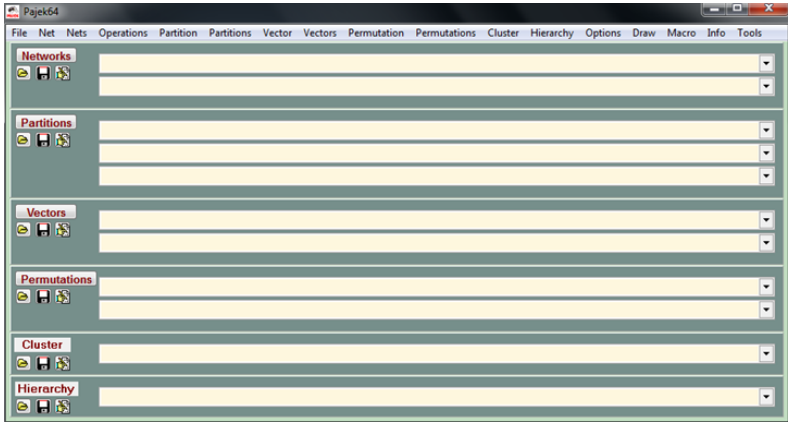
Social Network Analysis of Blogs with Pajek





Social Homophily

Social Network Analysis of Blogs with Pajek



- **Colors and bold simply for illustration**
- **Liberal Blogs (nos. 1-758)**
- **Conservative Blogs (nos. 759 – 1490)**

.NET Network File
(Vertex descriptions followed By Edge descriptions)

***vertices 1490**
1 100monkeystyping 0.0 0.0
2 12thharmonic 0.0 0.0
…
757 yoder 0.0 0.0
758 younglibs 0.0 0.0
…
759 84rules 0.0 0.0
760 a100wwe 0.0 0.0
…
1489 zeke01 0.0 0.0
1490 zeph1z 0.0 0.0
***edges**
1 23 1.0
1 55 1.0
…
757 65 1.0
760 854 1.0
760 855 1.0
…
1489 1431 1.0
1490 802 1.0

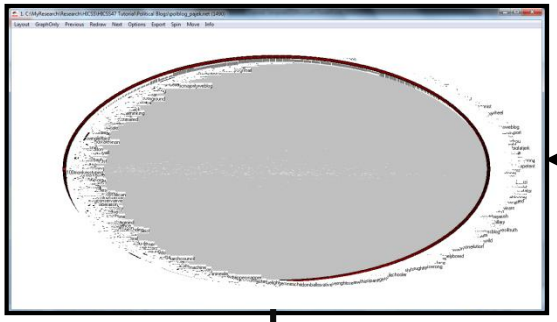
.CLU Partition File
(0-Lib; 1 – Cons
One entry per vertex)

***vertices 1490**
0
0
0
…
0
1
1
1
…
1

Social Homophily

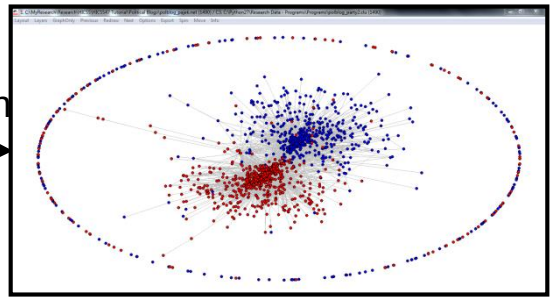
Social Network Analysis of Blogs with Pajek

Circular Layout w. labels

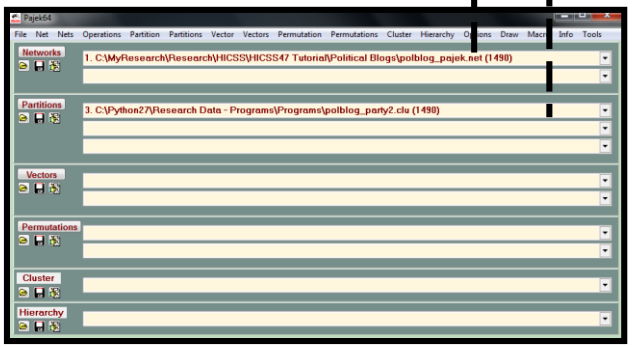


Draw Network

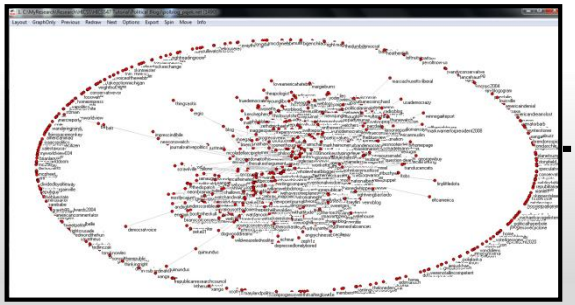
F-R 2D Layout w/o labels partitioned by pol. position



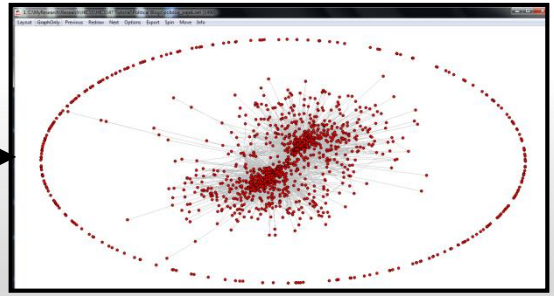
Draw Partition



Fructerman-Reingold 2D Layout with labels

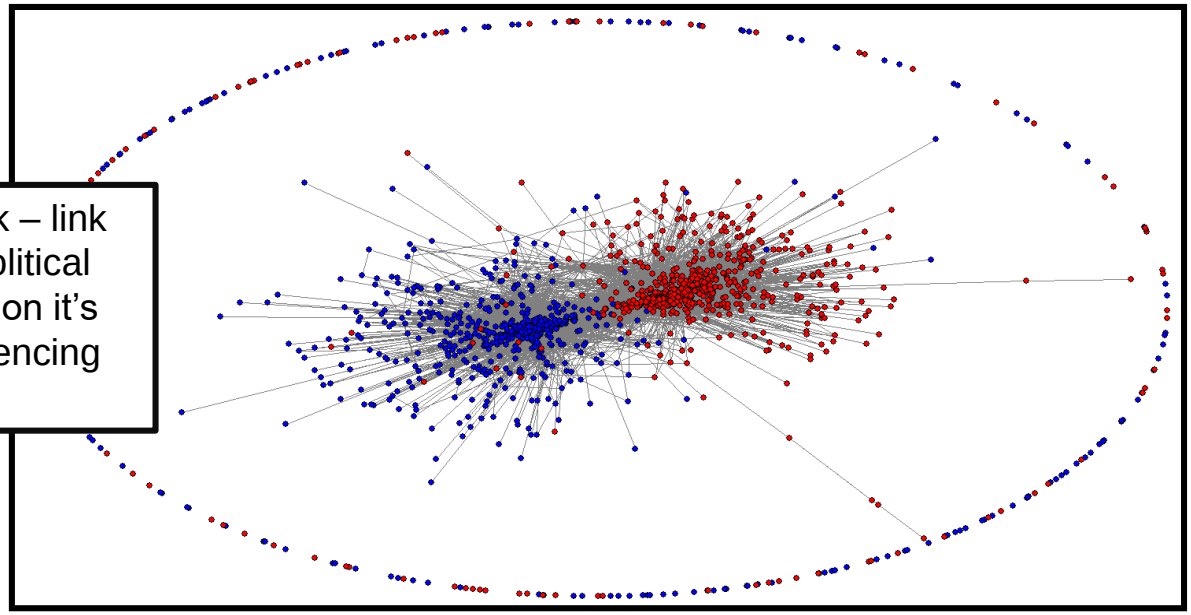


F-R 2D Layout w/o labels



Social Homophily

Cyberbalkanization of the Political Blogsphere?

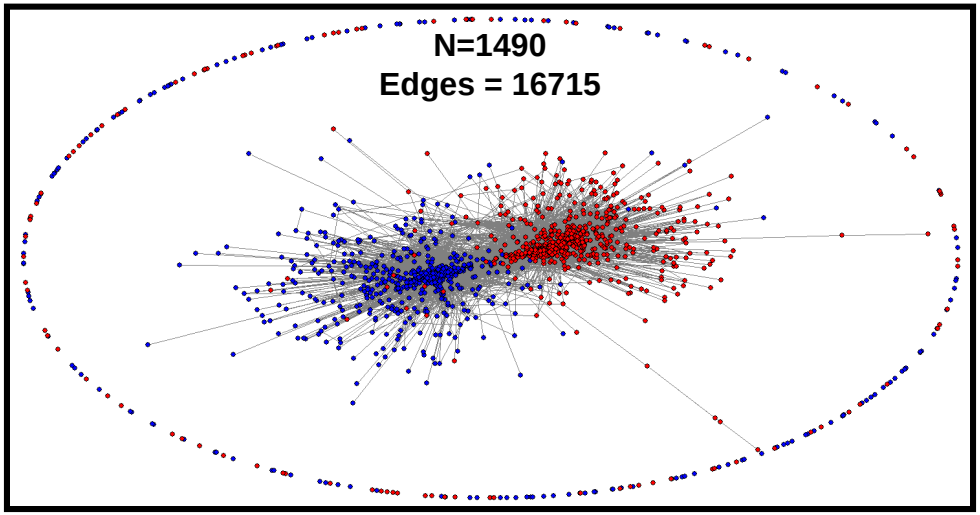


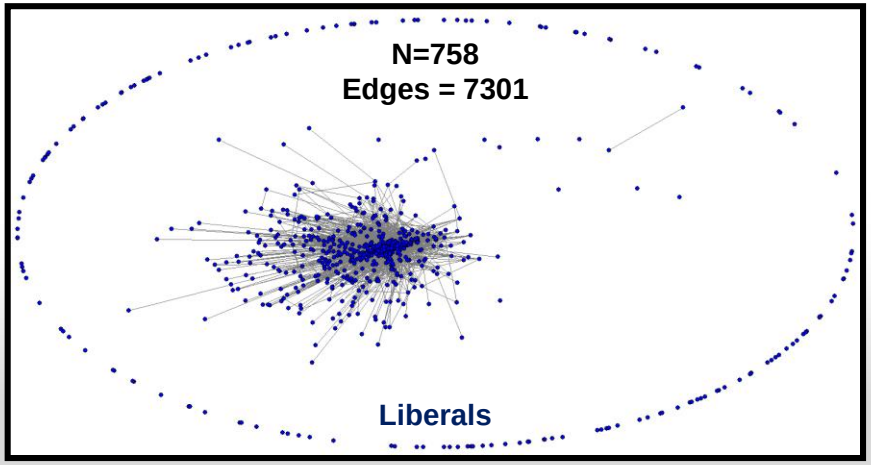
Citation Network – link implies that a political blog has a URL on it's front page referencing another blog

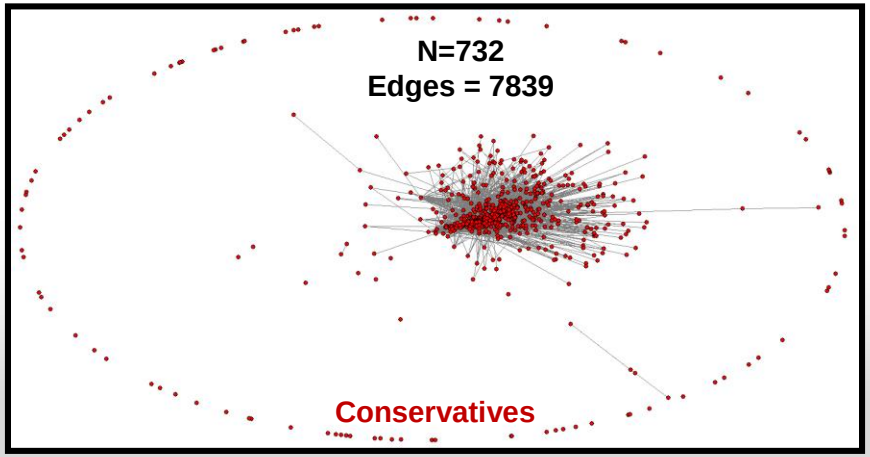
Proliferation of specialized online political news sources allows people with different political leanings to be exposed only to information in agreement with their previously held views?

