Implicit Affinity Networks

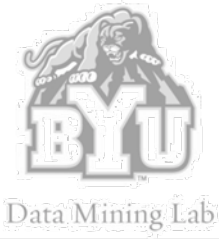
Matthew Smith
Christophe Giraud-Carrier
Brock Judkins (now at Amazon.com)
Department of Computer Science
Brigham Young University
Data Mining Lab - http://dml.cs.byu.edu

WITS 2007

Data Mining Lab

Outline

- Introduction & Motivation
- Project
  - Community Generation: IANs
  - Social Capital for Community Tracking
- Experiments & Observations
  - IAN Community
  - Blogosphere
- Conclusions and Future Work



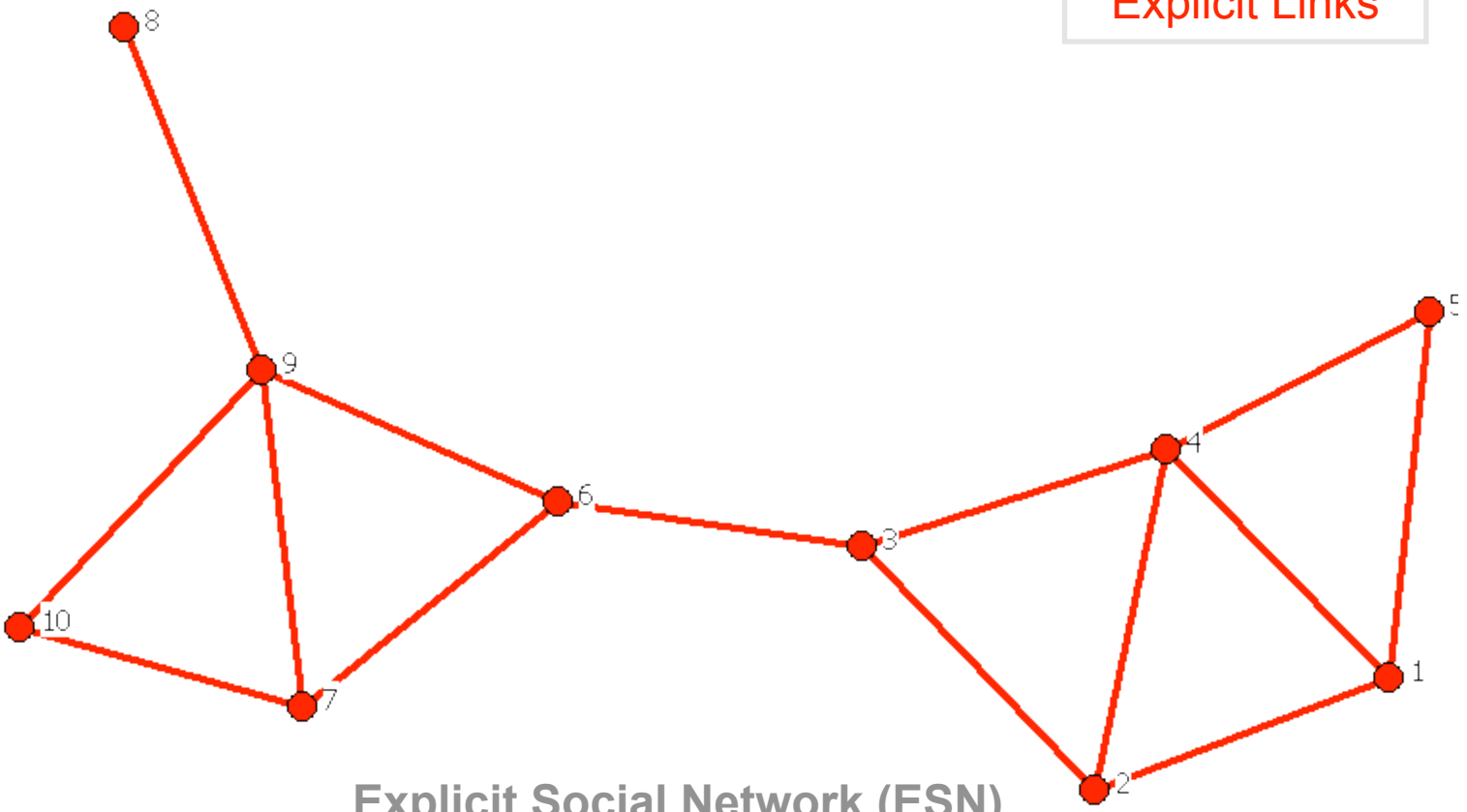
Introduction

- Online Communities
  - Continually emerging – many sites are adding this aspect
  - Like offline communities, they are complex and dynamic
- Examples
  - USENET (1980), Google Groups, Wikipedia
  - LinkedIn, Flickr, YouTube, MySpace, Facebook, etc.
  - Medical Communities (e.g., DailyStrength, NAAF)
  - Political Communities
  - Blogosphere – focus of experiments

