Carnegie Mellon

Big data, big research?

Opportunities and constraints for computer supported social science

Jürgen Pfeffer

Digital Methods
Vienna, Austria, November 2013

isr institute for SOFTWARE RESEARCH

Agenda

- Look and feel of big data research
- How is big data research different from traditional social science research?
- Methodological problems
  - Big data
  - Online social networks
- How big are big data?
- Technical/algorithmic problems

Goals

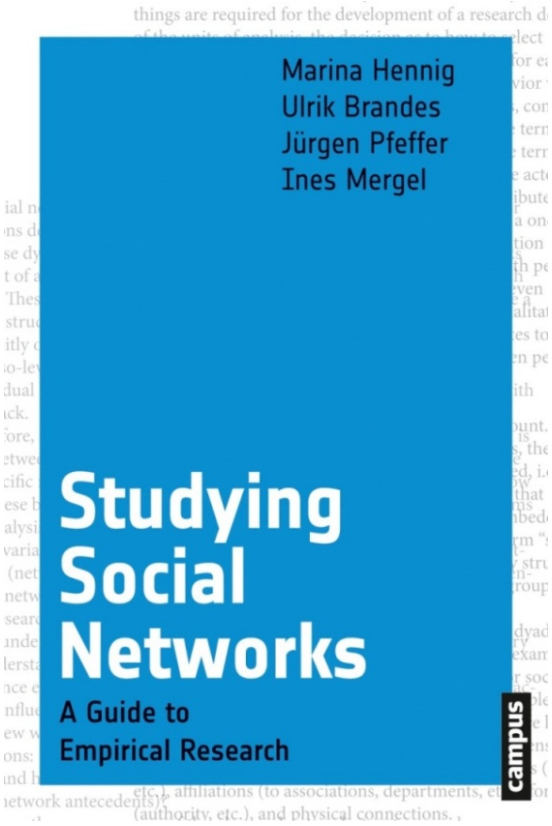
- Understanding big data research approach
- Seeing the current limitations
- Feeling the future potentials

Jürgen Pfeffer

- Assistant Research Professor
  School of Computer Science
  Carnegie Mellon University

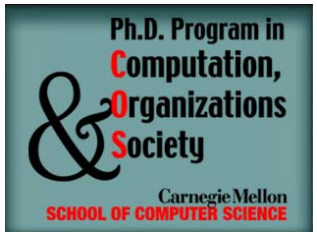
- Vienna University of Technology:
  - BA: Computer Science
  - PhD: Business Informatics

- Corporate Consultant, Freelancer
- Research Studios Austria
- Trainer for Rhetoric and
  Personal Performance



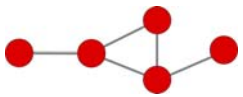
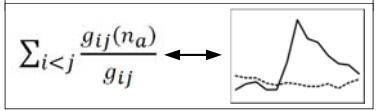
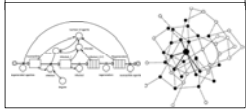
Jürgen Pfeffer

- Research focus:
  - Computational analysis of organizations and societies
  - Special emphasis on large-scale systems
- Methodological and algorithmic challenges
- Methods:
  - Network analysis theories and methods
  - Visual analytics, geographic information systems
  - Agent-based simulations, system dynamics



Center for Computational Analysis
of Social and Organizational Systems

Challenges for Analyzing Large-Scale Systems

Data Mining Text Mining	Data-to-Network Model	Algorithms Change Detection	Visual Analytics Geo Analysis	Modeling Simulation
				

- Mining of large amounts of diverse data
- Automated data-to-network processing
- Dynamic network analysis and change detection
- Visual analytics of network data
- Modeling and simulation of real world networks

Toward a Real Time Analysis of Large-Scale Dynamic Socio-Cultural Systems

Toward a Real Time Analysis of Large-Scale Dynamic Socio-Cultural Systems

Motivation & Hope

- "A field is emerging that leverages the capacity to collect and analyze data at a scale that may **reveal patterns of individual and group behaviors**. "
- "…access to terabytes of **data describing minute-by-minute interactions** and locations of entire populations of individuals… [will] offer qualitatively new perspectives on **collective human behavior**."

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). Computational social science. Science, 323, 721-723.

Motivation & Hope

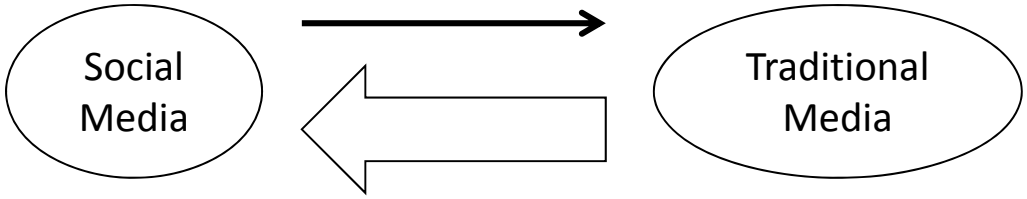
- "Social media offers us the opportunity for the first time to both observe **human behavior and interaction in real time** *and* on a global scale. "

Golder, S. A., & Macy, M. W. (2012, January). Social science with social media. ASA footnotes, 40(1), 7.