Social Homophily

Cyberbalkanization of the Political Blogsphere?



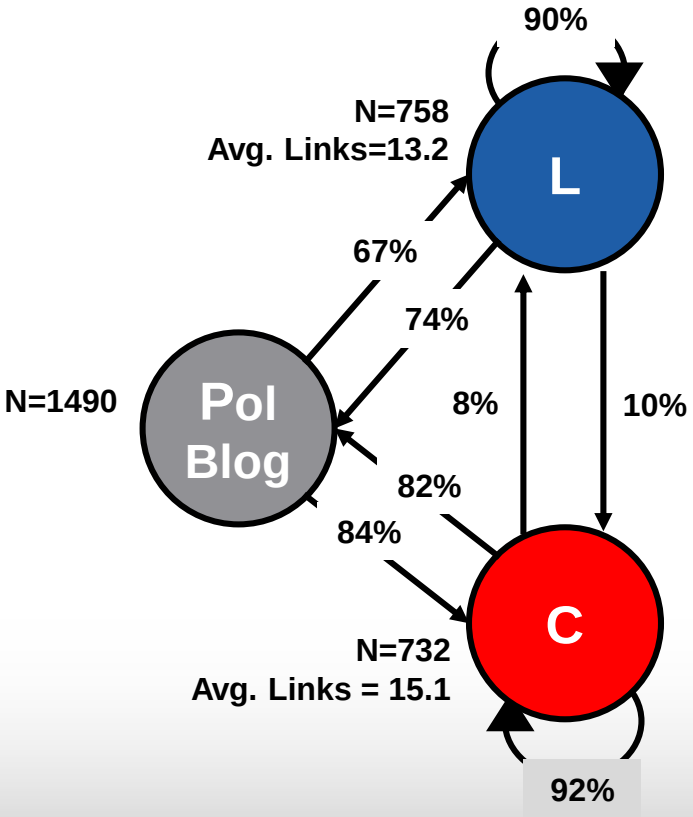




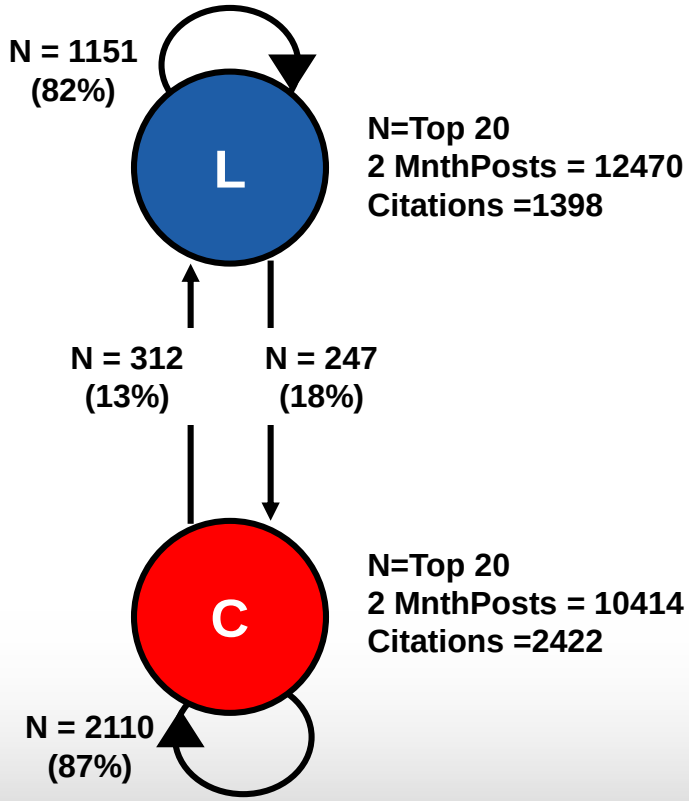
Social Homophily

Cyberbalkanization of the Political Blogsphere

Blogrolls



Page Citation



Social Homophily

Polarization on Twitter

Political Polarization on Twitter

M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini, F. Menczer
Center for Complex Networks and Systems Research
School of Informatics and Computing
Indiana University, Bloomington, IN, USA

Abstract

In this study we investigate how social media shape the networked public sphere and facilitate communication between communities with different political orientations. We examine two networks of political communication on Twitter, comprised of more than 250,000 tweets from the six weeks leading up to the 2010 U.S. congressional midterm elections. Using a combination of network clustering algorithms and manually-annotated data we demonstrate that the network of political retweets exhibits a highly segregated partisan structure, with extremely limited connectivity between left- and right-leaning users. Surprisingly this is not the case for the user-to-user mention network, which is dominated by a single politically heterogeneous cluster of users in which ideologically-opposed individuals interact at a much higher rate compared to the network of retweets. To explain the distinct topologies of the retweet and mention networks we conjecture that politically motivated individuals provoke interaction by injecting partisan content into information streams whose primary audience consists of ideologically-opposed users. We conclude with statistical evidence in support of this hypothesis.

1 Introduction

Social media play an important role in shaping political discourse in the U.S. and around the world (Bennett 2003; Benkler 2006; Sunstein 2007; Farrell and Drezner 2008; Aday et al. 2010; Tumasjan et al. 2010; O'Connor et al. 2010). According to the Pew Internet and American Life Project, six in ten U.S. internet users, nearly 44% of American adults, went online to get news or information about politics in 2008. Additionally, Americans are taking an active role in online political discourse, with 26% of internet users contributing comments or questions about the political process to social networking sites, blogs or other online forums (Pew Internet and American Life Project 2008).

Despite this, some empirical evidence suggests that politically active web users tend to organize into insular, homogenous communities segregated along partisan lines. Adamic and Glance (2005) famously demonstrated that political blogs preferentially link to other blogs of the same political ideology, a finding supported by the work of Hargittai, Gallo, and Kane (2007). Consumers of online political information tend to behave similarly, choosing to read blogs that share their political beliefs, with 26% more users doing so in 2008 than 2004 (Pew Internet and American Life Project 2008).

In its own right, the formation of online communities is not necessarily a serious problem. The concern is that when politically active individuals can avoid people and information they would not have chosen in advance, their opinions are likely to become increasingly extreme as a result of being exposed to more homogeneous viewpoints and fewer credible opposing opinions. The implications for the political process in this case are clear. A deliberative democracy relies on a broadly informed public and a healthy ecosystem of competing ideas. If individuals are exposed exclusively to people or facts that reinforce their pre-existing beliefs, democracy suffers (Sunstein 2002; 2007).

In this study we examine networks of political communication on the Twitter microblogging service during the six weeks prior to the 2010 U.S. midterm elections. Sampling data from the Twitter 'gardenhose' API, we identified 250,000 politically relevant messages (*tweets*) produced by more than 45,000 users. From these tweets we isolated two networks of political communication — the *retweet* network, in which users are connected if one has rebroadcast content produced by another, and the *mention* network, where users are connected if one has mentioned another in a post, including the case of tweet replies.

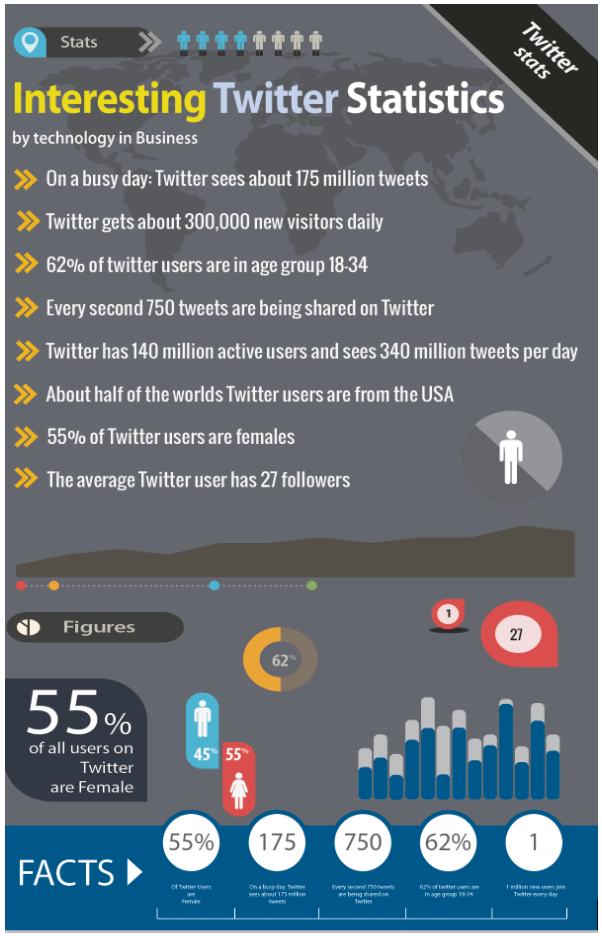
We demonstrate that the retweet network exhibits a highly modular structure, segregating users into two homogenous communities corresponding to the political left and right. In contrast, we find that the mention network does not exhibit this kind of political segregation, resulting in users being exposed to individuals and information they would not have been likely to choose in advance.

Finally, we provide evidence that these network structures result in part from politically motivated individuals annotating tweets with multiple *hashtags* whose primary audiences consist of ideologically-opposed users, a behavior also documented in the work of Yardi and boyd (2010). We argue that this process results in users being exposed to content they are not likely to rebroadcast, but to which they may respond using mentions, and provide statistical evidence in support of this hypothesis.

- Twitter communications six weeks prior to 2010 mid-term elections (9/14-11/1)
- Examined 250,000 “politically relevant” tweets (specific hashtags)
- Focused on two networks – retweets and mentions using “streaming API”

Social Homophily

An Aside – What is Twitter?



