Freie Universität Berlin

Class 1 - Introduction into course concept and organization

Course: Computational Network Analysis

Prof. Dr. Cl. Müller-Birn

Human-Centered Computing

Feb 22, 2016

Today's outline

- Network Science meets Web Science
- Application examples for analyzing networks
- About the course

Network Science

History of Network Science

- Jacob Moreno introduced the ideas and tools of sociometry in 1933
  - Sociometry: The study of positive and negative affective relations
  - Sociograms = diagrams of human interaction

Case study: attraction network

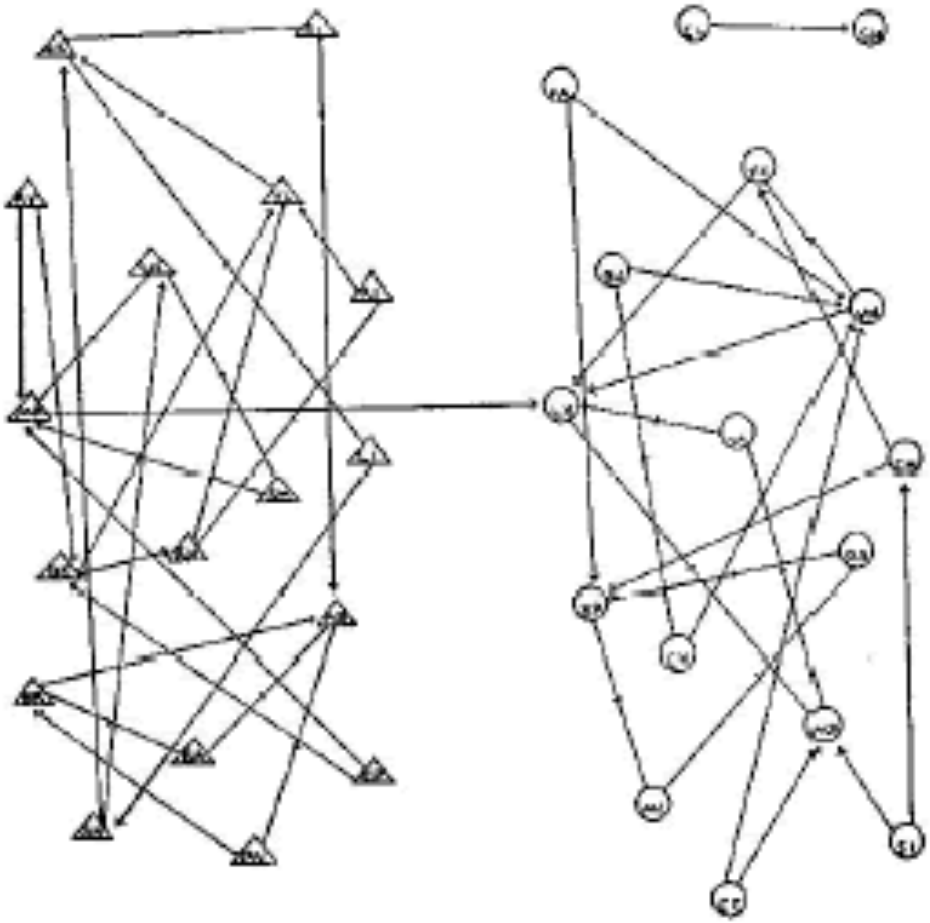


Figure 1. An Attraction Network in a Fourth Grade Class (from Moreno [19], p. 38).

History of Network Science

- Jacob Moreno introduced the ideas and tools of sociometry in 1933
  - Sociometry: The study of positive and negative affective relations
  - Sociograms = diagrams of human interaction

- In the 1950s, Alex Bavelas founded the Group Networks Laboratory at M.I.T to study the effectiveness of different communication patterns on patterns in helping small groups of people solve common tasks.

- A social network is *„a specific set of linkages among a defined set of persons, with the additional property that the characteristics of these linkages as a whole may be used to interpret the social behavior of these persons involved“* [Mitchell 1969, S.2]

- Milgram (1967): Small World Experiments.

From Network Analysis to Network Science

- Network Analysis is a set of relational methods for systematically understanding and identifying connections among actors.

- Network Analysis
  - is motivated by a structural intuition based on ties linking social actors
  - is grounded in systematic empirical data
  - draws heavily on graphic imagery
  - relies on the use of mathematical and/or computational models.

- Network Science embodies a range of theories relating types of observable "social" spaces and their relation to individual and group behavior.

Noshir Contractor: Web Science meets Network Science

Web Science

Web Science

- Web Science is the processing the information available on the web in similar terms to those applied to natural environment. (Shneiderman, 2007)

Computer Science	Web Science
Metrics	
Moore's Law	Page views
Order (n) algorithm analysis	Unique visitors/month
Gigabytes	Number of songs or videos
Topics	
Computer networks	Social networks
Packet switching	Voice over IP, music sharing
Information	Relationships
Programming languages	Wikis, blogs, tagging
Databases, operating systems, compilers	E-commerce, e-learning, e-government, medical informatics, financial analysis
3D graphics, rendering algorithms, computational geometry, object modeling	Creating and sharing video, animation, music, photos, maps
Focus	
Technology	Applications
Computers	Users
Supercomputers	Mobile devices
Proficient programmers	Universal usability

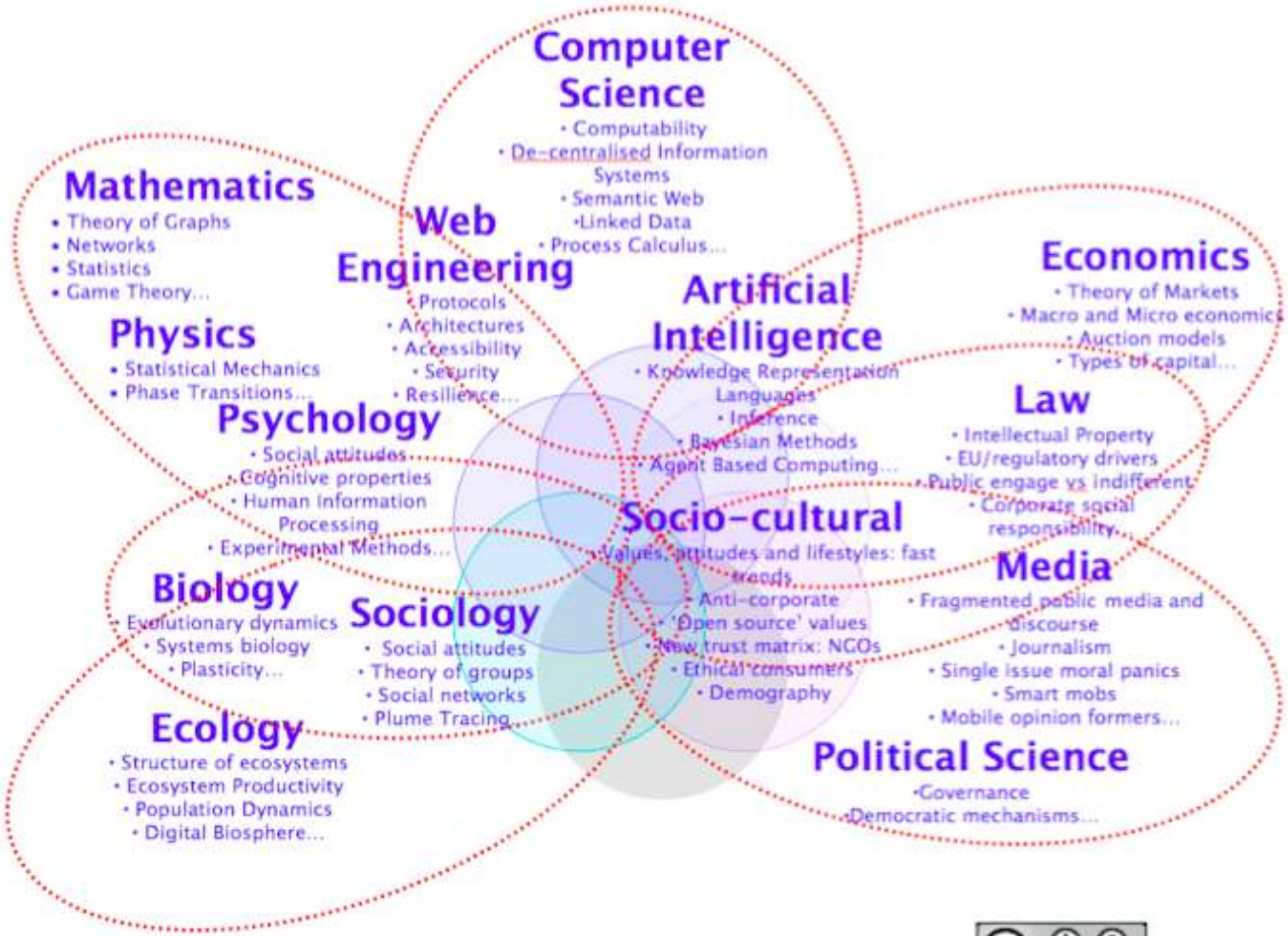
Ben Shneiderman. (2007). Web Science: A Provocative Invitation to Computer Science. Communication of the ACM, 50(6), 25–27.