Example: Interplay Social Media/Traditional Media

Offline and online media reinforce one another

- Social media are an important information source for traditional media (Diakopoulos et al., 2012).
- Twitter is used as “radar”
- Social media hooks are connected to the media story
- Significant amount of dynamics are “external events and factors outside the network” (Myers et al., 2012)
- Online firestorms:



→ *Cross media dynamics*

Interplay Social Media/Traditional Media

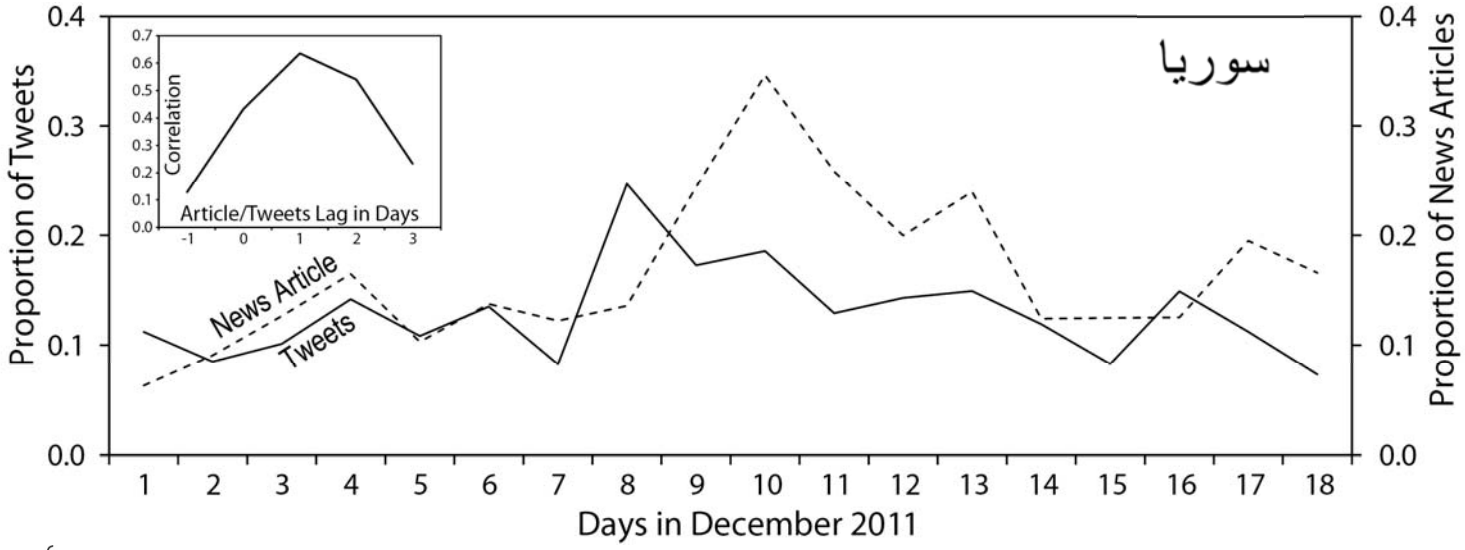
Traditional Social Science approaches:

- Survey Twitter/Facebook users
- Interview journalists
- Observe media web sites
- Content analysis
- Etc.

Interplay Social Media/Traditional Media

Data driven approach:

- Contrast *Arabic* tweets with *English* news articles (2 weeks):
  - 7,763 English news articles (“Syria”)
  - 61,633 Arabic written tweets from 10,186 users (“Syria”, “سوريا”)
    - Arabic written keywords related to humanitarian crisis, e.g. violence, death, food, shelter, etc. to reduce tweets



Pfeffer, J., Carley, K. M. (2012). Social Networks, Social Media, Social Change. Proceedings of the 2nd International Conference on Cross-Cultural Decision Making: Focus 2012, San Francisco, CA.

Interplay Social Media/Traditional Media

Data mining approach:

- Carlos Castillo (Qatar Computing Research Institute, Doha, Qatar)
- Mohammed El-Haddad (Al Jazeera, Doha, Qatar)
- Matt Stempeck (MIT Media Lab, Cambridge, USA)
- Jürgen Pfeffer (Carnegie Mellon University, Pittsburgh, USA)

Carnegie Mellon

Data Collection

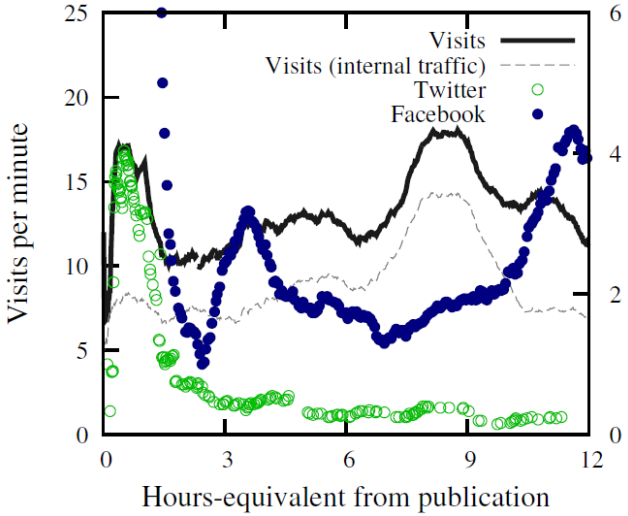
- AlJazeera.com
  - "beacon" embedded in all article pages
  - events are processed using Apache S4
  - collect and aggregate the visits with a 1-minute granularity
  - data is stored using a Cassandra NoSQL database
- Facebook.com
  - collect messages from Facebook discussing the articles
  - using the Facebook Query Language API
- Twitter.com
  - collect messages from Twitter discussing the articles
  - Using the Twitter Search API