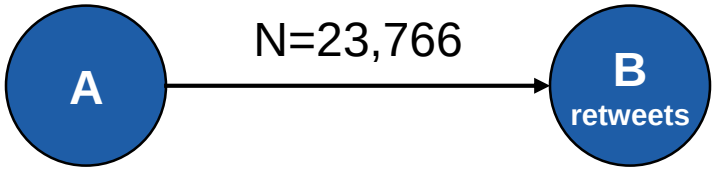
http://www.forbes.com/sites/marketshare/2013/01/23/the-number-one-mistake-retail-brands-make-when-it-comes-to-twitter/

- Each tweet <= 140 characters (avg. 10-15 words/message)
- Heavy presence of non-alpha symbols, abbrevs, misspellings and slang
- Tweets often include
  - Retweets (original tweet repeated)
  - Mentions (user mentions another user in a tweet)
  - Hashtags “#” (metadata annotation denoting topic or audience)

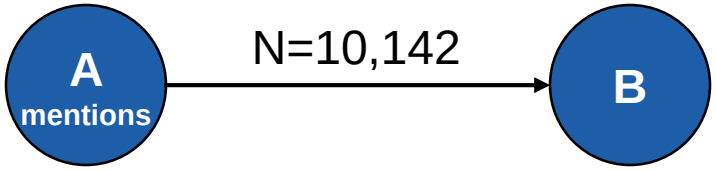
Social Homophily

Polarization on Twitter – Social Networks (in study)

Network 1:



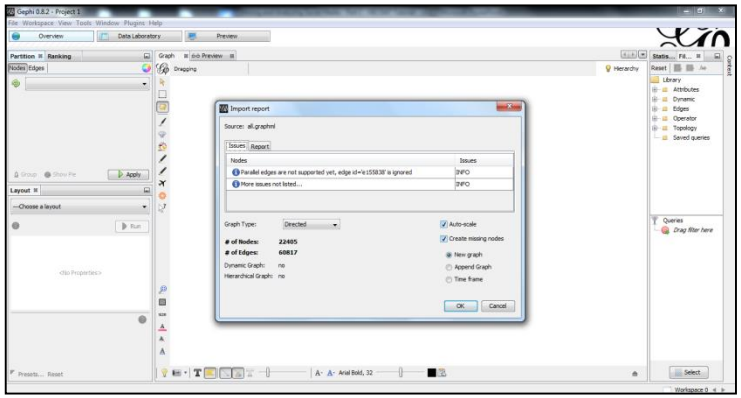
Network 2:



Both representing potential pathways
for information to flow from A to B

Social Homophily

Polarization on Twitter – Social Networks



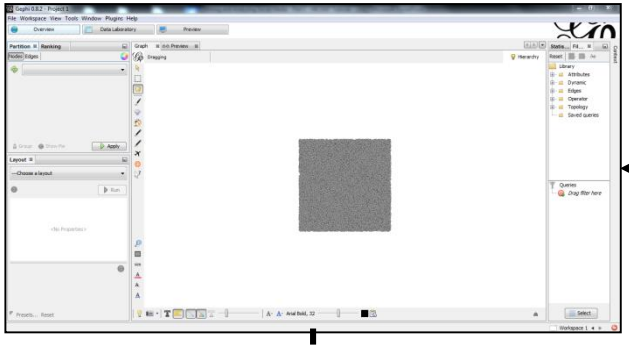
Gephi Representation
GraphicXML
Retweet Network

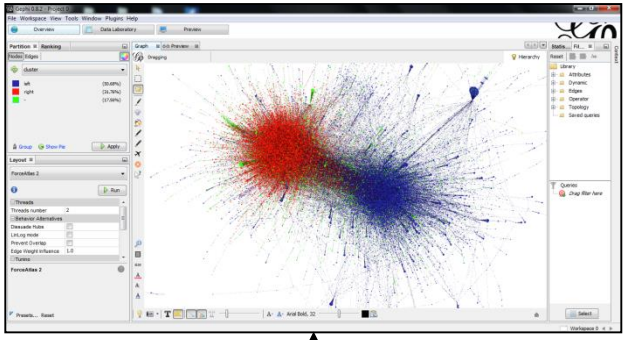
Source:
http://carl.cs.indiana.edu/data/icwsm/icwsm_polarization.zip

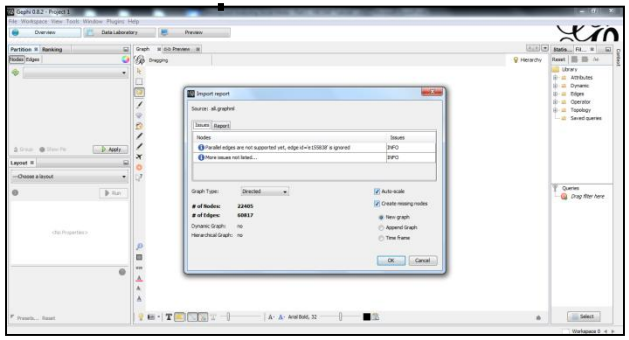
```
<?xml version="1.0" encoding="UTF-8"?>
<graphml xmlns="http://graphml.graphdrawing.org/xmlns"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
       xsi:schemaLocation="http://graphml.graphdrawing.org/xmlns
        http://graphml.graphdrawing.org/xmlns/1.0/graphml.xsd">
<!-- Created by igraph -->
  <key id="cluster" for="node" attr.name="cluster" attr.type="string"/>
  <key id="tags" for="edge" attr.name="tags" attr.type="string"/>
  <key id="type" for="edge" attr.name="type" attr.type="string"/>
  <key id="urls" for="edge" attr.name="urls" attr.type="double"/>
  <key id="time" for="edge" attr.name="time" attr.type="string"/>
  <graph id="G" edgedefault="directed">
    <node id="n0">
      <data key="cluster">right</data>
    </node>
    <node id="n2">
      <data key="cluster">left</data>
    </node>
 ...
    <edge source="n12464" target="n7349">
      <data key="tags">[&apos;#tcot&apos;, &apos;#tlot&apos;]</data>
      <data key="type">retweet</data>
      <data key="urls">1</data>
      <data key="time">1286901355</data>
    </edge>
...
  </graph>
</graphml>
```

Social Homophily

Polarization on Twitter – Gephi Analysis

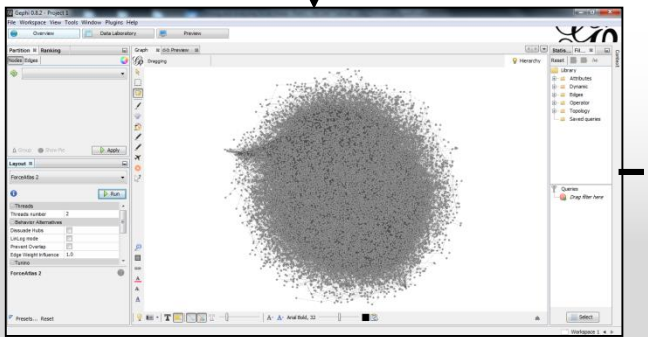


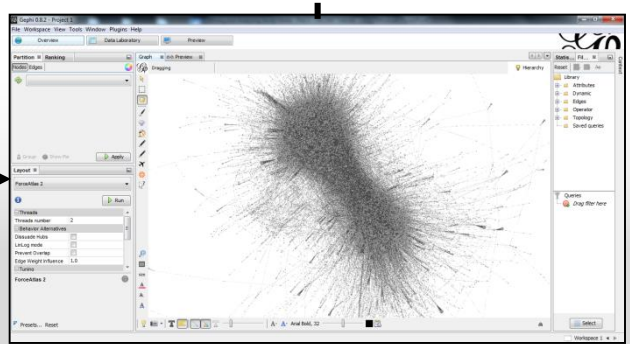




Forced Directed Layout (Force Atlas 2)

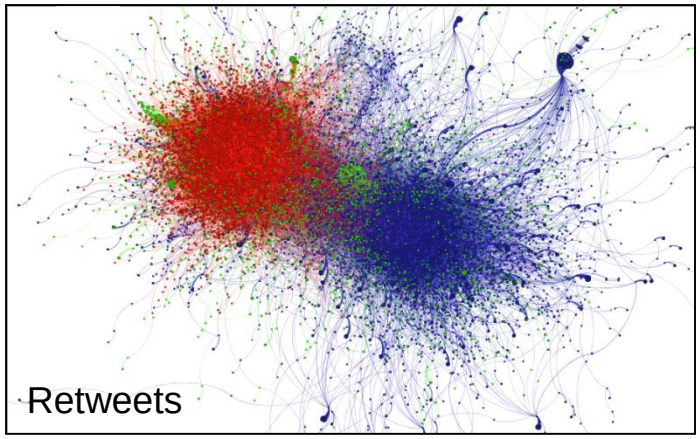
Nodes colored by Cluster (Political Ideology)

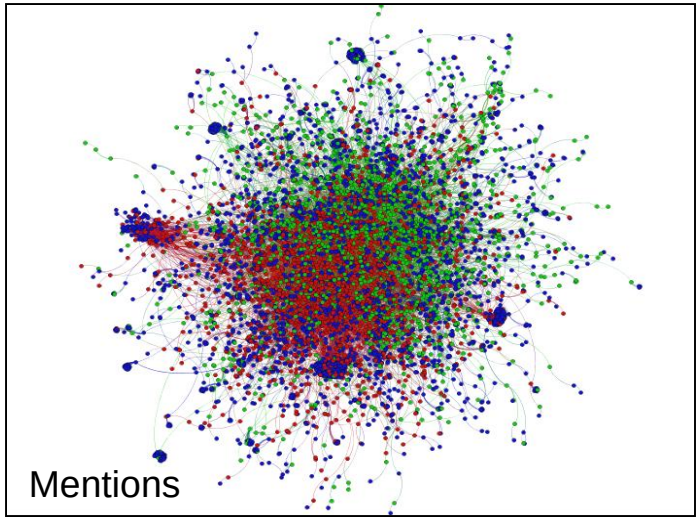




Social Homophily

Contributions of Twitter Study

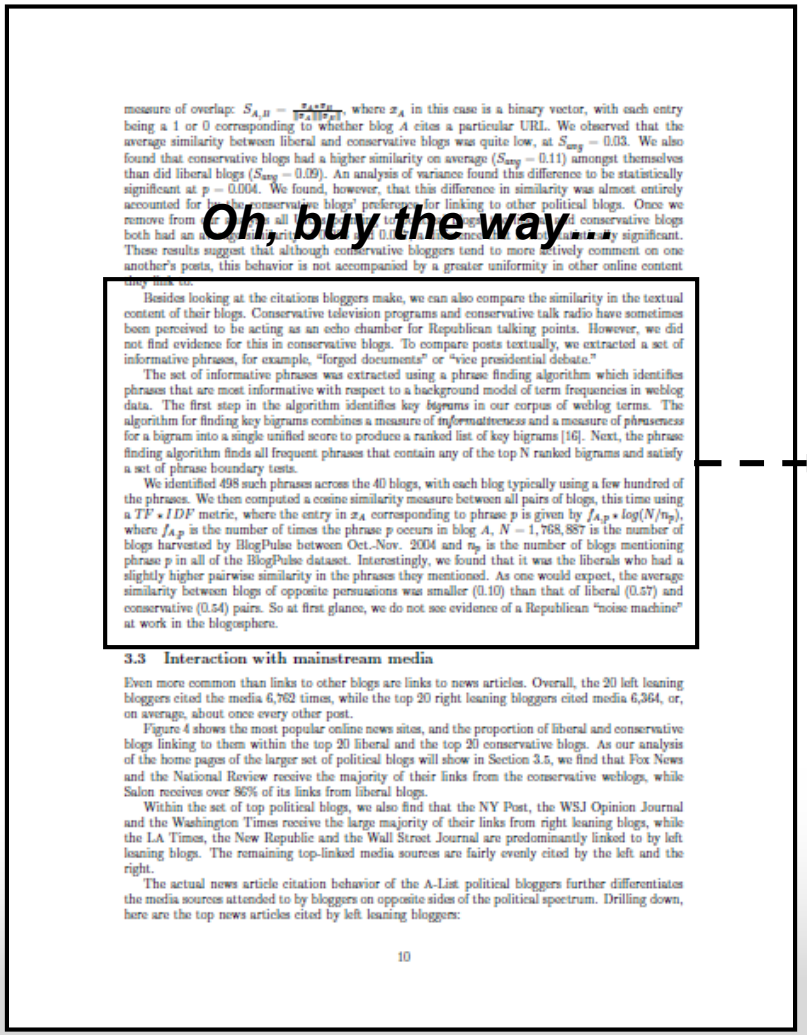




- Cluster analysis of networks derived from corpus shows that the networks of retweets exhibits clear segregation while the mentioned network is dominated by single cluster.
- Manual classification of Twitter user by political alignment, demonstrating that the retweet network clusters corresponding to political left and right. They also show the mention network is political.
- Interpretation of the observed community structures based on injection of partisan content into ideologically opposed hashtag information streams