Data Mining Lab

Motivation

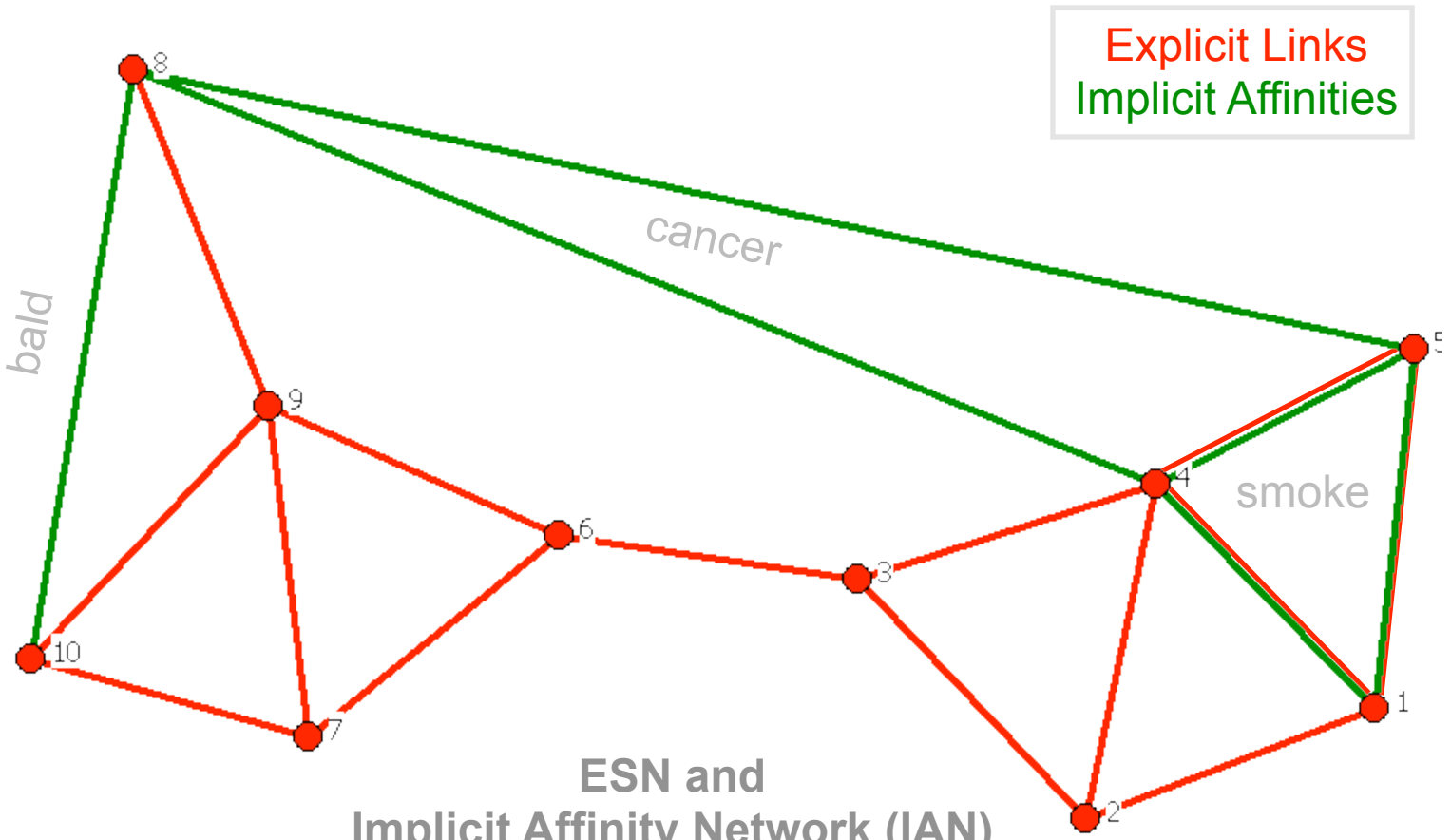
Explicit Links



Explicit Social Network (ESN)
Links: Friends, Web Links, etc.

Data Mining Lab

Motivation

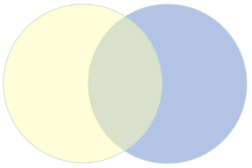


ESN and
Implicit Affinity Network (IAN)
Applications: Medical, Blogosphere, etc.