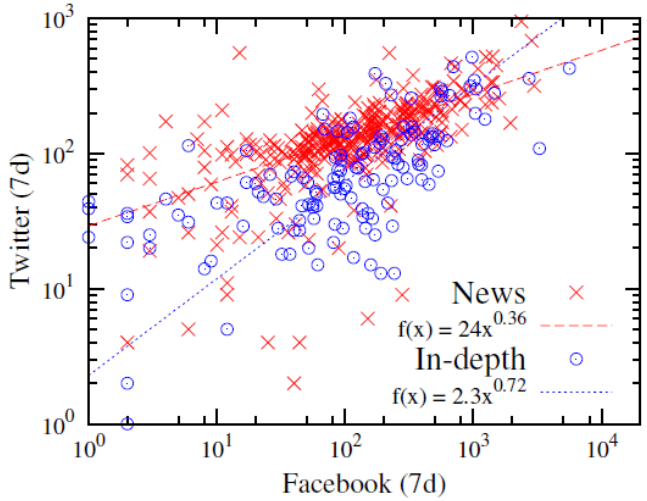
Data Collection

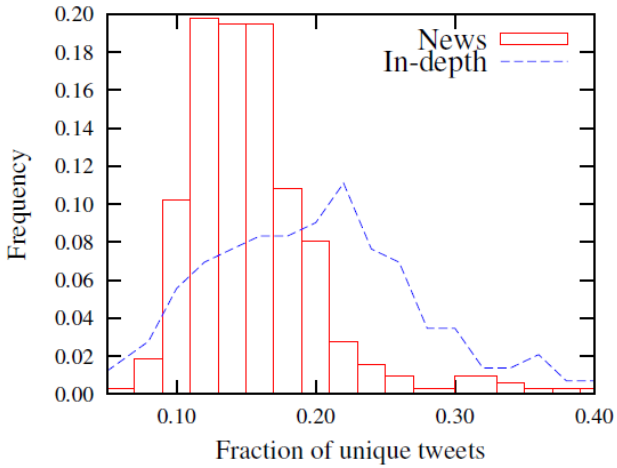
Case Study, 1 week of data:

- Number of articles 606
- Visits after 7 days 3.6 M
- Facebook shares 155 K#
- Tweets 80 K
- Where do the article visits come from

Interplay Social Media/Traditional Media





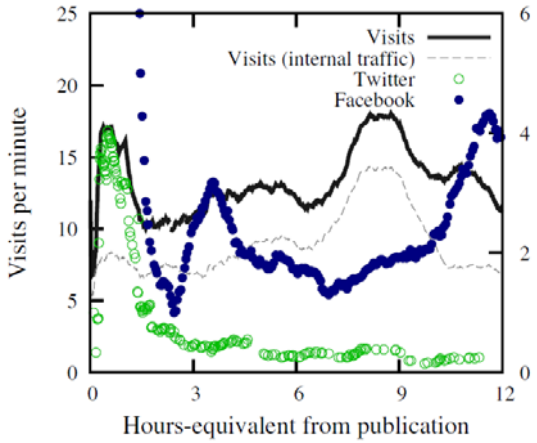


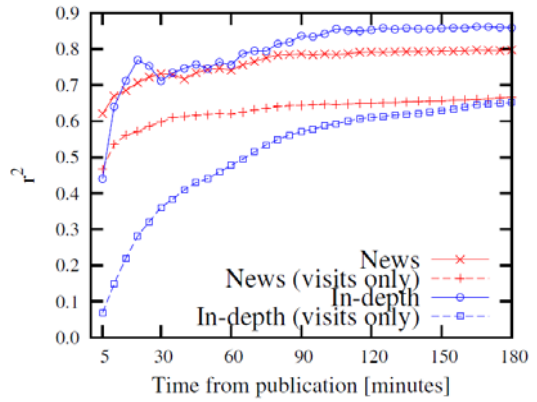
Castillo, Carlos & El-Haddad, Mohammed & Pfeffer, Jürgen & Stempeck, Mat (2014, forthcoming). Characterizing the Life Cycle of Online News Stories Using Social Media Reactions. 17th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW 2014), February 15-19, Baltimore, Maryland.

Interplay Traditional and Social Media

- Describing life cycle of online news stories
- Using early social media reactions
  - 20 minutes of Social Media activities
  - Can we estimate the 7-day visiting volume?
- Results:
  - Social media reactions can contribute substantially to the understanding of visitation patterns in online news.

After 20 Minutes	In-depth	News
Facebook shares	*	*
Twitter avg. followers	***	-
Volume of unique tweets	-	***
Twitter entropy	***	***



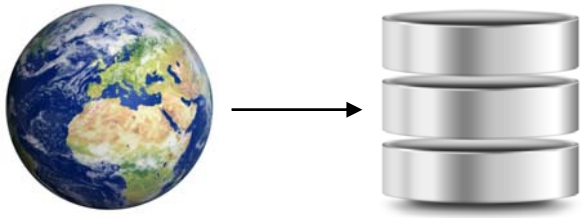


Al Jazeera Web Analytics Platform

- Al Jazeera launches predictive web analytics platform based on our research
- Media coverage:
  - Qatar Tribune
  - Doha News
  - Gulf Times
  - Fana News
  - Albawaba
  - Wan-Ifra
  - Rapid TV News
  - Etc.

Big Data Principles: Collect All Data

- Collect all available data
- No sampling, N = all
- There are no unrelated data
- Messy data and bad data is good
- Thousands of ("independent") variables
- We (the system) can decide later what is useful and what not



Data Driven Research Processes

Social Science

1. Problem
2. Research Question/ Hypotheses
3. Theories
4. Methods
5. Data
6. Analysis
7. Result Presentation

Typical Big Data Analysis

1. Methods
2. Data
3. Analysis
4. Result Presentation
5. Problem

Correlation not Cause: Babies and Storks

Social Science

- Collect other (socio-demographic) variables
- Build hypotheses about underlying variables
- Figure out that education is a good predictor for babies and storks (non-cities)
- Question: “Why?”