Social Homophily

Polarization – The other side of the coin



measure of overlap: $S_{A,B} = \frac{x_A \cdot x_B}{\|x_A\| \|x_B\|}$, where x_A in this case is a binary vector, with each entry being a 1 or 0 corresponding to whether blog A cites a particular URL. We observed that the average similarity between liberal and conservative blogs was quite low, at $S_{avg} = 0.03$. We also found that conservative blogs had a higher similarity on average ($S_{avg} = 0.11$) amongst themselves than did liberal blogs ($S_{avg} = 0.09$). An analysis of variance found this difference to be statistically significant at $p = 0.004$. We found, however, that this difference in similarity was almost entirely accounted for by the conservative blogs' preference for linking to other political blogs. Once we remove from [illegible] all links to [illegible] conservative blogs both had an [illegible] significant. These results suggest that although conservative bloggers tend to more actively comment on one another's posts, this behavior is not accompanied by a greater uniformity in other online content they link to.

Oh, buy the way . . .

Besides looking at the citations bloggers make, we can also compare the similarity in the textual content of their blogs. Conservative television programs and conservative talk radio have sometimes been perceived to be acting as an echo chamber for Republican talking points. However, we did not find evidence for this in conservative blogs. To compare posts textually, we extracted a set of informative phrases, for example, "forged documents" or "vice presidential debate."

The set of informative phrases was extracted using a phrase finding algorithm which identifies phrases that are most informative with respect to a background model of term frequencies in weblog data. The first step in the algorithm identifies key *bigrams* in our corpus of weblog terms. The algorithm for finding key bigrams combines a measure of *informativeness* and a measure of *phraseness* for a bigram into a single unified score to produce a ranked list of key bigrams [16]. Next, the phrase finding algorithm finds all frequent phrases that contain any of the top N ranked bigrams and satisfy a set of phrase boundary tests.

We identified 498 such phrases across the 40 blogs, with each blog typically using a few hundred of the phrases. We then computed a cosine similarity measure between all pairs of blogs, this time using a $TF * IDF$ metric, where the entry in x_A corresponding to phrase p is given by $f_{A,p} * log(N/n_p)$, where $f_{A,p}$ is the number of times the phrase p occurs in blog A, $N = 1,768,887$ is the number of blogs harvested by BlogPulse between Oct.-Nov. 2004 and n_p is the number of blogs mentioning phrase p in all of the BlogPulse dataset. Interestingly, we found that it was the liberals who had a slightly higher pairwise similarity in the phrases they mentioned. As one would expect, the average similarity between blogs of opposite persuasions was smaller (0.10) than that of liberal (0.57) and conservative (0.54) pairs. So at first glance, we do not see evidence of a Republican "noise machine" at work in the blogosphere.

3.3 Interaction with mainstream media

Even more common than links to other blogs are links to news articles. Overall, the 20 left leaning bloggers cited the media 6,762 times, while the top 20 right leaning bloggers cited media 6,364, or, on average, about once every other post.

Figure 4 shows the most popular online news sites, and the proportion of liberal and conservative blogs linking to them within the top 20 liberal and the top 20 conservative blogs. As our analysis of the home pages of the larger set of political blogs will show in Section 3.5, we find that Fox News and the National Review receive the majority of their links from the conservative weblogs, while Salon receives over 86% of its links from liberal blogs.

Within the set of top political blogs, we also find that the NY Post, the WSJ Opinion Journal and the Washington Times receive the large majority of their links from right leaning blogs, while the LA Times, the New Republic and the Wall Street Journal are predominantly linked to by left leaning blogs. The remaining top-linked media sources are fairly evenly cited by the left and the right.

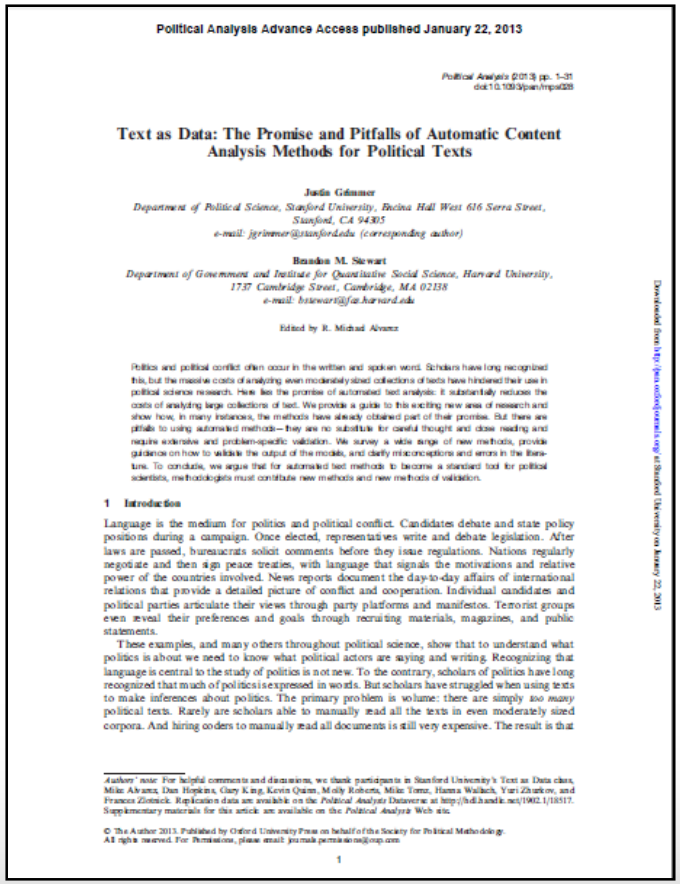
The actual news article citation behavior of the A-List political bloggers further differentiates the media sources attended to by bloggers on opposite sides of the political spectrum. Drilling down, here are the top news articles cited by left leaning bloggers:

10

- In addition to citations, Adamic and Glance also considered the similarity in the textual content of the blogs
- Identified 498 “bigrams” across the Top 40 blogs
- Computed cosine similarity measure between *TD-IDF* scores for all pairs of blogs
- Average similarity between opposites was .10, among conservatives was .54 and liberals was .57.

Social Homophily

Mining Text Content



Political Analysis Advance Access published January 22, 2013

Political Analysis (2013) pp. 1–31
doi:10.1093/pan/mps028

Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts

Justin Grimmer
Department of Political Science, Stanford University, Encina Hall West 616 Serra Street, Stanford, CA 94305
e-mail: jgrimmer@stanford.edu (corresponding author)

Brandon M. Stewart
Department of Government and Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138
e-mail: bstewart@fas.harvard.edu

Edited by R. Michael Alvarez

Politics and political conflict often occur in the written and spoken word. Scholars have long recognized this, but the massive costs of analyzing even moderately sized collections of texts have hindered their use in political science research. Here lies the promise of automated text analysis: it substantially reduces the costs of analyzing large collections of text. We provide a guide to this exciting new area of research and show how, in many instances, the methods have already obtained part of their promise. But there are pitfalls to using automated methods—they are no substitute for careful thought and close reading and require extensive and problem-specific validation. We survey a wide range of new methods, provide guidance on how to validate the output of the models, and clarify misconceptions and errors in the literature. To conclude, we argue that for automated text methods to become a standard tool for political scientists, methodologists must contribute new methods and new methods of validation.

1 Introduction

Language is the medium for politics and political conflict. Candidates debate and state policy positions during a campaign. Once elected, representatives write and debate legislation. After laws are passed, bureaucrats solicit comments before they issue regulations. Nations regularly negotiate and then sign peace treaties, with language that signals the motivations and relative power of the countries involved. News reports document the day-to-day affairs of international relations that provide a detailed picture of conflict and cooperation. Individual candidates and political parties articulate their views through party platforms and manifestos. Terrorist groups even reveal their preferences and goals through recruiting materials, magazines, and public statements.

These examples, and many others throughout political science, show that to understand what politics is about we need to know what political actors are saying and writing. Recognizing that language is central to the study of politics is not new. To the contrary, scholars of politics have long recognized that much of politics is expressed in words. But scholars have struggled when using texts to make inferences about politics. The primary problem is volume: there are simply *too many* political texts. Rarely are scholars able to manually read all the texts in even moderately sized corpora. And hiring coders to manually read all documents is still very expensive. The result is that

Authors' note: For helpful comments and discussions, we thank participants in Stanford University's Text as Data class, Mike Alvarez, Dan Hopkins, Gary King, Kevin Quinn, Molly Roberts, Mike Tomz, Hanna Wallach, Yuri Zhukov, and Frances Zlotnick. Replication data are available on the *Political Analysis* Dataverse at http://hdl.handle.net/1902.1/18517. Supplementary materials for this article are available on the *Political Analysis* Web site.

- "Scholars have long recognized that much of politics is expressed in words."
- "To understand what politics is about we need to know what political actors are saying and writing."
- "…The assumption is that ideology dominates the language used in the text."
- *Ergo, another way to study the polarization of political blogs is to compare what they say.*

Social Homophily

Mining Text Content - Political Blogs Example

Conservative or Liberal?

Republican nightmare begins: Obamacare is 'a godsend' for people getting coverage

U.S. President Barack Obama delivers remarks alongside Human Services Secretary Kathleen Sebelius (R) and other Americans the White House says will benefit from the opening of health insurance marketplaces under the Affordable Care Act, in the Rose Garden.

The health care exchanges—federal and state—are now functioning and not sucking up all the oxygen around the implementation of Obamacare. Finally, the "good news" news stories are finally being told, like this one in The New York Times.

…

That's the kind of story that makes Republicans quake in their shoes. For more of them, go below the fold.