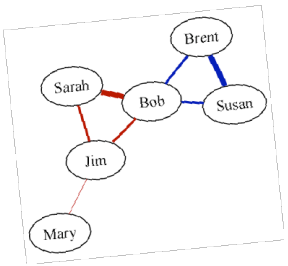
Data Mining Lab

Implicit Affinity

- Affinity:
  - The overlapping of attributes-values for any common attribute

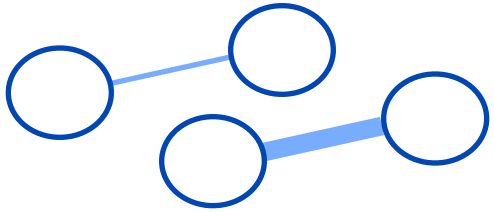


- Community:
  - Set of individuals characterized by attributes
  - Linked by affinities rather than explicit relationships



Data Mining Lab

Affinity Scoring



- Affinity score for a particular attribute – (Jaccard's Index)

$$AffScore_{A_k}(X,Y) = \frac{|\ \mathcal{V}_{A_k}(X) \cap \mathcal{V}_{A_k}(Y)\ |}{|\ \mathcal{V}_{A_k}(X) \cup \mathcal{V}_{A_k}(Y)\ |} \times \alpha_{A_k}$$

- Affinity score for all attributes

$$AffScore(X,Y) = \frac{\sum_{A_i \in Attr(X,Y)} AffScore_{A_i}(X,Y)}{|\ Attr(X,Y)\ |}$$

Let $Attr(X,Y) = Attr(X) \cap Attr(Y)$

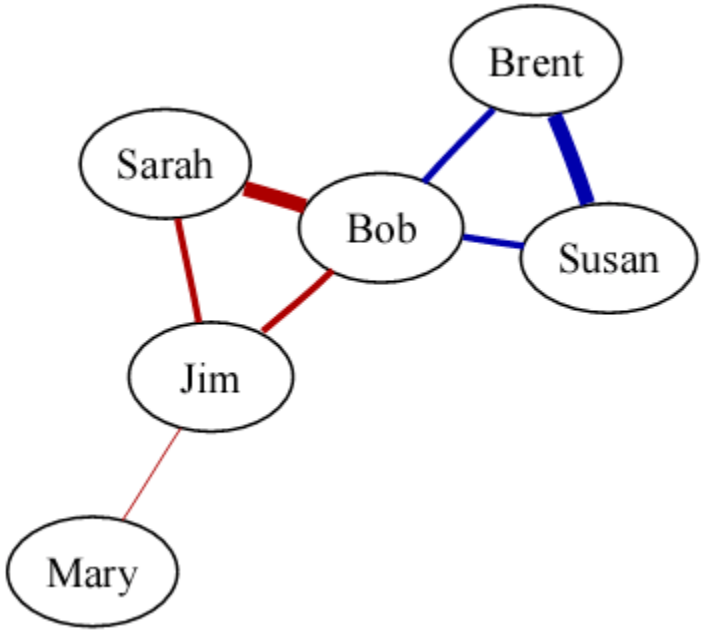
Data Mining Lab

Affinity Network Building

Sample of Individuals and their Attributes

Individual	Attributes
Jim	A: $\{a_1, a_2, a_3\}$
Sarah	A: $\{a_1, a_2\}$
Mary	A: $\{a_3\}$
Bob	A: $\{a_1, a_2\}$ B: $\{b_1, b_2\}$
Susan	B: $\{b_1, b_2, b_3\}$
Brent	B: $\{b_1, b_2, b_3\}$

Sarah										
1	0	Bob								
$\frac{2}{3}$	0	$\frac{2}{3}$	0	Jim						
0	0	0	0	$\frac{1}{3}$	0	Mary				
0	0	$\frac{2}{3}$	$\frac{2}{3}$	0	0	0	0	Susan		
0	0	$\frac{1}{3}$	$\frac{1}{3}$	0	0	0	0	0	1	Brent
A	B	A	B	A	B	A	B	A	B	



IAN

Data Mining Lab

Social Capital for Community Tracking

- **Social Capital**

  The advantage available through connections between individuals within a particular network