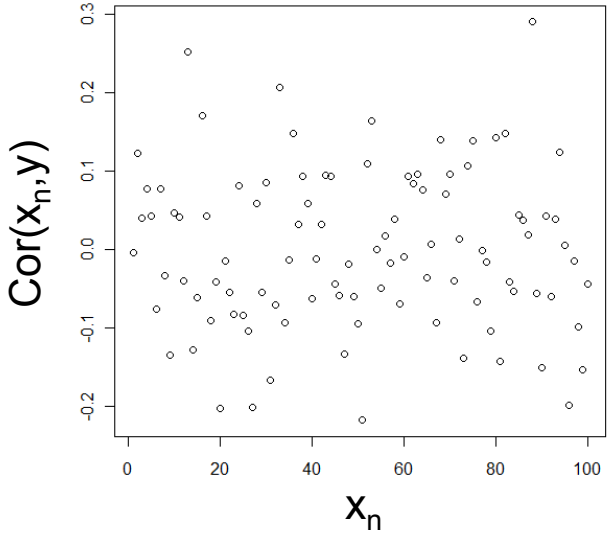
Big Data Analysis

- Include ~1,200 variables in a regression-like model.
- Number of storks and avg. car gas consumption are good enough predictors for number of babies
- Goodness of fit

Many Variables: Statistical Issues I

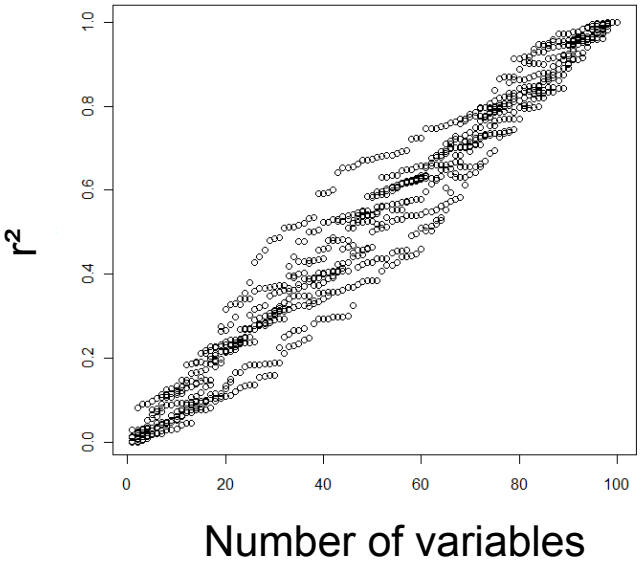
- 1st example:
  - 1 variable y, 100 elements, random 0-1
  - 1 variable x, 100 elements, random 0-1
  - Cor(x,y) = ~0.00
- 2nd example:
  - 1 variable y, 100 elements, random 0-1
  - 100 variable x_n, 100 elements, random 0-1
  - Cor(x_n,y) = ?

→ Something always correlates



Many Variables: Statistical Issues II

- 1st example:
  - 1 variable y, 100 elements, random 0-1
  - 1 variable x, 100 elements, random 0-1
  - r^2 - lm(x,y) = ~.0
- 2nd example:
  - 1 variable y, 100 elements, random 0-1
  - 100 variable x_n, 100 elements, random 0-1
  - r^2 - lm(x_1…x_n,y) = ?



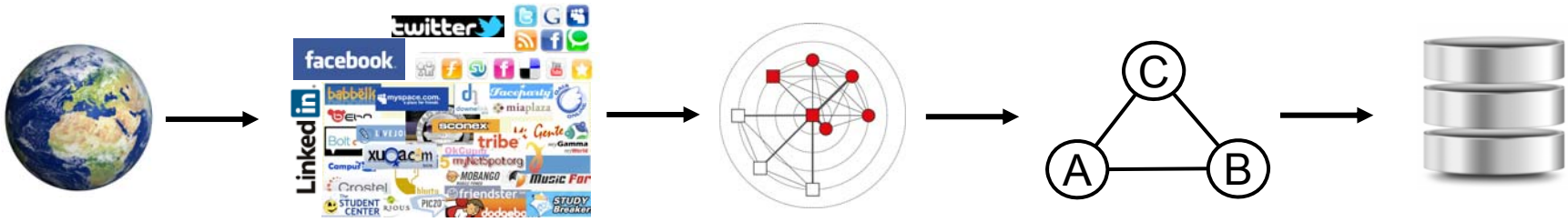
→ If you use enough variables, your r^2 is always high

N = All

- Is it all?
- All of what?
- Is it all of what we want?
- Is it all of what we think it is?

Multi-Level Bias Problem

1. Do the people online represent society?
2. Do the people that are online behave like offline?
3. Do the created data represent human behavior?
4. Do the analyzed data represent the created data?