Social Homophily

Mining Text Content – Political Blogs Example

Conservative or Liberal?

The Colorado #obamacare exchange is a disaster.

I'm sorry, but there's no other way to describe these enrollment numbers. Which are, by the way, from a state exchange – meaning that the problems of healthcare.gov should theoretically be irrelevant to the conversation:

Enrollment have grown to 15,074, marketplace officials announced Monday. Marketplace officials set a goal of 136,000 people covered on exchanged-based plans by the end of 2014, but so far the exchange has failed to reach even worst-case enrollment projections.

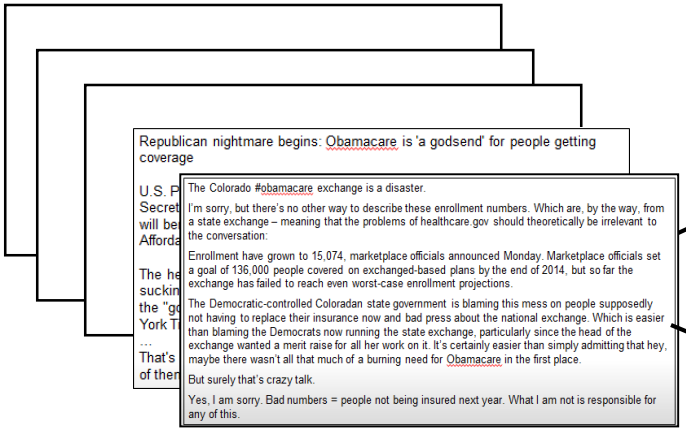
The Democratic-controlled Coloradan state government is blaming this mess on people supposedly not having to replace their insurance now and bad press about the national exchange. Which is easier than blaming the Democrats now running the state exchange, particularly since the head of the exchange wanted a merit raise for all her work on it. It's certainly easier than simply admitting that they, maybe there wasn't all that much of a burning need for Obamacare in the first place.

But surely that's crazy talk.

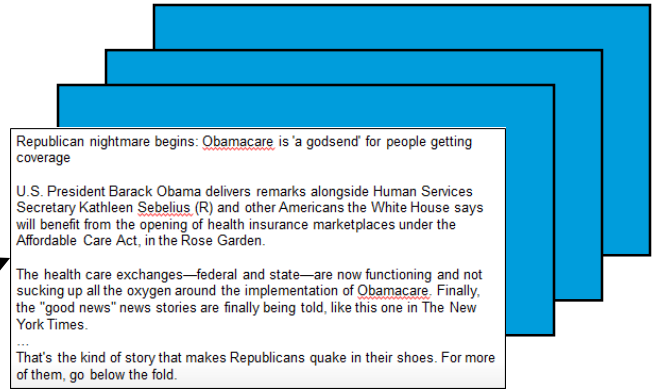
Yes, I am sorry. Bad numbers = people not being insured next year. What I am not is responsible for any of this.

Social Homophily

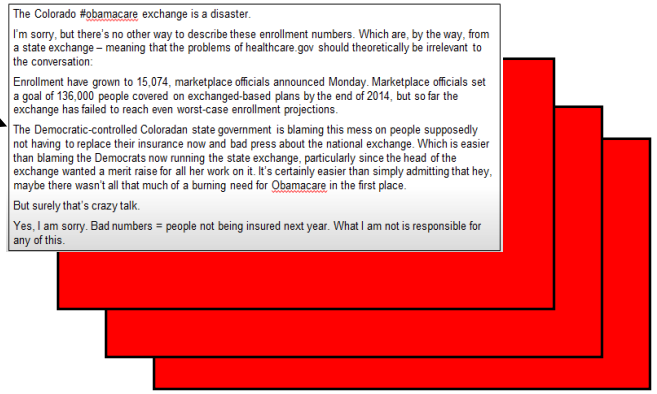
Mining Text Content – Political Blogs Example



Sample of 50 blog entries about "Obamacare" from popular Liberal & Conservative Political Blogs

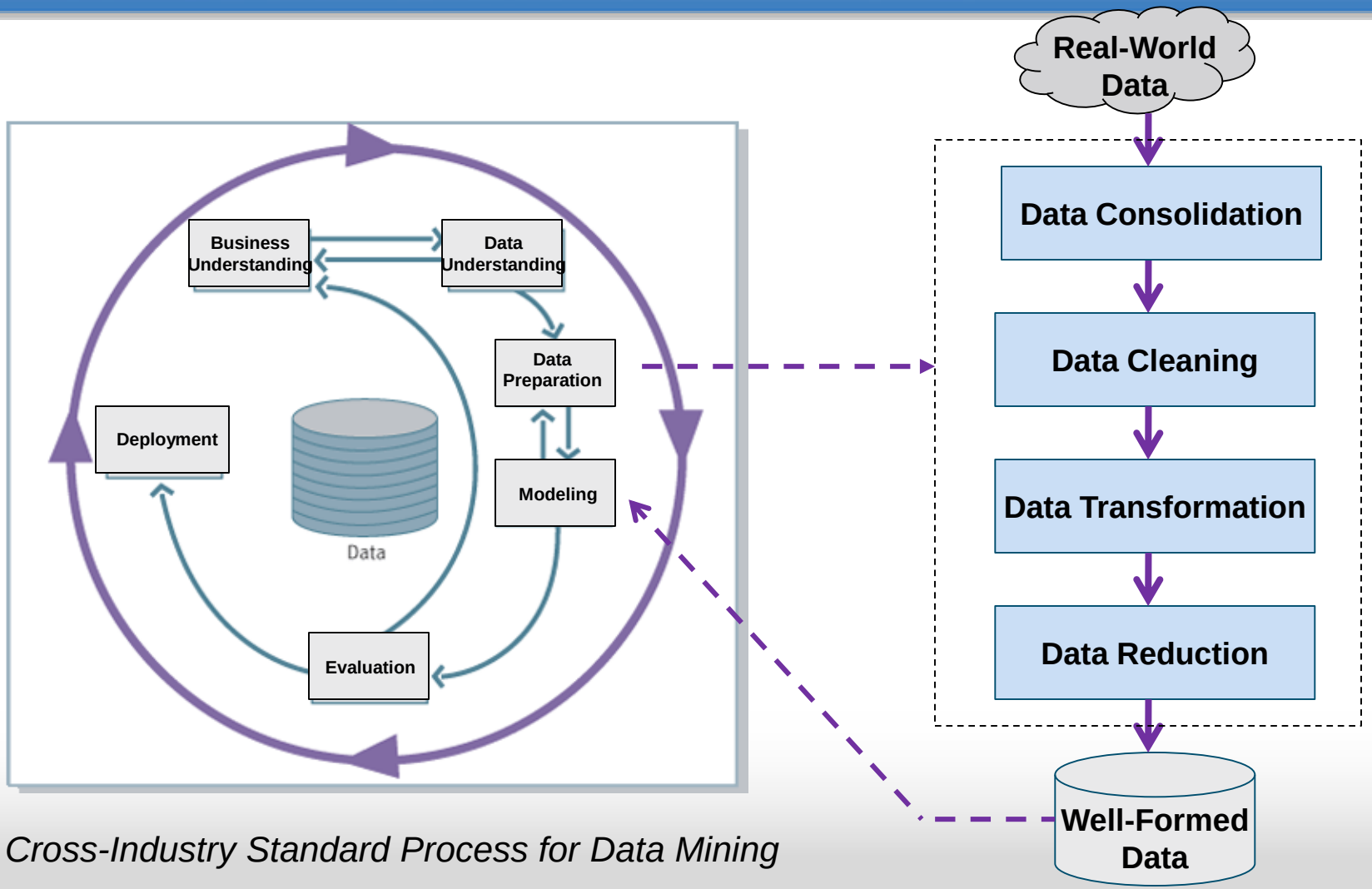


Are they separated into 2 distinct clusters based on their content?



Social Homophily

Mining Text Content vs. Data Mining



Cross-Industry Standard Process for Data Mining

70

Social Homophily

Mining Text Content vs. Data Mining

Some Basic DM Tasks

- Anomaly Detection
- Association Rule Learning
- Clustering
- Classification
- Regression
- Summarization

Social Homophily

Mining Text Content vs. Data Mining

DM Data is usually

- Structured
- Transformed
- Well-formed

Social Homophily

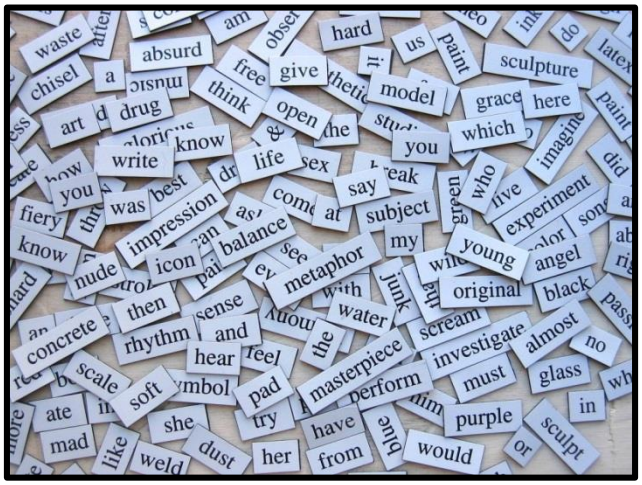
Mining Text Content – The Process

Unstructured Data

- No specified format
- Variable length
- Variable spelling
- Punctuation and non-alphanumeric characters
- Contents are not predefined and no predefined set of values

Social Homophily

Mining Text Content – The Process

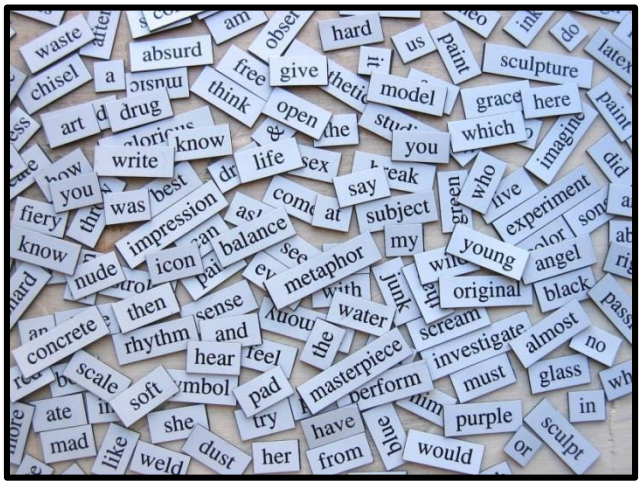


NLP

The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods.