Web Science

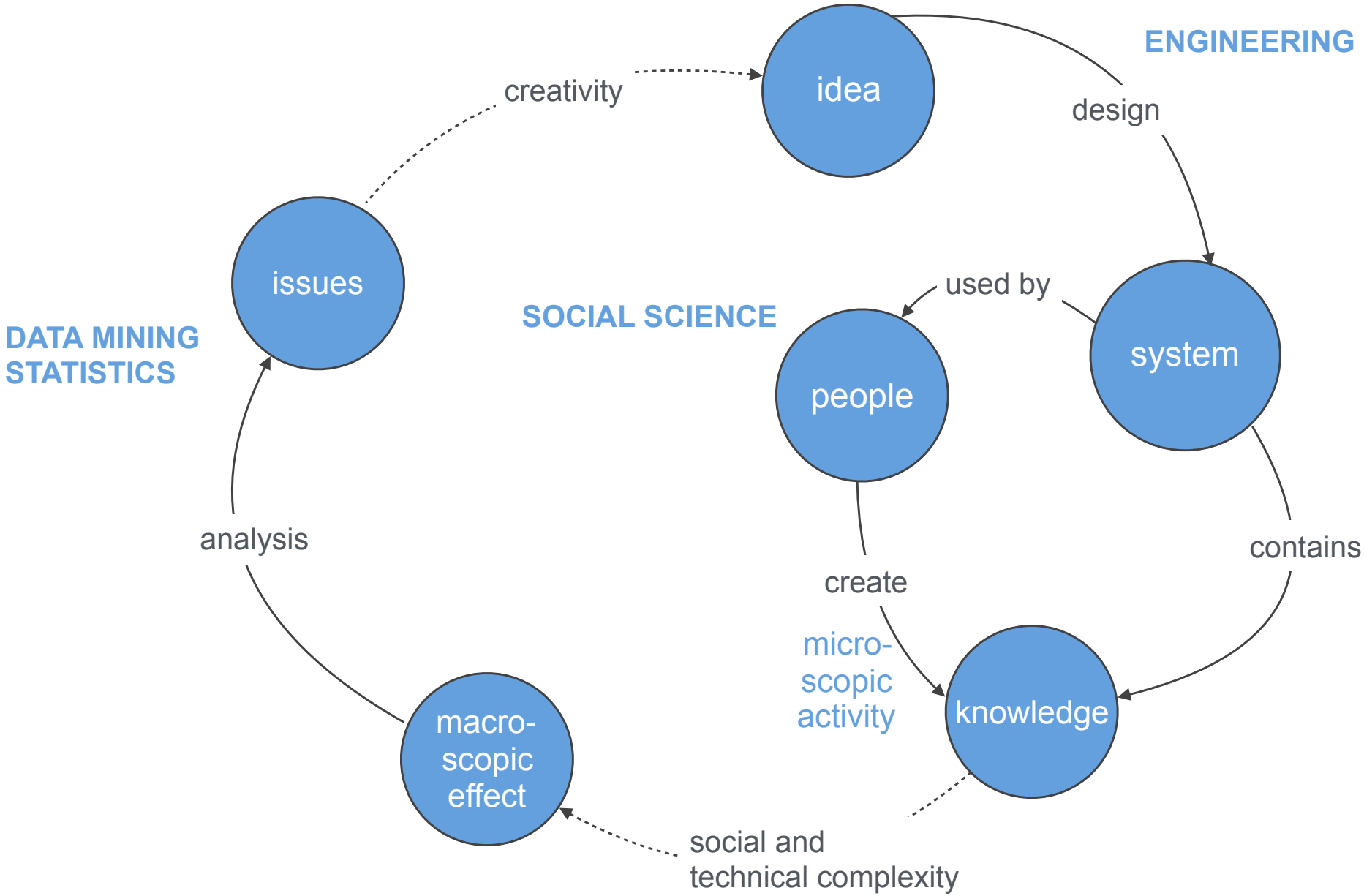
- Web Science is the processing the information available on the web in similar terms to those applied to natural environment. (Shneiderman, 2007)

- Web Science is a socio-technical science that investigates how the Web evolves given the regulations, technology and content imposed, engineered and contributed, respectively, as an effect of human behavior and how the Web affects human behavior. (Staab, 2012)

Web Science: Components

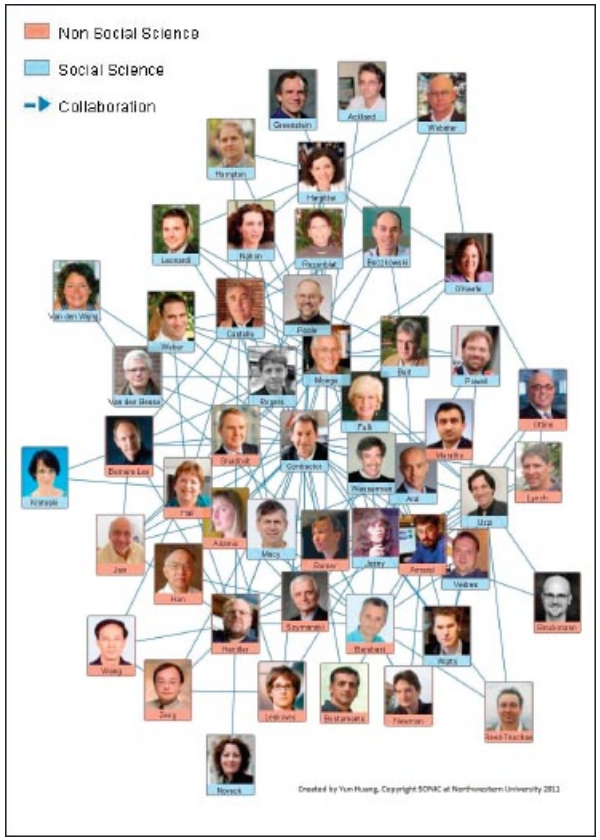


The (adapted) process of Web Science



Web Science meets Network Science

Wright, Alex. "Web science meets network science." Communications of the ACM 54.5 (2011): 23-23.



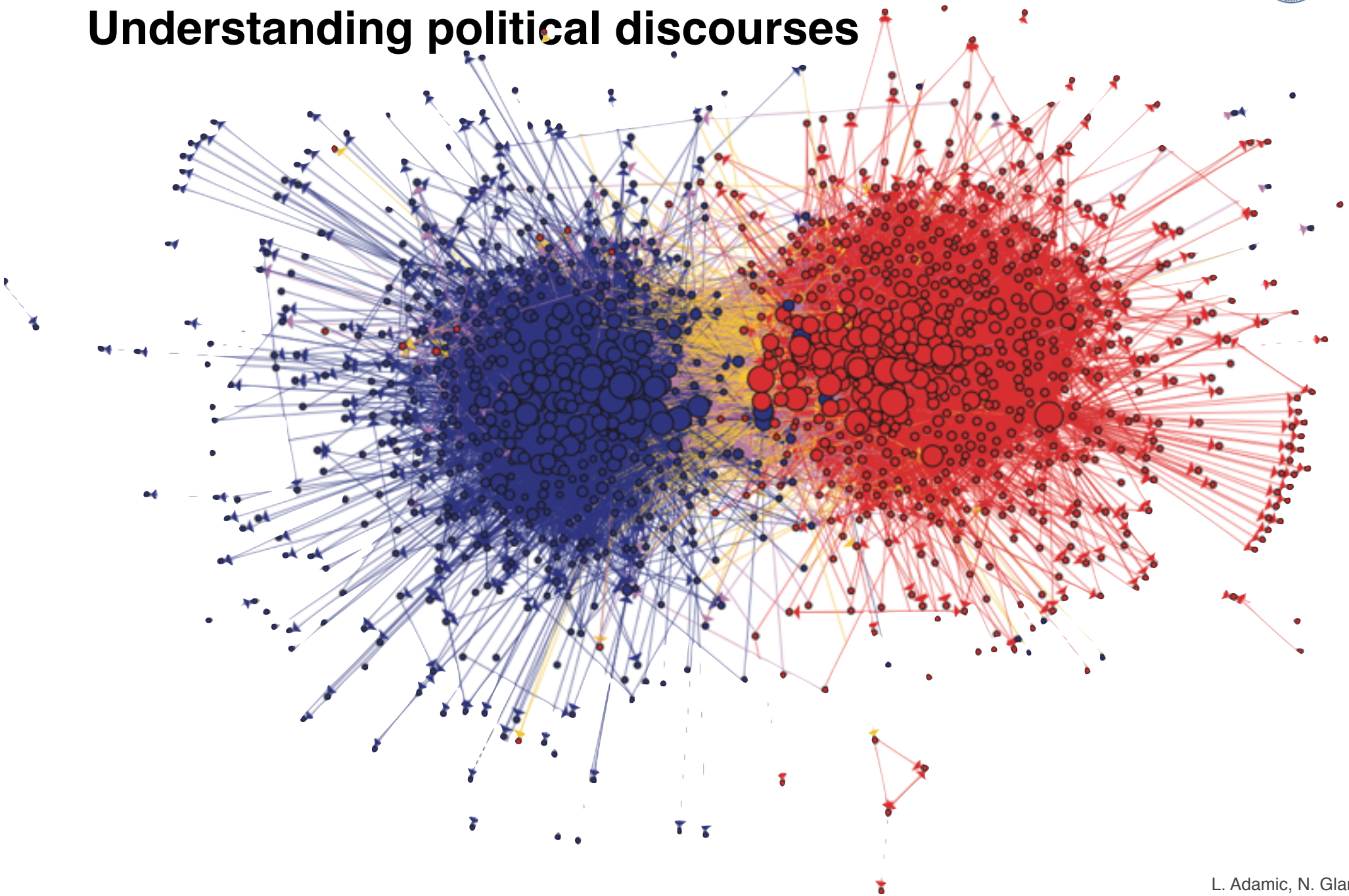
Where do these two emerging fields overlap?