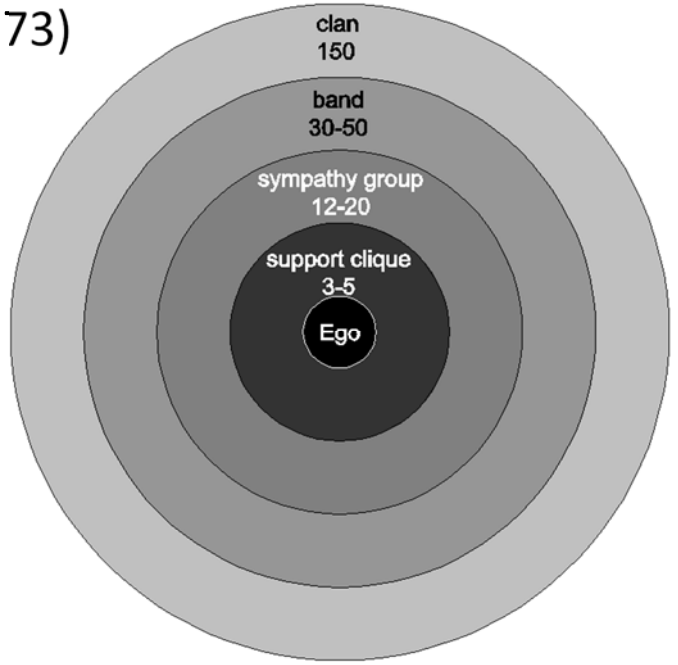
Do Created Data Represent Human Behavior?

Pfeffer, J. & Zorbach, T. & Carley, K.M. (2013). Understanding online firestorms: Negative word of mouth dynamics in social media networks. Journal of Marketing Communications

Empirical Observations/Factors

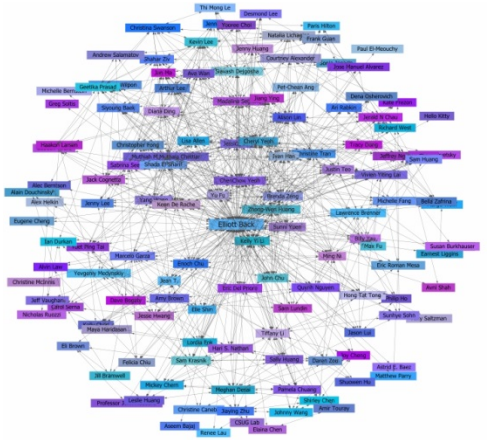
Hundreds of "friends" create many information

- Offline: Hierarchical groups of alters (Zhou et al., 2005)
- Strength of ties
  - amount of time, the emotional intensity, the intimacy, and the reciprocal service (Granovetter, 1973)
- In social media, every connection gets the same amount of attention



→ *Massive unrestrained information flow*

Empirical Observations/Factors



Amplified epidemic spreading, network clusters

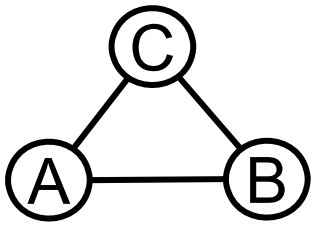
- Average Facebook user Ann: 130 friends
- Ben posts a very interesting piece of information
- Ben's friends like what Ben says (Homophily)
- Ben's friends are also friends with Ann (Transitivity)
- Ann receive a large amount of posts to one topic
- Amplifying effects of opinion-forming: echo chambers (Key, 1966)

→ *Network clusters & echo chambers*

Empirical Observations/Factors

Amplified epidemic spreading, network clusters

- Transitive link creations (Heider, 1946)
- Interpersonal communication networks have significant local clustering (e.g. Pfeffer and Carley, 2011)
- Local clusters are important for diffusion (e.g. Pfeffer and Carley, 2013)



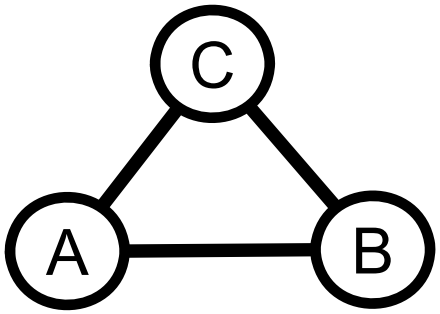
Pfeffer, Jürgen & Carley, Kathleen M. (2013). The Importance of Local Clusters for the Diffusion of Opinions and Beliefs in Interpersonal Communication Networks. International Journal of Innovation and Technology Management 10 (5)

Pfeffer, Jürgen & Carley, Kathleen M. (2011). Modeling and Calibrating Real World Interpersonal Networks. Proceedings of the IEEE NSW 2011, 1st International Workshop on Network Science, 9-16.

Do Created Data Represent Human Behavior?

- Transitive link creations (Heider, 1946)
- Programmers of social media know this
  - E.g. ~70% of new links on LinkedIn are triadic closure
  - Groups to follow, etc.
- Social media systems intensify this effect with link suggestions

→ Do we analyze society or a software implementation?

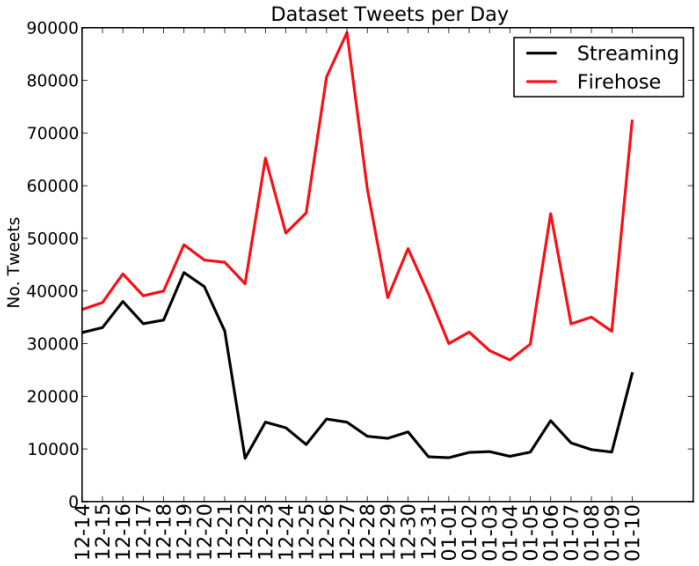


Do the Analyzed Data Represent the System?