Social Homophily

Mining Text Content – The Process

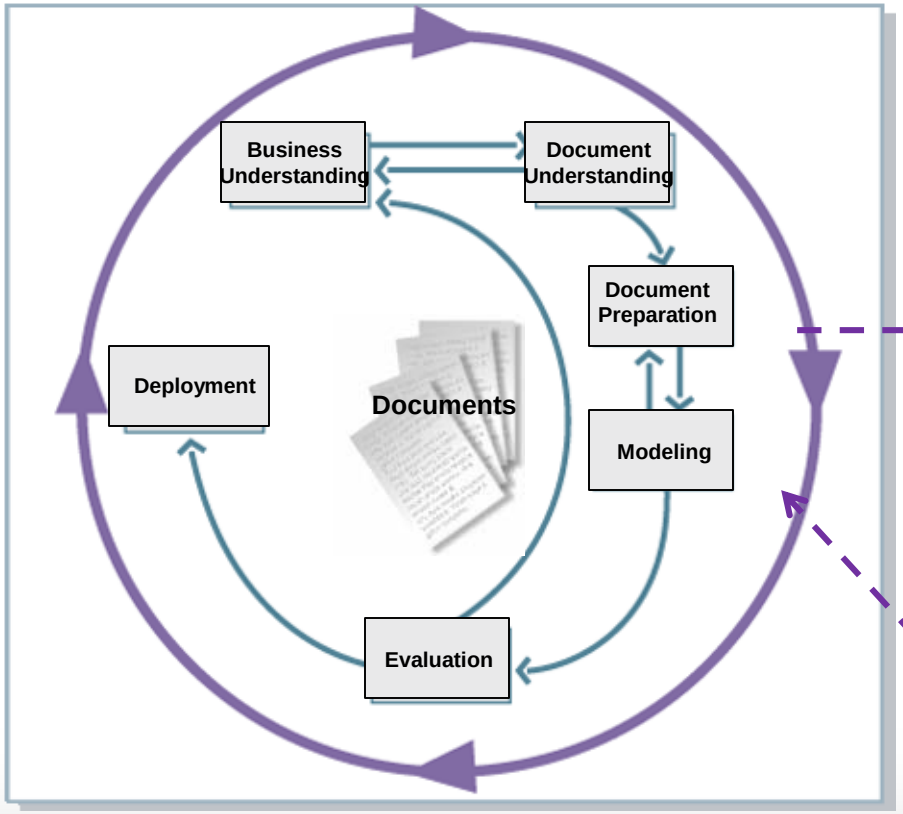


NLP

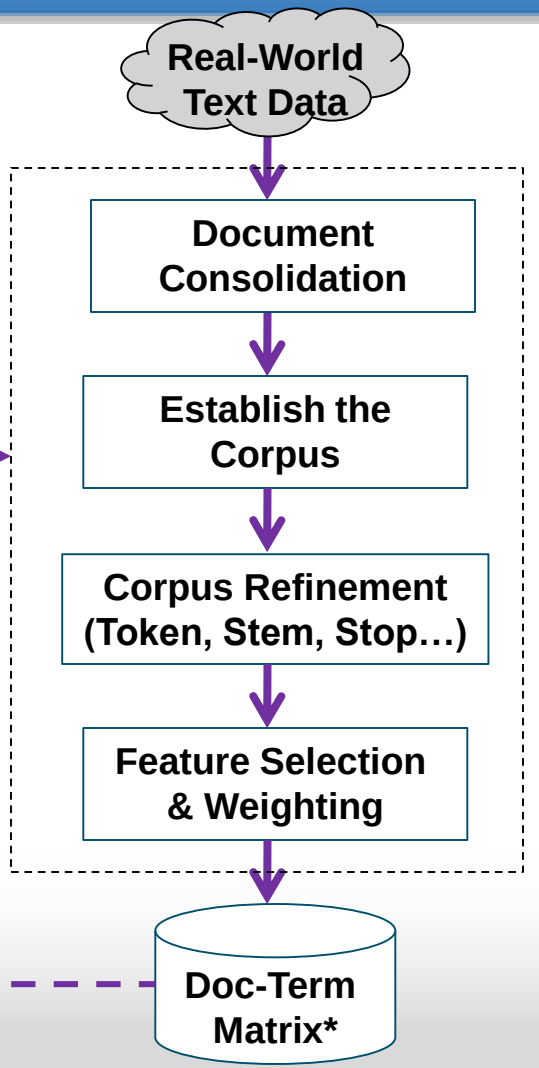
"The most consequential, and shocking, step we will take is to discard the order in which words occur in documents. We will assume documents are a *bag of words*, where order does not inform our analyses… If this assumption is unpalatable, we can retain some word order by including bigrams (word pairs) or trigrams.."

Social Homophily

Mining Text Content – The Process



CRISP-Like Processes



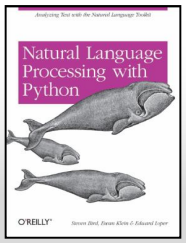
Social Homophily

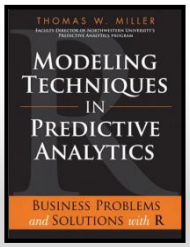
Mining Text Content – The Process

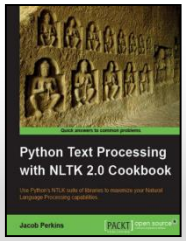
Common representation of tokens within and between documents

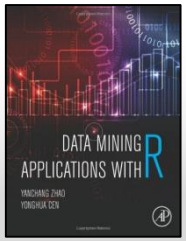
Tokenization → **Normalize** → **Eliminate Stop Words** → **Stemming**

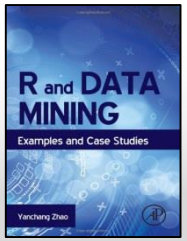
- Tokenization —Parse the text to generate terms. Sophisticated analyzers can also extract phrases from the text.
- Normalize — Convert them to lowercase.
- Eliminate stop words — Eliminate terms that appear very often (e.g. the, and, …).
- Stemming — Convert the terms into their stemmed form—remove plurals and different word forms (e.g. achieve, achieves, achieved – achiev) [note: word about synonyms – WordNet Synset]

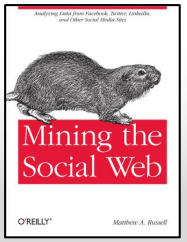






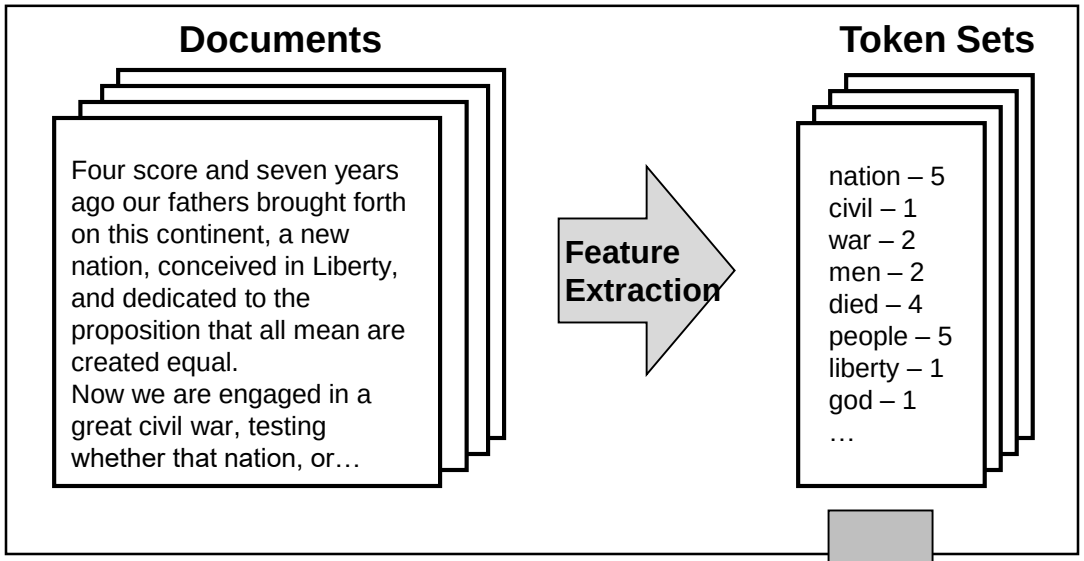






Social Homophily

Mining Text Content – The Process



Feature Extraction & Weighting

"Bag of Words, Terms or Tokens"

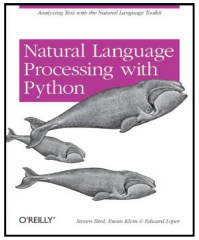
Doc/Token Matrix:
Vectors of Words, Terms or Tokens by Doc

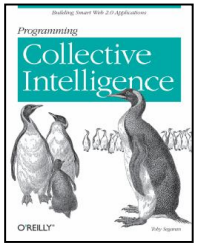
	Token1	Token2	Token3	Token4	…
Doc1	1	2	2	4	
Doc2	4	2	3	0	
Doc3	1	1	1	0	
Doc4	1	1	1	2	
…					

"Bag of Words" (BOW) or Vector Space Model (VSM): Words or Tokens are attributes and documents are examples

Social Homophily

Mining Text Content – Political Blog Example





Analysis from Python NLTK Program

The Colorado #obamacare exchange is a disaster.

I'm sorry, but there's no other way to describe these enrollment numbers. Which are, by the way, from a state exchange – meaning that the problems of healthcare.gov should theoretically be irrelevant to the conversation:

Enrollment have grown to 15,074, marketplace officials announced Monday. Marketplace officials set a goal of 136,000 people covered on exchanged-based plans by the end of 2014, but so far the exchange has failed to reach even worst-case enrollment projections.

The Democratic-controlled Coloradan state government is blaming this mess on people supposedly not having to replace their insurance now and bad press about the national exchange. Which is easier than blaming the Democrats now running the state exchange, particularly since the head of the exchange wanted a merit raise for all her work on it. It's certainly easier than simply admitting that hey, maybe there wasn't all that much of a burning need for Obamacare in the first place.

But surely that's crazy talk.

Yes, I am sorry. Bad numbers = people not being insured next year. What I am not is responsible for any of this.

For each sentence

Tokens: ['The', 'Colorado', '#', 'obamacare', 'exchange', 'is', 'a', 'disaster', '.']
Remove Stopwords: ['Colorado', '#', 'obamacare', 'exchange', 'disaster', '.']
Lower case: ['colorado', '#', 'obamacare', 'exchange', 'disaster', '.']
Alpha only: ['colorado', 'obamacare', 'exchange', 'disaster']
Stems: ['colorado', 'obamacar', 'exchang', 'disast'] ...