- "Pure" scientific approach of Network Science vs. the more applied, engineering-oriented tactics of Web Science
- Web science is focused on how we could do things better, while network science is more focused on how things work.
- In this course, we focus more on the former but the latter matters too.

Wright, A. (2011). Web science meets network science. Commun. ACM, 54, 23–23. http://doi.org/http://doi.acm.org/10.1145/1941487.1941497

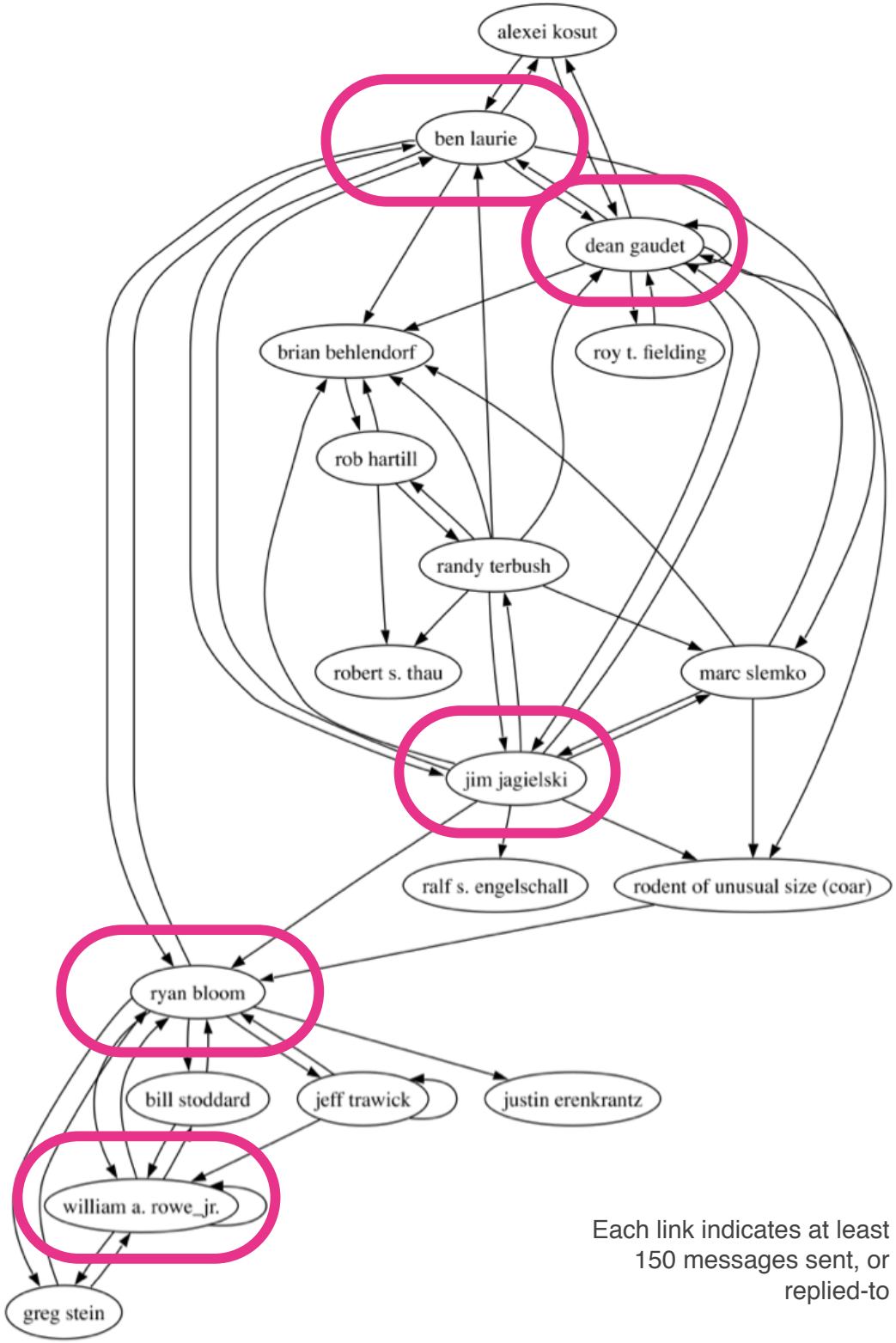
Application examples

Understanding political discourses



L. Adamic, N. Glance:
The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. 2005.

Understanding participation in OSS projects



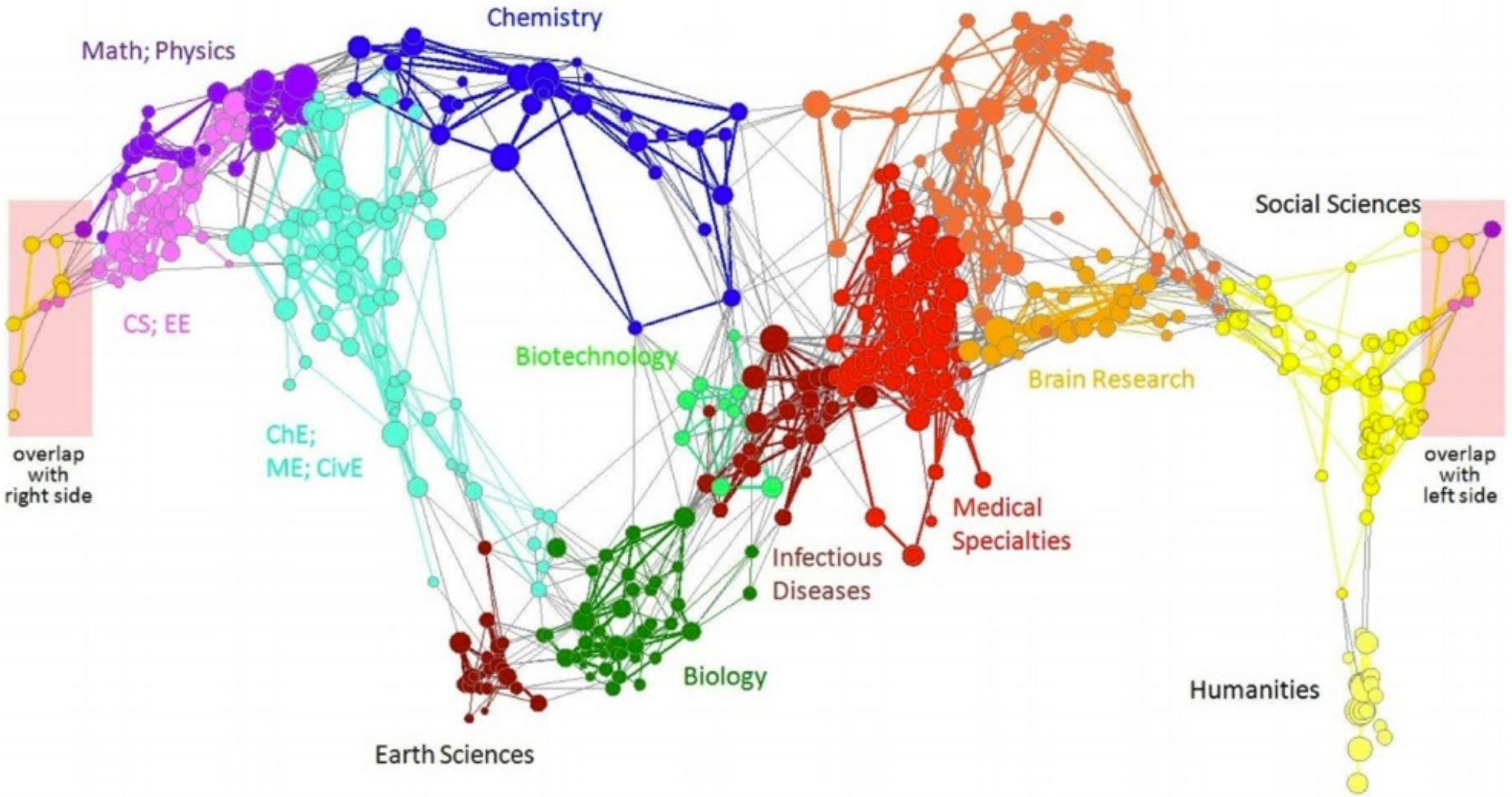
Each link indicates at least 150 messages sent, or replied-to

Christian Bird, Alex Gourley, Prem Devanbu, Michael Gertz, and Anand Swaminathan. 2006. Mining email social networks. In Proceedings of the 2006 international workshop on Mining software repositories (MSR '06). ACM, New York, NY, USA, 137-143.