Morstatter, F. & Pfeffer, J.& Liu, Huan & Carley, K.M. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. ICWSM, Boston, MA.

Sampling Twitter Data

- "Firehose" feed - 100% - costly.
- "Streaming API" feed - 1% - free.
- We don't know how Twitter samples data.
- Is the sampled data from the Streaming API representative of the true activity on Twitter's Firehose?



Sampling Twitter Data

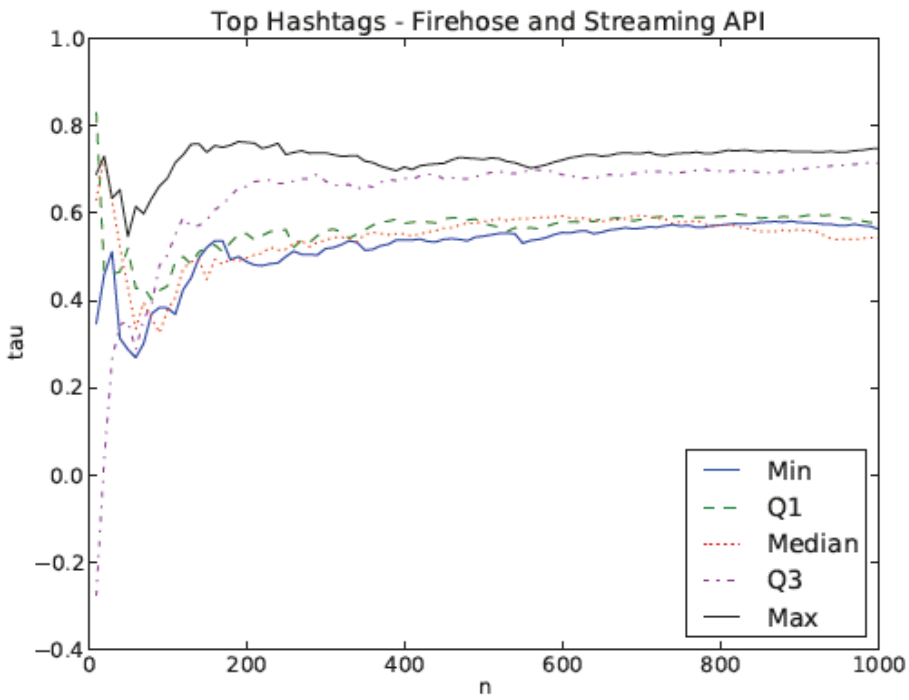


Figure 4: Relationship between n - number of top hashtags, and the correlation coefficient, τ_β.

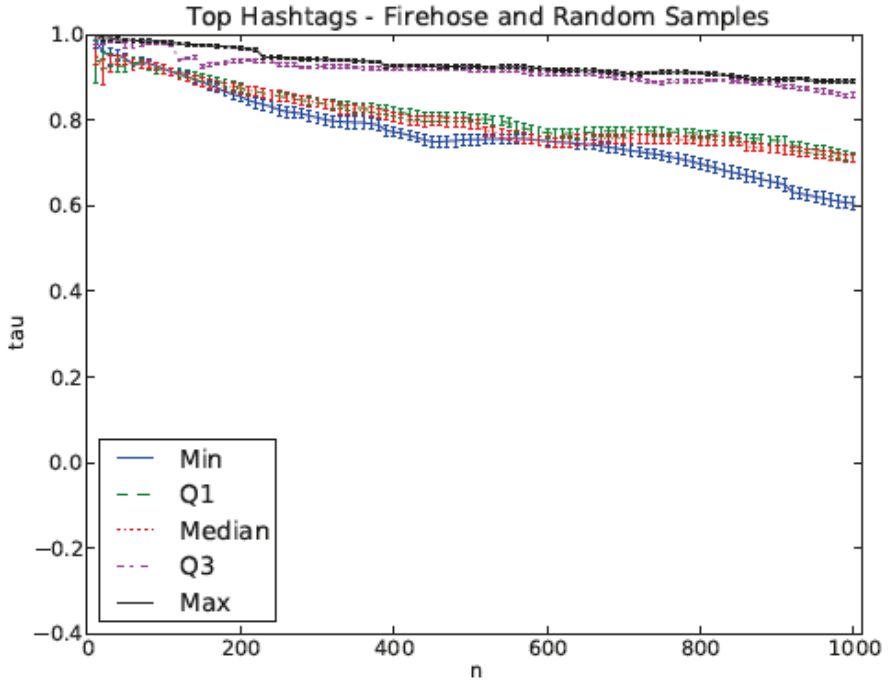


Figure 5: Random sampling of Firehose data. Relationship between n - number of top hashtags, and τ_β - the correlation coefficient for different levels of coverage.