Vector of stems for all sentences in blog posting

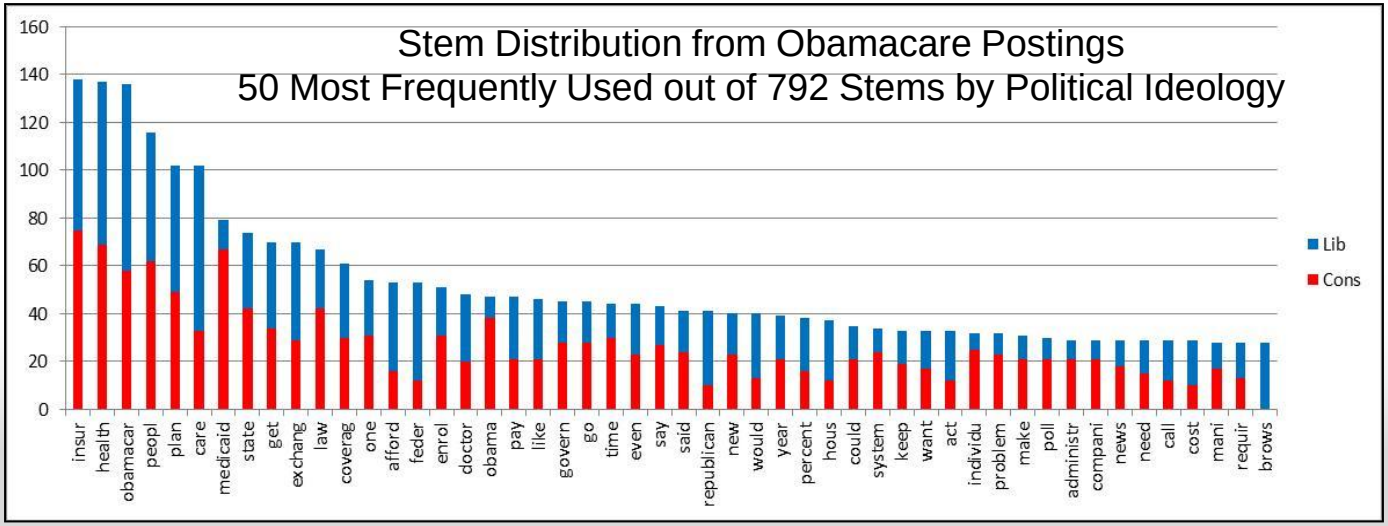
{'one-year': 1, 'offici': 2, 'merit': 1, 'particularli': 1, 'coloradan': 1, 'explain': 2, 'theoret': 1, 'certainli': 1, 'easier': 2, 'mayb': 1, 'far': 1, 'nation': 1, 'disast': 1, 'govern': 1, 'press': 1, 'democrat': 1, 'bad': 2, 'mid-rang': 1, 'mean': 1, 'set': 1, 'replac': 1, 'respons': 1, 'fail': 1, 'even': 1, 'decre': 2, 'state': 3, 'estim': 1, 'democratic-control': 1, 'run': 1, 'obamacar': 2, 'burn': 1, 'blame': 2, 'let': 1, 'enrol': 6, 'worst-cas': 2, 'could': 1, 'unenforc': 1, 'admit': 1, 'place': 1, 'presid': 1, 'opinion': 1, 'first': 1, 'simpli': 1, 'colorado': 2, 'one-tim': 1, 'exchang': 5, 'number': 2, 'cancel': 1, 'announc': 1, 'supposedli': 1, 'describ': 1, 'would': 1, 'pictur': 1, 'hey': 1, 'next': 1, 'much': 1, 'way': 2, 'head': 1, 'offer': 1, 'peopl': 4, 'compani': 1, 'exchanged-bas': 1, 'loom': 1, 'whether': 1, 'work': 2, 'project': 2, 'sinc': 1, 'problem': 1, 'aw': 1, 'site': 1, 'crazi': 1, 'want': 1, 'need': 1, 'grown': 1, 'irrelev': 1, 'goal': 1, 'renew': 1, 'marketplac': 2, 'sure': 1, 'insur': 3, 'monday': 1, 'mess': 1, 'reach': 1, 'rais': 1, 'plan': 1, 'date': 1, 'end': 1, 'mid-novemb': 1, 'scenario': 1, 'cover': 1, 'talk': 1}

Social Homophily

Mining Text Content – Political Blog Example

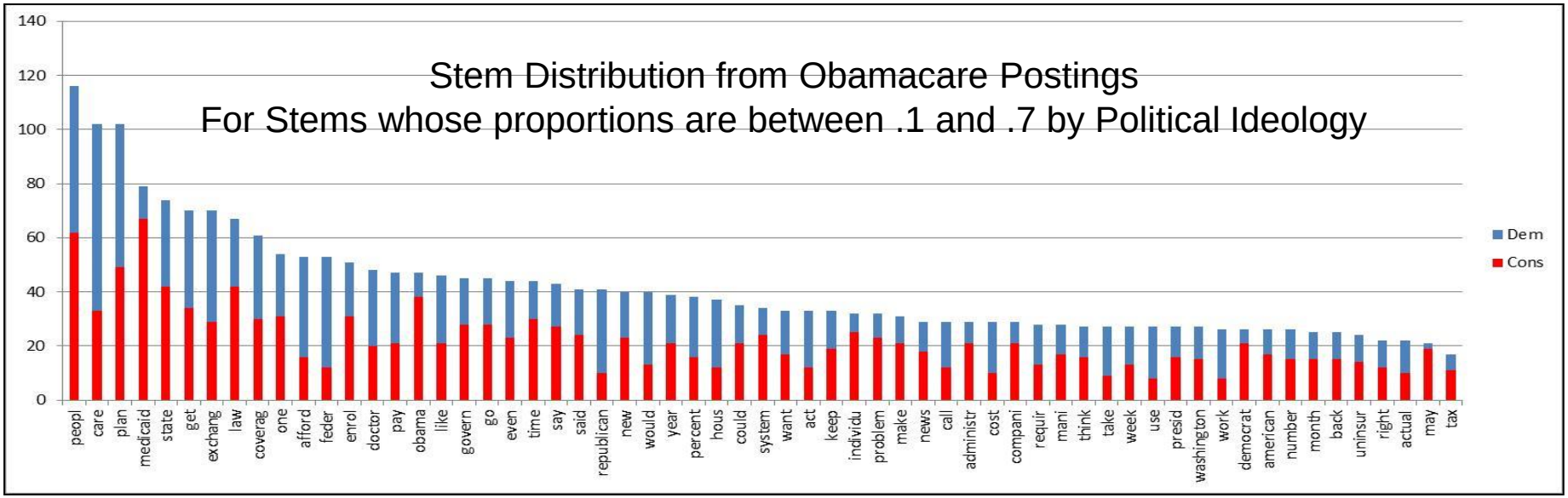
Doc-Stem Matrix (from Python NLTK Program)

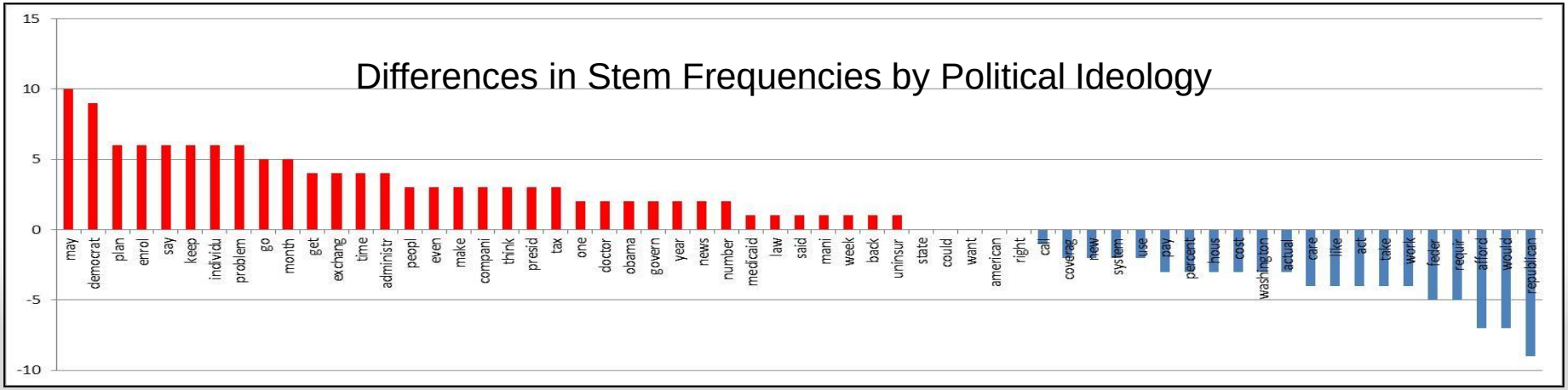
Blog	consider	voter	tweet	worth	digit	physician	rise	govern	wednesday	reimburs	...	detail	futur	er	monday	fulli	anti-obamacar	repeal	rule	emerg	decemb	Total Words
c1-1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0	54
c1-2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	4	1	0	0	0	57
c1-3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	1	1	0	0	181
c1-4	0	0	0	0	0	0	0	5	0	0	...	0	0	0	0	1	0	0	1	0	0	244
c1-5	0	0	0	0	0	2	0	1	0	3	...	1	0	0	0	0	0	0	0	0	0	113
c2-6	0	0	0	0	0	0	1	1	0	0	...	0	0	0	0	0	0	0	0	0	0	83
c2-7	0	0	0	0	0	0	0	3	0	0	...	0	0	0	0	0	0	0	0	0	0	74
c2-8	0	0	0	0	0	0	0	0	0	0	...	0	0	0	2	0	0	0	0	0	2	136
c2-9	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	1	0	0	0	190
c2-10	0	0	0	0	0	0	0	1	0	0	...	0	0	0	0	0	0	0	0	0	2	160
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
L4-16	0	0	0	0	0	0	1	1	0	0	...	0	0	0	0	0	0	0	2	0	0	184
L4-17	0	0	0	0	0	0	0	3	2	0	...	0	0	0	0	0	1	0	0	0	0	151
L4-18	0	0	0	0	0	0	0	1	0	0	...	0	0	0	0	0	0	0	0	0	3	333
L4-19	0	0	0	0	0	0	0	0	0	0	...	1	0	0	0	0	0	0	0	0	0	182
L4-20	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	0	269
L5-21	0	0	2	0	0	0	0	0	0	0	...	0	0	0	0	0	0	1	0	0	0	83
L5-22	0	1	0	0	0	0	0	1	0	0	...	0	0	0	0	0	0	4	0	0	0	133
L5-23	0	0	0	0	0	0	0	4	0	0	...	0	0	0	1	0	0	1	0	0	0	131
L5-24	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	1	0	1	0	0	0	226
L5-25	2	0	0	3	0	0	0	0	0	0	...	0	0	0	1	0	1	0	0	0	0	146
Total	2	5	3	4	2	7	4	45	3	3	...	7	4	4	7	4	7	18	8	6	10	



Social Homophily

Mining Text Content – Political Blog Example





Social Homophily

Mining Text Content – The Process

Transforming Frequencies and Weights

- Binary Frequencies: tf =1 for tf>0; otherwise 0
- Term Frequencies: tf(i,j)/Sum of tf(i,j) in Doc K
- Log Frequencies: 1 + log(tf) for tf>0; otherwise 0
- Normalized Frequencies: Divide each frequency by SQRT of Sum of Squares of the frequencies within the vector (column)
- Term Frequency–Inverse Document Frequency
  - TF: Freq of term for given doc/ sum of words for given doc
  - Inverse Document Frequency: log(N/(1+D)) where N is total number of docs and D is number with term
  - TF * IDF

Social Homophily

Mining Text Content – The Process

Stem Frequencies	govern	even	new	enrol	would	Total
c1-1	0	0	0	2	0	2
c1-2	0	0	0	1	0	1
c1-3	0	2	1	1	0	4
c1-4	5	1	0	0	1	7
L5-23	4	2	0	3	2	11
L5-24	0	2	2	0	0	4
L5-25	0	1	0	0	2	3
#M	2	5	2	4	3	
#D	10	10	10	10	10	
IDF=1+LN(#D/#M)	2.61	1.69	2.61	1.92	2.20	