Understanding collaboration and coordination patterns

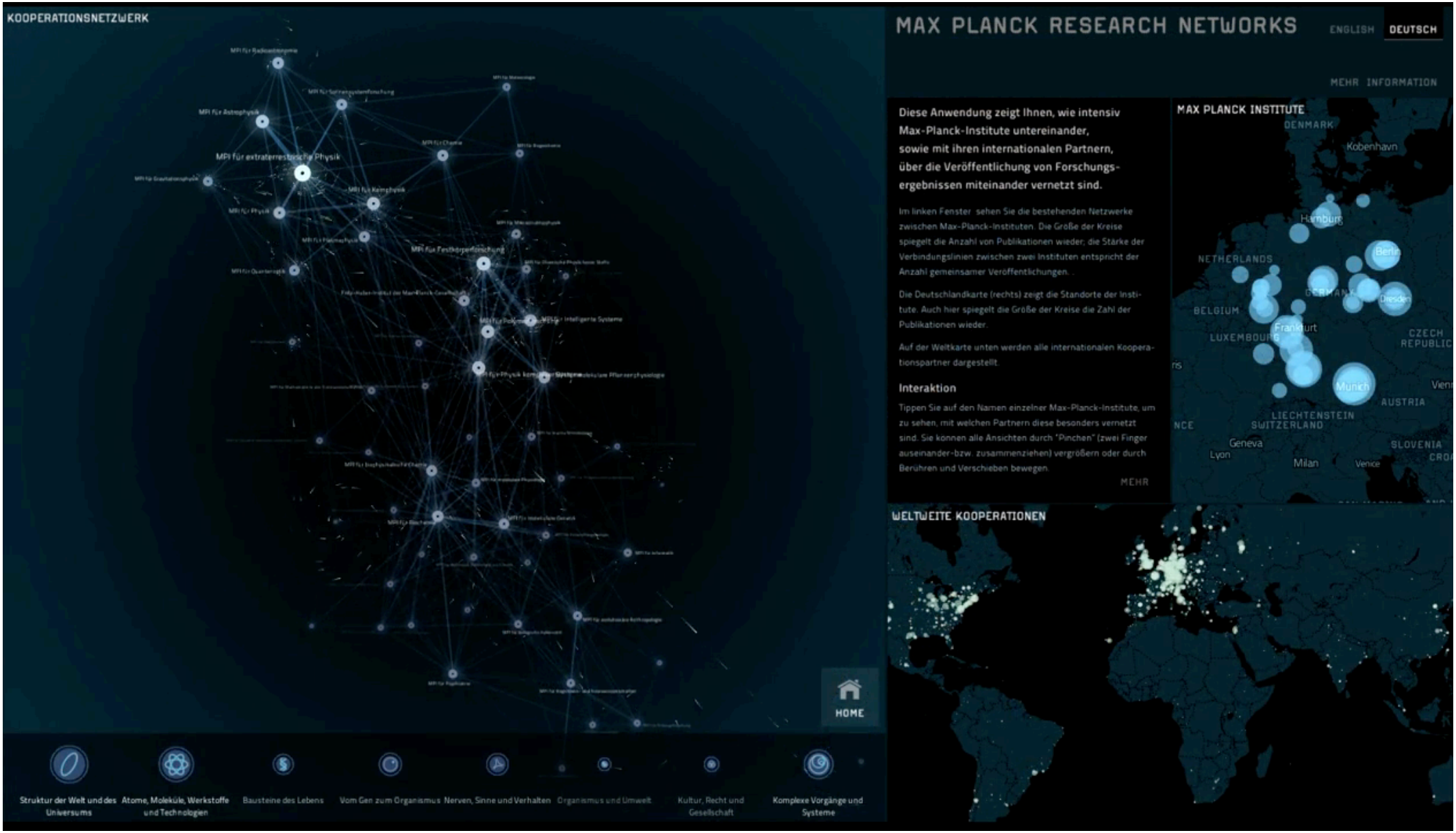
Uncloaking Terrorist Networks by Valdis E. Krebs
First Monday, Volume 7 Number 4 - 1 April 2002
http://firstmonday.org/ojs/index.php/fm/article/view/941/863

Understanding research topics

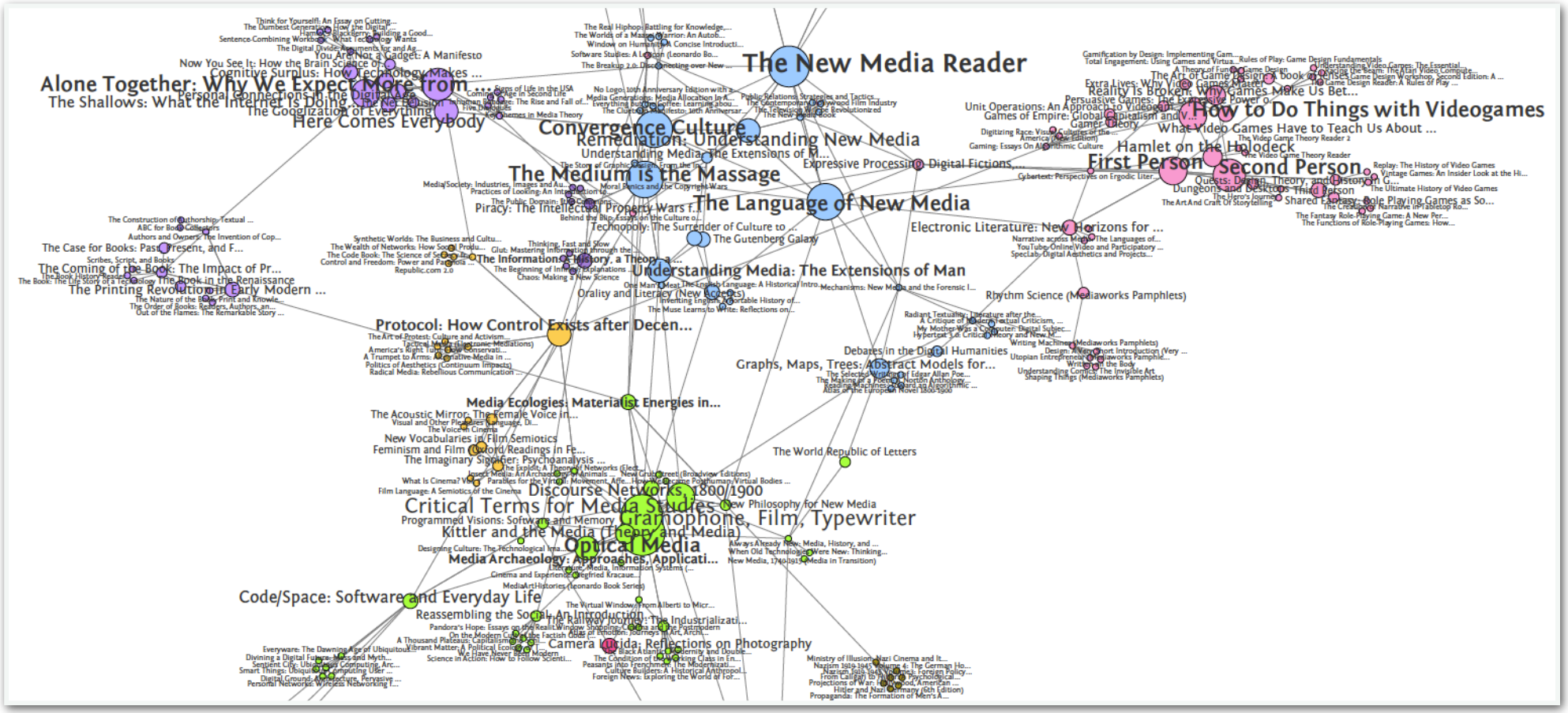


Börner K, Klavans R, Patek M, Zoss AM, et al. (2012) Design and Update of a Classification System: The UCSD Map of Science. PLoS ONE 7(7): e39464. doi:10.1371/journal.pone.0039464