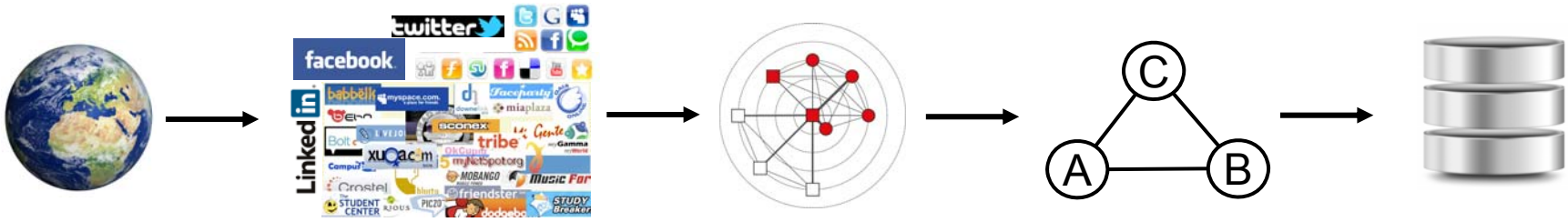
Sampling Twitter Data

- 42% Overall Coverage
- Daily Coverage from 17% to 89%.
- Can we find the right key-player?

Measure	*k*	Average Agreement (min-max)	All 28 Days
In-Degree	10	4.21 (0-9)	4
In-Degree	100	53.4 (36-82)	73
Betweenness	100	54.8 (41-81)	55
Potential Reach	100	59.2 (32-83)	80

Multi-Level Bias Problem

- ? Do the people online represent society?
- ? Do the people that are online behave like offline?
- ? Do the created data represent human behavior?
- ? Do the analyzed data represent the created data?



How Big are Big Data?

- Facebook is collecting your data — 500 terabytes a day
  - 2.5 billion status updates, posts, photos, videos, comments per day
  - 2.7 billion Likes per day
  - 300 million photos uploaded per day
  - $10M-$20M/year

http://gigaom.com/2012/08/22/facebook-is-collecting-your-data-500-terabytes-a-day/

- Total world Internet traffic in 2012: 1.1 Exabytes per day (1000petabytes = 1 million terabytes = 1 billion gigabytes)

http://www.cisco.com/web/solutions/sp/vni/vni_forecast_highlights/index.html

→ **Store just metadata!**

Big Networks: Memory Space

- 1 billion vertices, 100 billion edges
  - 111 PB adjacency matrix
  - 2.92 TB adjacency list
  - 2.92 TB edge list

Burkhardt & Waring, An NSA Big Graph experiment

Top Supercomputer Installations

- Titan Cray XK7 at ORNL — #1 Top500 in 2012
  - 0.5 million cores
  - 710 TB memory
  - 8.2 Megawatts
  - 4300 sq.ft.
- Sequoia IBM Blue Gene/Q at LLNL — #1 Graph500 in 2012
  - 1.5 million cores
  - 1 PB memory
  - 7.9 Megawatts
  - 3000 sq.ft.
- $7 million per year energy costs

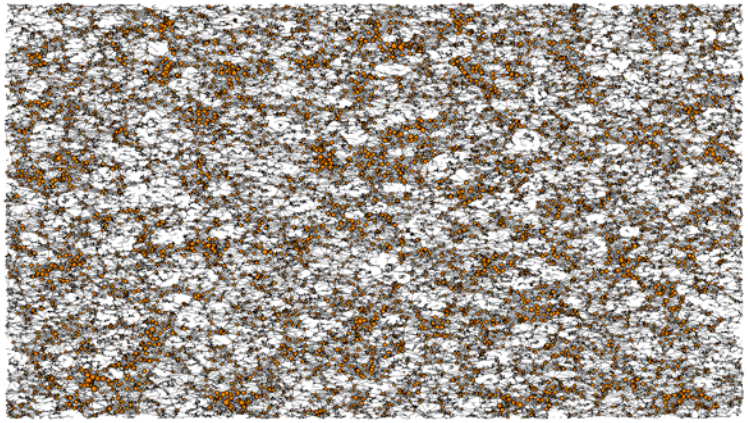
Source: Burkhardt & Waring, An NSA Big Graph experiment

Calculation Time

- State of the art algorithms for SNA metrics
  - require $\geq \Theta(n+m)$ space and
  - run in $\geq \Theta(nm)$ time, some in $\Theta(n^2)$ or $\Theta(n^3)$
  - with n = number of nodes, m = number of edges.

$$C^B(n_a) = \sum_{i<j} \frac{g_{ij}(n_a)}{g_{ij}}$$

- Example: 50k nodes, 193k edges
  - Betweenness centrality (Freeman 1979)
  - 1 processor, laptop: 51.23 min



Calculation Time

- Betweenness Centrality with Facebook
  - 264,399,256,813 min (500k years)
  - With 1,000,000 cores: 0.5 years
  - With 10x faster cores: 18.4 days

→ Approximations and localized algorithms

Large Networks?

Social Networks, 1 (1978/79) 215–239

Centrality in Social Networks Conceptual Clarification

Linton C. Freeman

*Lehigh University**

The intuitive background for measures of structural centrality in social networks is reviewed and existing measures are evaluated in terms of their consistency with intuitions and their interpretability.

Three distinct intuitive conceptions of centrality are uncovered and existing measures are refined to embody these conceptions. Three measures are developed for each concept, one absolute and one relative measure of the centrality of positions in a network, and one reflecting the degree of centralization of the entire network. The implications of these measures for the experimental study of small groups is examined.

Large Networks?