Term-Frequencies	govern	even	new	enrol	would
c1-1	0.00	0.00	0.00	1.00	0.00
c1-2	0.00	0.00	0.00	1.00	0.00
c1-3	0.00	0.50	0.25	0.25	0.00
c1-4	0.71	0.14	0.00	0.00	0.14
L5-23	0.36	0.18	0.00	0.27	0.18
L5-24	0.00	0.50	0.50	0.00	0.00
L5-25	0.00	0.33	0.00	0.00	0.67

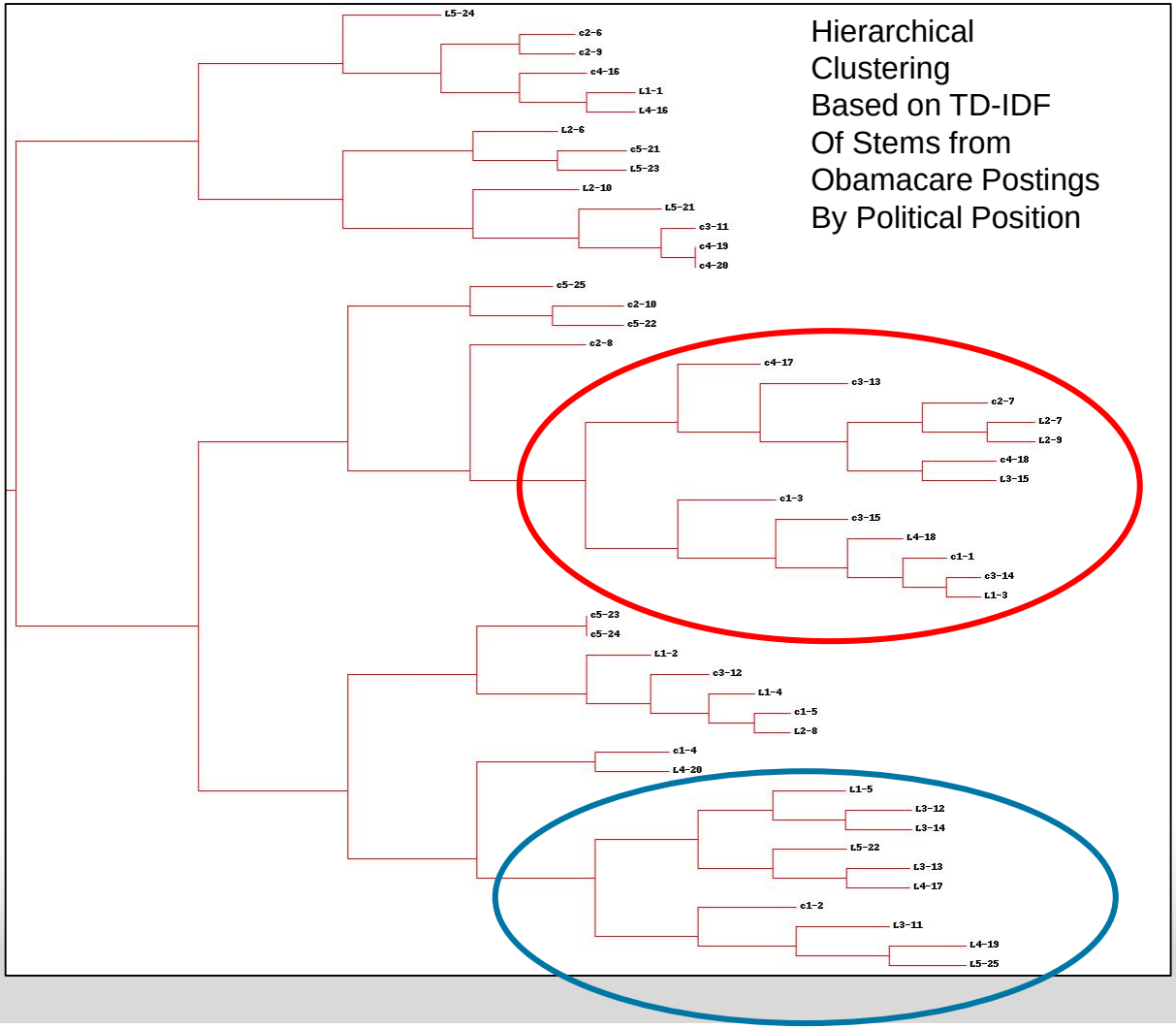
TF*IDF	govern	even	new	enrol	would
c1-1	0.00	0.00	0.00	1.92	0.00
c1-2	0.00	0.00	0.00	1.92	0.00
c1-3	0.00	0.85	0.65	0.48	0.00
c1-4	1.86	0.24	0.00	0.00	0.31
L5-23	0.95	0.31	0.00	0.52	0.40
L5-24	0.00	0.85	1.30	0.00	0.00
L5-25	0.00	0.56	0.00	0.00	1.47

Social Homophily

Mining Text Content – Political Blog Example

Average
Cosine Similarities

Cons = .25
Lib = .27
Cons-Lib = .26



Hierarchical
Clustering
Based on TD-IDF
Of Stems from
Obamacare Postings
By Political Position

Social Homophily

Mining Text Content – Retweets and Mentions

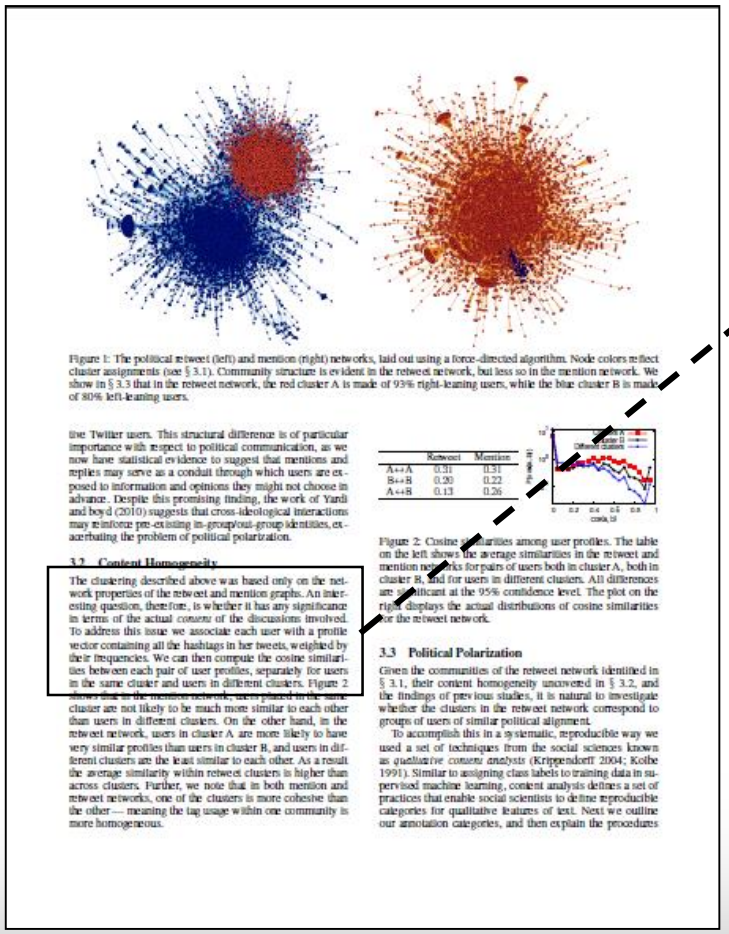


Figure 1: The political retweet (left) and mention (right) networks, laid out using a force-directed algorithm. Node colors reflect cluster assignments (see § 3.1). Community structure is evident in the retweet network, but less so in the mention network. We show in § 3.3 that in the retweet network, the red cluster A is made of 93% right-leaning users, while the blue cluster B is made of 80% left-leaning users.

the Twitter users. This structural difference is of particular importance with respect to political communication, as we now have statistical evidence to suggest that mentions and replies may serve as a conduit through which users are exposed to information and opinions they might not choose in advance. Despite this promising finding, the work of Yardi and boyd (2010) suggests that cross-ideological interactions may reinforce pre-existing in-group/out-group identities, exacerbating the problem of political polarization.

3.2 Content Homogeneity

The clustering described above was based only on the network properties of the retweet and mention graphs. An interesting question, therefore, is whether it has any significance in terms of the actual *content* of the discussions involved. To address this issue we associate each user with a profile vector containing all the hashtags in her tweets, weighted by their frequencies. We can then compute the cosine similarities between each pair of user profiles, separately for users in the same cluster and users in different clusters. Figure 2 shows that in the mention network, users placed in the same cluster are not likely to be much more similar to each other than users in different clusters. On the other hand, in the retweet network, users in cluster A are more likely to have very similar profiles than users in cluster B, and users in different clusters are the least similar to each other. As a result the average similarity within retweet clusters is higher than across clusters. Further, we note that in both mention and retweet networks, one of the clusters is more cohesive than the other — meaning the tag usage within one community is more homogeneous.

	Retweet	Mention
A↔A	0.31	0.31
B↔B	0.20	0.22
A↔B	0.13	0.26

Figure 2: Cosine similarities among user profiles. The table on the left shows the average similarities in the retweet and mention networks for pairs of users both in cluster A, both in cluster B, and for users in different clusters. All differences are significant at the 95% confidence level. The plot on the right displays the actual distributions of cosine similarities for the retweet network.

3.3 Political Polarization

Given the communities of the retweet network identified in § 3.1, their content homogeneity uncovered in § 3.2, and the findings of previous studies, it is natural to investigate whether the clusters in the retweet network correspond to groups of users of similar political alignment.

To accomplish this in a systematic, reproducible way we used a set of techniques from the social sciences known as *qualitative content analysis* (Krippendorff 2004; Kolbe 1991). Similar to assigning class labels to training data in supervised machine learning, content analysis defines a set of practices that enable social scientists to define reproducible categories for qualitative features of text. Next we outline our annotation categories, and then explain the procedures

Content Homogenity

- An interesting question, therefore, is whether it has any significance in terms of the actual content of the discussions involved.
- To address this issue we associate each user with a profile vector containing all the hashtags in her tweets, weighted by their frequencies.
- We can then compute the cosine similarities between each pair of user profiles, separately for users in the same cluster and users in different clusters.

Social Homophily

Mining Text Content – Retweets and Mentions

Troll of Arc @TurdBandito 10m

I hope #**democrats** realize their party is right now having an ironically even bigger hand in the destruction of freedom than the #GOP did

Expand Reply Retweet Favorite More

Tweets	#tcot	#p2	#gop	#teaparty	#hcr	#democrats	#cookiemonster	...
User1	1	0	1	1	0	0	0	...
User2	0	1	0	0	0	1	0	...
User3	1	1	0	0	1	1	1	...
User4	0	0	1	0	0	1	0	...
...	...	...	...	...	...	...	...	...

Cosine Similarities

Average Similarities

	Retweet	Mention
A↔A	0.31	0.31
B↔B	0.20	0.22
A↔B	0.13	0.26