Showing scientific collaboration

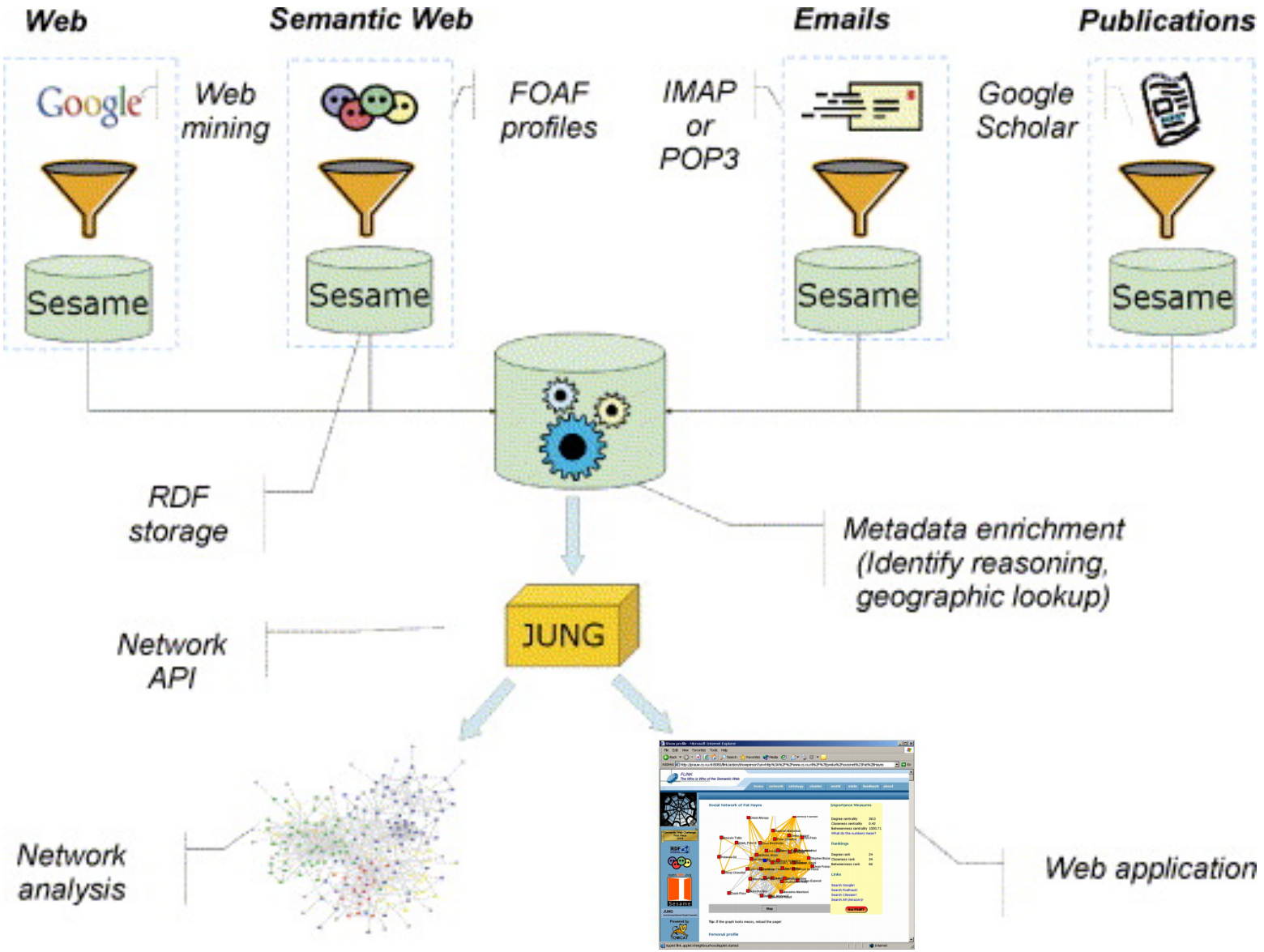


Exploring recommendations on Amazon



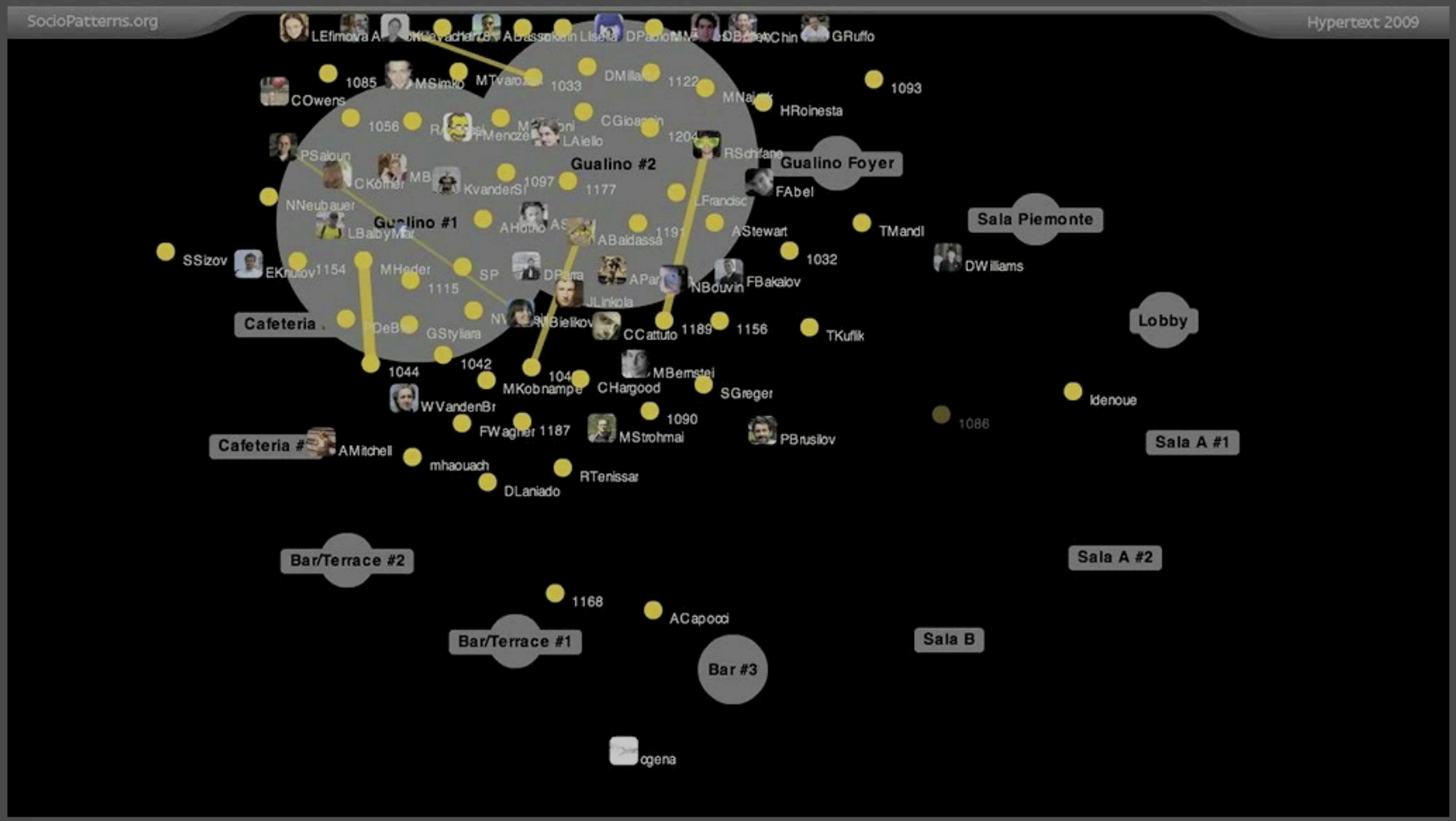
Data by Anne Helmond
Code by http://christopherwarnow.com/portfolio/?p=278

Extending social networks by semantic data



Mika, P. 2005, 'Flink: Semantic Web technology for the extraction and analysis of social networks', Web Semantics: Science, Services and Agents on the World Wide Web, 3, 2–3, 211--223.