Figure 1. *A graph with five points and five edges.*

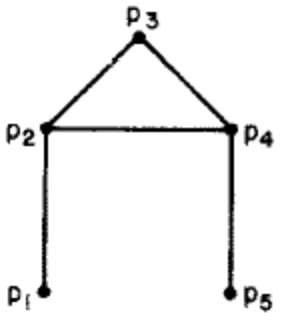


Figure 2. *A star or wheel with five points.*

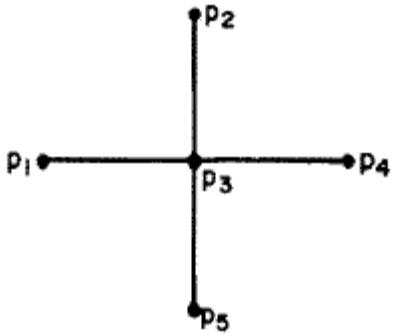
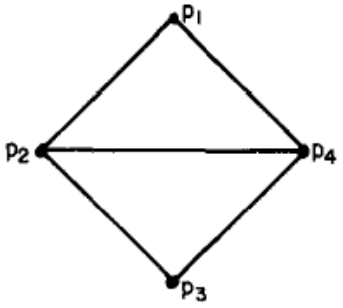
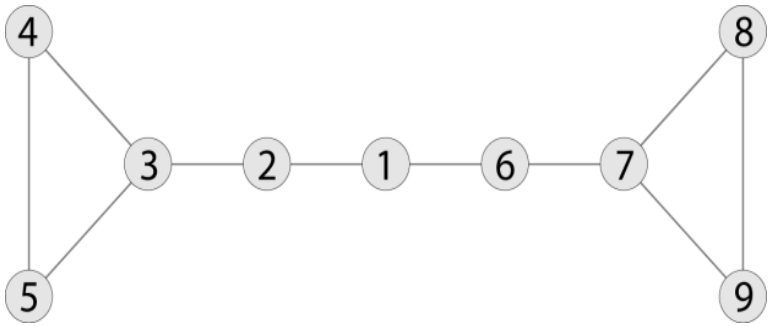


Figure 4. *A graph with four points and five edges.*



Large Networks: New Algorithms?

- What are the interpretations of our traditional measures for large networks?
  - E.g., does node nr. 1 sit in the center or rather on the fence?
  - What does it mean to be the most central actor on Facebook?
  - Approximation algorithms are new metrics!



- What are the research questions for large networked data at all?

Summary

- Big hopes and dreams related to big data
- … especially to social media data
- Data driven research is different
- Combining this with traditional social science research is not trivial
- Multi-level bias problem of social media data
  - Sampling issues
  - Representative
- Big data are really big
- But it is possible to handle big data/networks
- Many metrics have been developed for small groups, validity for big data is often not guaranteed

Other Issues

- Privacy
- Surveillance
- You are the product not the customer
- When correlations lead to predictions and interventions
  - Predictive policing

What needs to be done?

- Don't be blinded by big data
- Ask questions:
  - What do we learn from a study?
  - Do the authors ask "why?"
  - Good old research process is still important
- Don't be satisfied with one needle (especially, when you dream of the haystack)
- Let's utilize big data! But with care.

Conclusions: Mixed Methods

"The digital records of online behavior and social interaction hold the promise of opening up **a new era in the social and behavioral sciences**, but when and whether this opportunity is realized may depend on the involvement and leadership **of sociologists with the necessary technical and computational skills**."

"Online data should therefore be viewed as a **complement to, and not substitute for, data collected by traditional methods**. Indeed, in many cases, the value of online data may depend on opportunities to integrate with data obtained from surveys."

Golder, S. A., & Macy, M. W. (2012, January). Social science with social media. *ASA footnotes, 40*(1), 7.

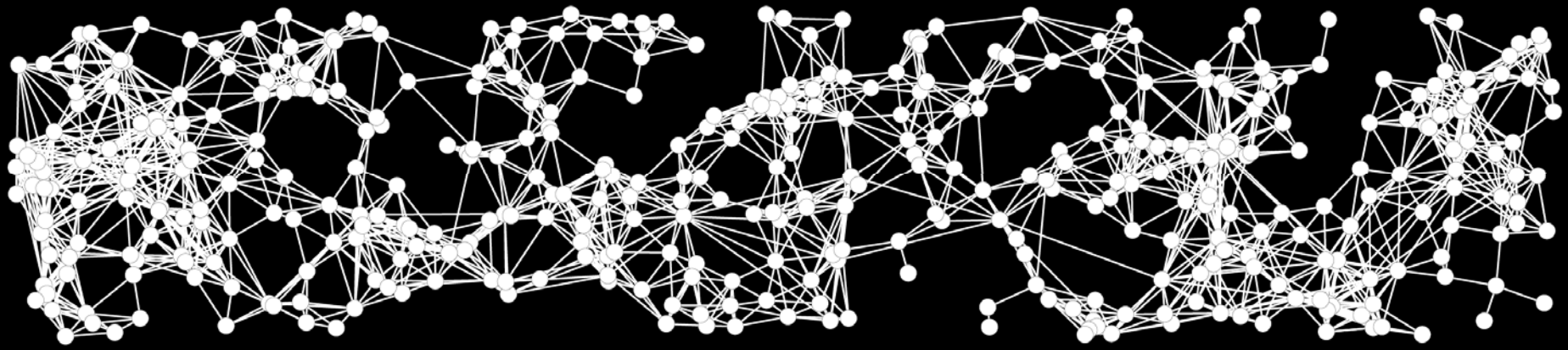
Conclusions

"Initially, computational social science needs to be the work of ***teams of social and computer scientists****. In the long run, the question will be whether academia should nurture computational social scientists, or* ***teams of computationally literate social scientists and socially literate computer scientists****."*

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). Computational social science. *Science, 323*, 721-723.

“Our mission is to go forward, and it has only just begun. There's still much to do, still so much to learn. Engage!”

Jean-Luc Picard, TNG Season 1 Ep. 26



Jürgen Pfeffer

jpfeffer@cs.cmu.edu