Live Social Semantics…



Social dynamics in conferences collected by wearable sensors

Freie Universität Berlin

Contact characteristics	ESWC 2009	HT 2009	ESWC 2010
Number of contact events	16258	9875	14671
Average contact length (s)	46	42	42
Fraction of contacts $\leq$ 1 mn	0.87	0.89	0.88
Fraction of contacts $\leq$ 2 mn	0.94	0.96	0.95
Fraction of contacts $\leq$ 5 mn	0.99	0.99	0.99
Fraction of contacts $\leq$ 10 mn	0.998	0.998	0.998

Network characteristics	ESWC 2009	HT 2009	ESWC 2010
Number of users	175	113	158
Average degree	54	39	55
Average strength (s)	8590	7374	7807
Average edge weight (s)	159	189	141
Fraction of weights $\leq$ 1 mn	0.7	0.67	0.74
Fraction of weights $\leq$ 5 mn	0.9	0.89	0.93
Fraction of weights $\leq$ 10 mn	0.95	0.94	0.96

Characteristics	all participants, 2009	all participants, 2010	common participants, 2009	common participants, 2010
Average degree	55	54	73	62
Average strength	8590	7807	16426	13216
Average weight	159	141	416	404
Average contact duration in seconds	46	42	52	57
Average number of contact events per edge	3.44	3.37	8	7

Barrat, Alain, et al. "Social dynamics in conferences: analyses of data from the Live Social Semantics application." The Semantic Web–ISWC 2010. Springer Berlin Heidelberg, 2010. 17-33.

Where can you apply what you have learned?

Understanding users and developing Web applications for them

Sociology & Web Engineering

Consulting corporations about social media activities

Economics & Web analytics

Political decision making processes on the Web

Web engineering and analytics realizing your ideas as Web service and starting your own company

Entrepreneurship & Web analytics Research

Developing the next generation of the Web

New forms of collaboration

Linked data web

Fostering democracy

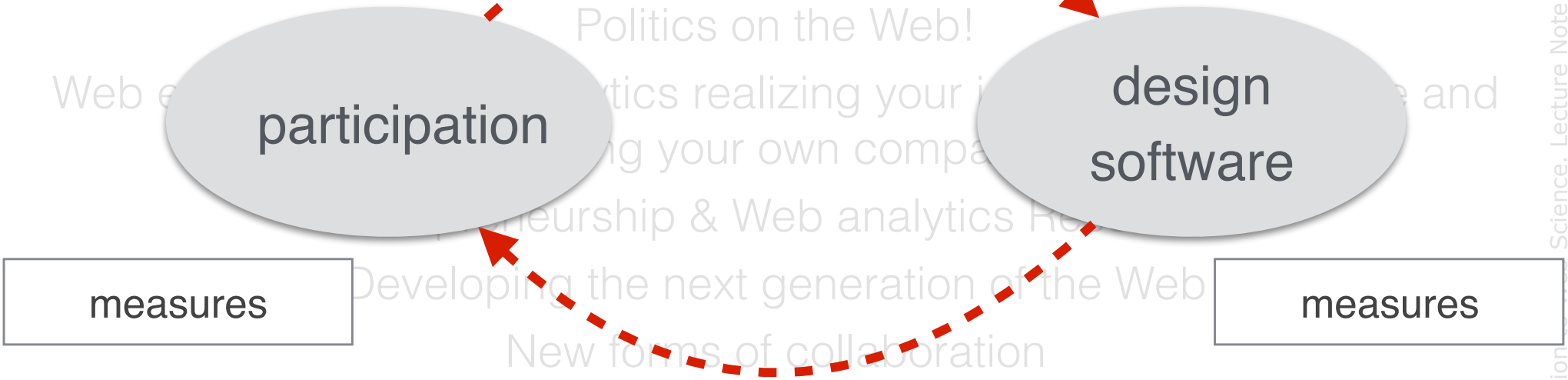
Improving security

Improving health care

....

Where can you apply what you have learned?

Understanding users and developing Web applications for them!



Sociology & Web Engineering
Consulting corporations about social media activities
Economics & Web analytics
Politics on the Web!
Developing the next generation of the Web
New forms of collaboration
Linked data web
Fostering democracy
Improving security
Improving health care

Dimensions of Analyzing Networks

Semantic Dimension of Network Analysis

Presentation of qualitative data (e.g. tabular with frequency, bar chart)	Presentation of quantitative data (e.g. histogram)	Measures of central tendency and variability (e.g. mean, range)

Syntactical Dimension of Network Analysis

Local structure	*Global structure*	*Partitions*
• Degree • Degree Centrality • Closeness Centrality • Betweenness Centrality • Local CC	• Mean degree • Degree distribution • Density • Network Centralization • Components	• Local definition, such as clique, k-core, k-plex • Global definition with null model (modularity with modularity optimization and edge betweenness)

About the course

In this course

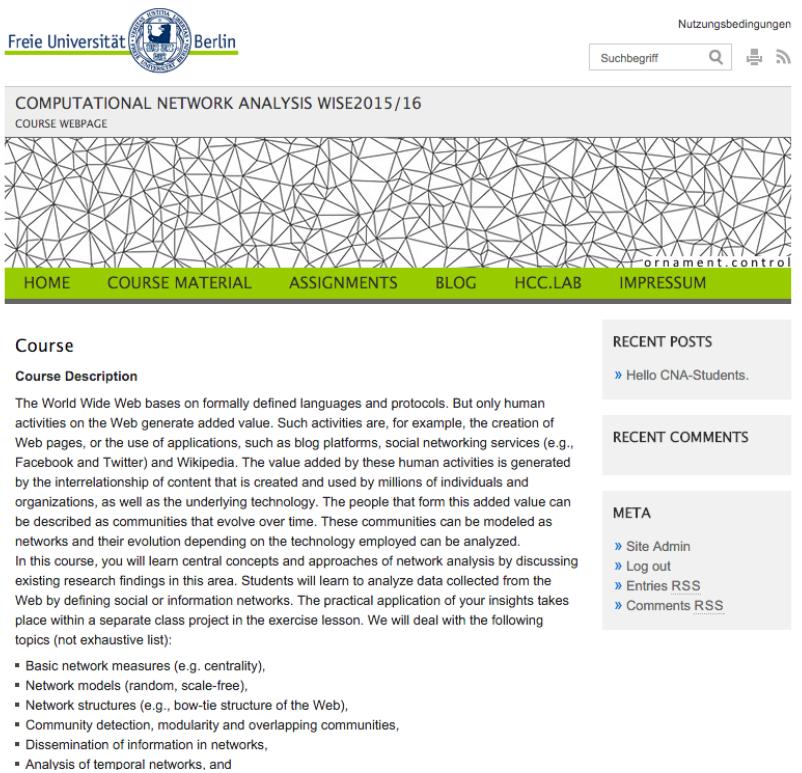
- You will learn about models and metrics which allow us to understand phenomena which have just discussed.
- You will investigate a large datasets of real social or technological networks.
- You will discuss your insights and explain, how your results can be used for.

Structure of Course

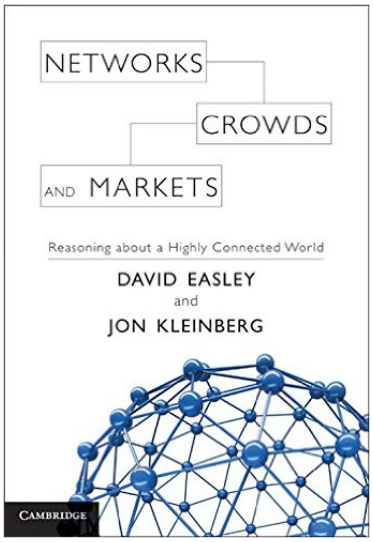
- General information
  - Block course from 2016-02-22 to 2016-03-11
  - Course No. 19620, 2+2 SWS hours
  - 051 Seminarraum (Takustr. 9)

- Lecture: 10:15 AM - 11:45 AM given by Prof. Birn
- Lab: 12:15 PM - 2:00 PM given by Kadir Tugan

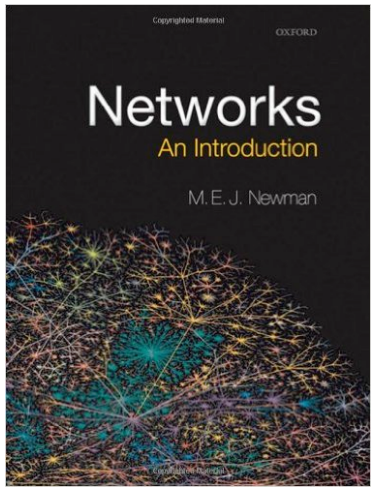
- Our website:
  http://blogs.fu-berlin.de/cna2016/



Main ressources



Easley, David, Kleinberg, Jon: Network, crowds, and markets. Cambridge, 2010.



Newman, Mark: Networks: An Introduction. Oxford University Press, 2010.

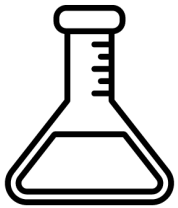
List of Lectures

- Lecture 1: Networks are everywhere
- Lecture 2: Collecting, Presenting and Visualizing Networks
- Lecture 3: Paths, Small World and Connectivity
- Lecture 4: Clustering Coefficient and Network Models
- Lecture 5: Affiliation Networks and Scientific Collaboration
- Lecture 6: Centrality and Application
- Lecture 7/8: Preparing Research Project with Q&A Session
- Lecture 9: Temporal Networks
- Lecture 10: Global Community Detection
- Lecture 11: Local Community Detection
- Lecture 12: Network Visualization
- Lecture 13: Diffusion in Networks
- Lecture 14: Poster Presentation
- Lecture 15: Writing-up Research Report

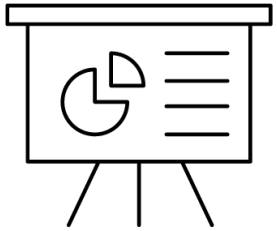
Assignments

- Your final grade (Portfolio-Prüfung) is based on three contributions:

  - Lab Participation 10%



  - Poster Presentation 30%



  - Research paper report 60%

Assignment 1: Lab Participation

- Participation in the lab in mandatory.
- One or more students will present the results of their homework at the beginning of each lab session.
- Only in week 1 you have to carry out homework. From week 2 you can work on your research project but tasks will be provided to guide you through the analysis process.
- Each student should present her/his homework at least once.
- Another possibility for participation are “lessons learned” or “tips and tricks” - what “aha”-effect did you have while solving the homework or working on your class project.

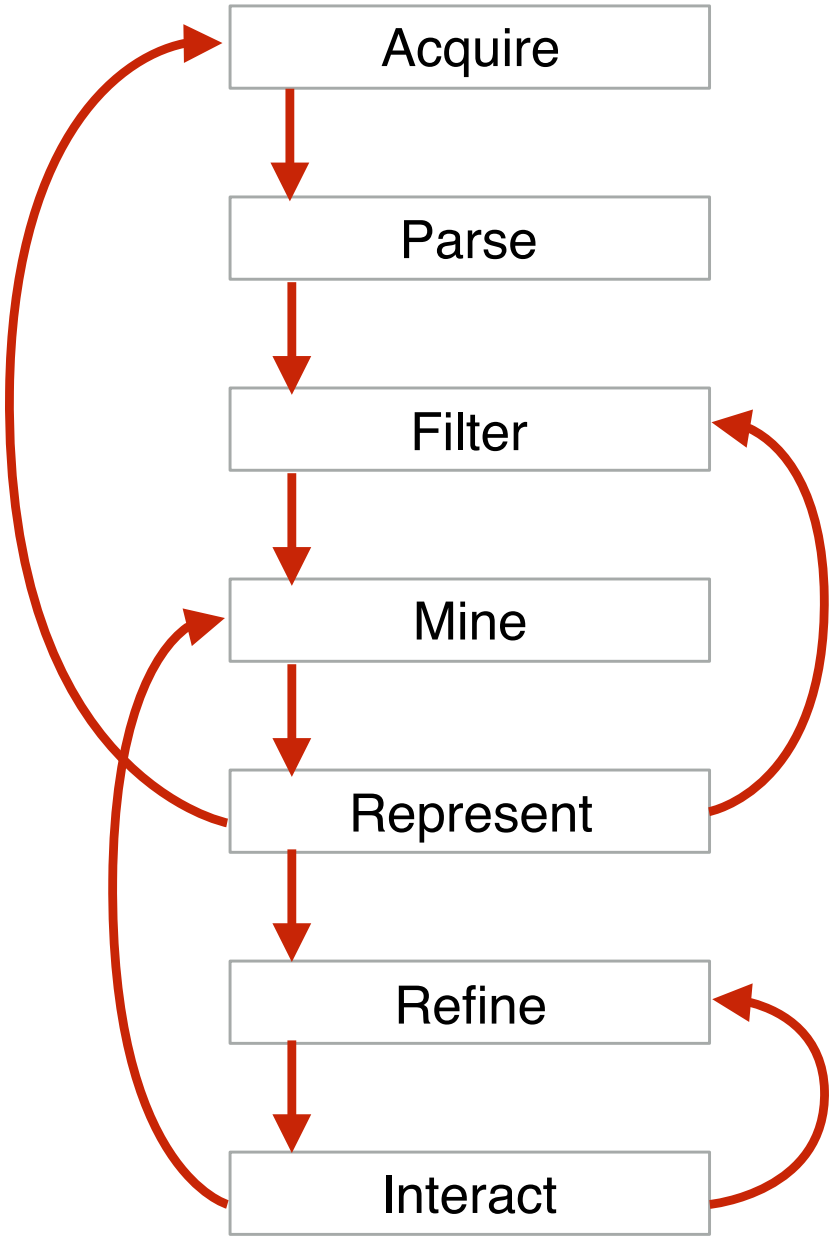
Assignment 2/3: Data Analysis Project

- Every student will complete a data analysis project which consists of the analysis of a dataset.

- The goal is to apply practically the learned concepts and methods in a real setting. Please keep in mind, understanding the inner structure of the data and interpreting the results by considering the context of the data origin is the core of the project.

- Each student is encouraged to come up with original project proposals which may be inspired by the data sets shown on the following slides.

Data Analysis Procedure



Obtain the data, whether from an API

Provide some structure for the data's meaning, and order it into categories

Remove all but the data of interest

Apply methods from network analysis in a way to discern patterns or place the data in mathematical context

Choose if needed a basic visual model, such as table, bar graph, list, or network to represent your results

Improve the applied methods and representation to make it clearer and more visually engaging

Manipulate the data or control what features are visible

(adapted from Fry 2008)

Possible data sets

Community/Information Hub	Reference
Semantic MediaWiki	Yolanda Gil, Angela Knight, Kevin Zhang, Larry Zhang, Ricky Sethi. (2013). An Initial Analysis of Semantic Wikis. Proceedings of the ACM International Conference on Intelligent User Interfaces.
Bibsonomy/del.icio.us	Mika, P. 2007, 'Ontologies are us: A unified model of social networks and semantics', Web Semantics: Science, Services and Agents on the World Wide Web, 5, 1, 5--15.
Github	F. Thung, T. F. Bissyande, D. Lo, and L. Jiang, Network Structure of Social Coding in GitHub. CSMR 2013
Historical Data Europeana	B.R. Erick Peirson, Julia Damerow, Manfred D. Laubichler, Don't Panic! A research system fornetwork-based digital history &philosophy of science, Unpublished Manuscript (Mar, 2014)
Scientific Collaboration	Miguel, Sandra, Zaida Chinchilla−Rodriguez, and Félix de Moya−Anegón. "Open access and Scopus: A new approach to scientific visibility from the standpoint of access." Journal of the American Society for Information Science and Technology 62.6 (2011): 1130-1145. Data Source: https://www.elsevier.com/solutions/scopus
Mailman	Bird, C., Pattison, D., D'Souza, R., Filkov, V., & Devanbu, P. 2008, 'Latent Social Structure in Open Source Projects', OR http://journal.r-project.org/archive/2011-1/RJournal_2011-1_Bohn~et~al.pdf
Reddit	Salman Jamali and Huzefa Rangwala. 2009. Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis. In Proceedings of the 2009 International Conference on Web Information Systems and Mining (WISM '09). IEEE Computer Society, Washington, DC, USA, 32-38.

Possible data sets

Community	References
Stackoverflow	Amiangshu Bosu, Christopher S. Corley, Dustin Heaton, Debarshi Chatterji, Jeffrey C. Carver, and Nicholas A. Kraft. 2013. Building reputation in StackOverflow: an empirical investigation. In Proc. of the 10th Working Conference on Mining Software Repositories (MSR '13).
dbpedia	Zhou, L., Ding, L., & Finin, T. 2011, 'How is the Semantic Web evolving? A dynamic social network perspective', Computers in Human Behavior, 27, 4, 1294 - 1302.
freebase	Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. 2008, 'Freebase: a collaboratively created graph database for structuring human knowledge', , 1247--1250.
wordpress/blogger/ thumble	Adamic, L. A. & Glance, N. 2005, 'The political blogosphere and the 2004 U.S. election: divided they blog', , 36--43.
Instagram	Hu, Y., Manikonda, L., Kambhampati, S., & others 2014, 'What we instagram: A first analysis of instagram photo content and user types', Proc. AAAI ICWSM.
foursquare	Noulas, A., Scellato, S., Mascolo, C., & Pontil, M. 2011, 'An Empirical Study of Geographic User Activity Patterns in Foursquare.', ICwSM, 11, 70--573.
Open Street Map	Peter Mooney & Padraig Corcoran: How social is OpenStreetMap?. Proceedings of the AGILE'2012

Some further inspirations

- Stanford Large Network Dataset Collection
  - Contains online social networks, email communication networks, co-authorship networks, web graphs, but also Amazon networks
  - Available at: http://snap.stanford.edu/data

- KONECT (the Koblenz Network Collection)
  - Is a project to collect large network datasets of all types in order to perform research in network science and related fields
  - Available at: http://konect.uni-koblenz.de/

- **Please do not use these data sets since they often provide no context information and are already cleaned :)**

Assignment 2: Poster Presentation

- As part of your data analysis project, you will create a poster that summarizes your work in progress. The idea and intention is to reflect on the study procedure and the results.

- During the last week of class, students will present their posters.

- Our class sessions will function as research conferences, with student presenters reviewing poster content and answering questions as small groups of viewers visit their poster areas.

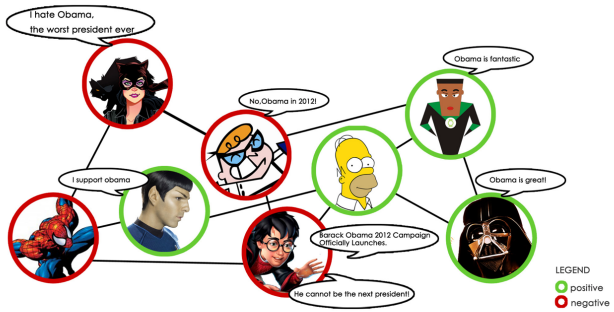
- Presenters will also collect viewer feedback on their research report.

- Additional information will be provided.

User-Level Sentiment Analysis Incorporating Social Networks

Chenhao Tan[1], Lillian Lee[1], Jie Tang[2], Long Jiang[3], Ming Zhou[3], Ping Li[1]
1. Cornell University, 2. Tsinghua University, 3. Microsoft Research Asia

Motivation



User-level Sentiment Analysis

- **Users** are the final target

In most cases, we want to study what each user has in mind

- More information user-wise

In contrast with document-level sentiment classification, we get more information for each user to make beter prediction in user-level

Network information

- Accessibility

User-relationship information is now more easily obtainable in social media

- Homophily or Attention ?

Users create a connection for a personal relationship or a desire to pay attention

Problem setting

Concrete problem in Twitter

Our task is to classify each user's sentiment on a specific topic into one of two polarities: "Positive" and "Negative"
It is hard to get full labels
We formulate a semi-supervised learning problem in Twitter:
Given a graph and labels of some nodes in the graph, try to classify the other users in the graph

Graph types

Directed t-follow graph Mutual t-follow graph
Directed @ graph Mutual @ graph

Data Analysis

Total data crawled: 1,414,340 users, 1,414,211 user profiles, 480,435,500 tweets, 274,644,047 t-follow edges, 58,387,964 @-edges.

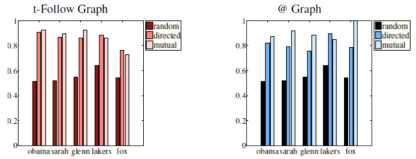
We employ user biographical information to label.

Table 1: Statistics for our main datasets.

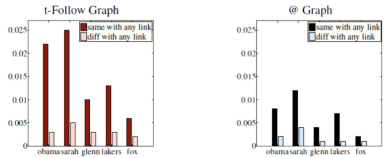
Topic	# users	#t-follow edges		#@ edges		# on-topic tweets
		dir.	mutual	dir.	mutual	
Obama	889	7,838	2,949	2,358	302	128,373
Sarah Palin	310	1,003	264	449	60	21,571
Glenn Beck	313	486	159	148	17	12,842
Lakers	640	2,297	353	1,167	127	35,250
Fox News	231	130	32	37	5	8,479

Sentiment and connectivity are correlated

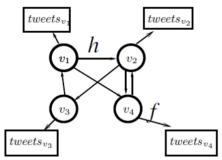
Shared sentiment in different types of graphs



Connectedness conditioned on labels



Approach



$$h_{k,\ell}(y_{v_i}, y_{v_j}) = \begin{cases} \frac{w_{\text{relation}}}{|Neighbors_{v_i}|} & y_{v_i} = k, y_{v_j} = \ell \\ 0 & \text{otherwise} \end{cases}$$

$$f_{k,\ell}(y_{v_i}, \hat{y}_t) = \begin{cases} \frac{w_{\text{labeled}}}{|tweets_{v_i}|} & y_{v_i} = k, \hat{y}_t = \ell, v_i \text{ labeled} \\ \frac{w_{\text{unlabeled}}}{|tweets_{v_i}|} & y_{v_i} = k, \hat{y}_t = \ell, v_i \text{ unlabeled} \\ 0 & \text{otherwise} \end{cases}$$

$$\log P(\mathbf{Y}) = \Big(\sum_{v_i \in V} \Big[\sum_{t \in tweets_{v_i}, k, \ell} \mu_{k,\ell} f_{k,\ell}(y_{v_i}, \hat{y}_t) + \sum_{v_j \in Neighbors_{v_i}, k, \ell} \lambda_{k,\ell} h_{k,\ell}(y_{v_i}, y_{v_j}) \Big] \Big) - \log Z,$$

- Parameter Estimation

Direct estimation from simple statistics
SampleRank

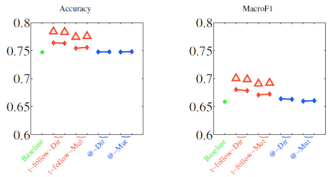
- Inference

Loopy belief propagation

Experiments

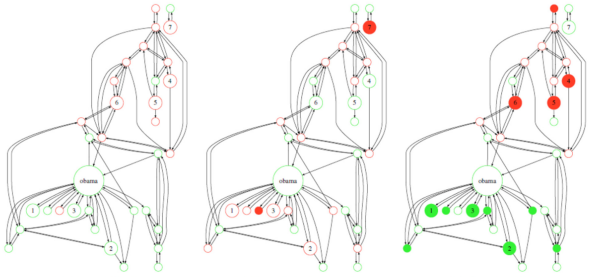
- Overall Performance

Beat Baseline!
Follow better than @
Directed better than Undirected
NoLearning same as Learning



- Case study

Our result is more clustered
It corrects many tweets that are hard to classify

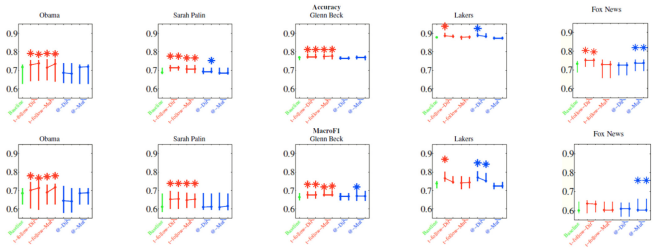


(a) Ground Truth (b) Text-Only Approach (c) Our algorithm

User ID	SVM Vote	HGM	True	Tweet
1	NEG	POS	POS	Obama is making the repubs look silly and petty. #hrc
2	NEG	POS	POS	Is happy Obama is President Obama collectable http://tinyurl.com/c5u7jf
4	POS	NEG	NEG	RT @TeaPartyProtest Only thing we have 2 fear is Obama himself & Pelosi & Cong & liberal news & Dems &... http://ow.ly/15M9Xv RT @GlennBeckClips: Barack Obama can no more disown ACORN than he could disown his own grandmother. #TCOT
5	POS	NEG	NEG	RT @JosephACaillaut Twilongerr: Suppose I wanted to Immigrant to Mexico? A Letter to President Obama.. http://ll.gd/1kr5rh George Bush was and acted like a war time President. Obama is on a four year power grab and photo op. #tcot

- Topical performance

Hard to interpret: Sparseness of graph? Size of graph?
#Tweets per user? SVM Classier Performance?



Future Work

Study larger datasets and different online social networks，find which parts of the whole network are helpful in different topics, build a theory of why and how users correlate on different topics in different networks.

References, related work, more experiment results in the paper!

Acknowledgment to Rong Chen for the design of the poster

Assignment 3: Research Report

- Report results in a research project description
- Submit report by: **Friday March 11, 2016, 23:59 PM**
  (later submissions will be accepted without penalties until March 20, 2016, 23:59 PM)
- Upload your research report to Sakai

General requirements for the class report

- Use the ACM template for the report (please use LaTeX option 1) http://www.acm.org/sigs/publications/proceedings-templates
- Your report can be written either in English or German
- Number of pages: at least 5 but not more than 8 pages (including at least five references and any excluding code documentation, use pictures and tables with care)
- Each student must submit her/his own report but working on the same data set is possible
- Additional information regarding the content will be provided.

General structure of your research report

1. Introduction
2. Related work
3. Description of used data (social web community)
4. Results 1: Data characteristics
5. Results 2: Network analysis
6. Discussion of insights
7. Conclusions and future work
8. References

Questions?