- **Bonding and Bridging Metrics**

$$BondingPotential = \frac{2}{N(N-1)} \sum_{\{X,Y\} \in E} AffScore(X,Y)$$

$$BridgingPotential = 1 - BondingPotential$$

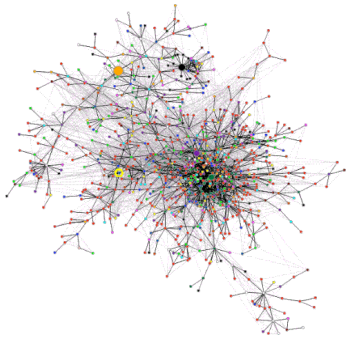
Data Mining Lab

Experiments & Observations

1. IAN Community Experiment

   Online at: http://dml.cs.byu.edu/IAN

2. Blogosphere Experiment



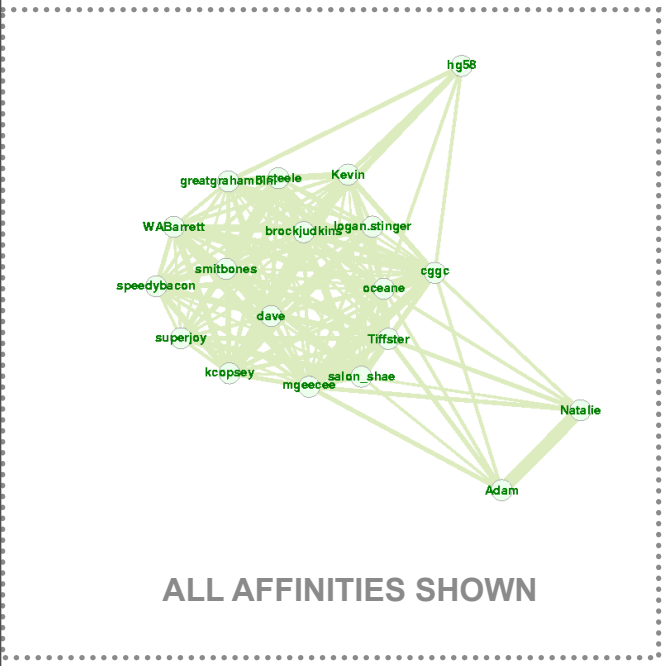
Data Mining Lab

IAN Experiment

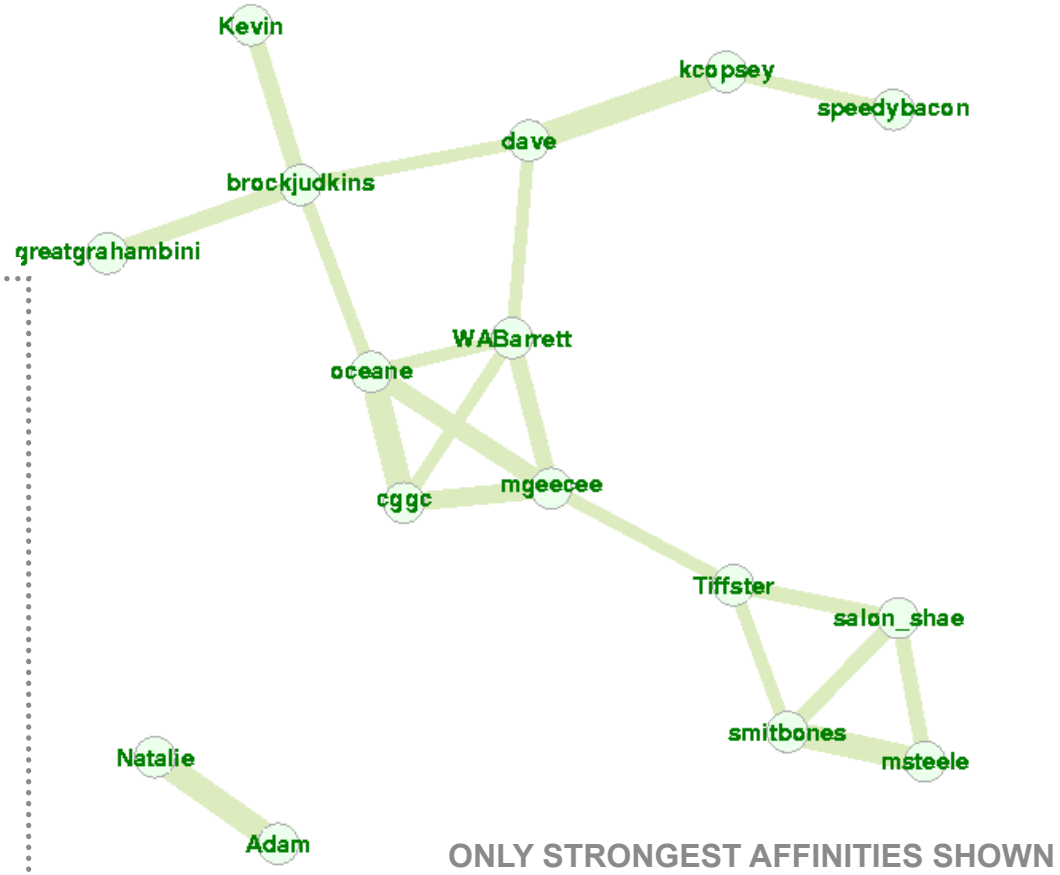
- 221 days
- 65 users participated in the experiment
- During first 183 days, on average, a user
  - was active for 39 days
  - visited every 8 days
  - added 2.38 attribute-values to profile per visit
  - had 93 attribute-values across 21 attributes

Data Mining Lab

Languages Spoken



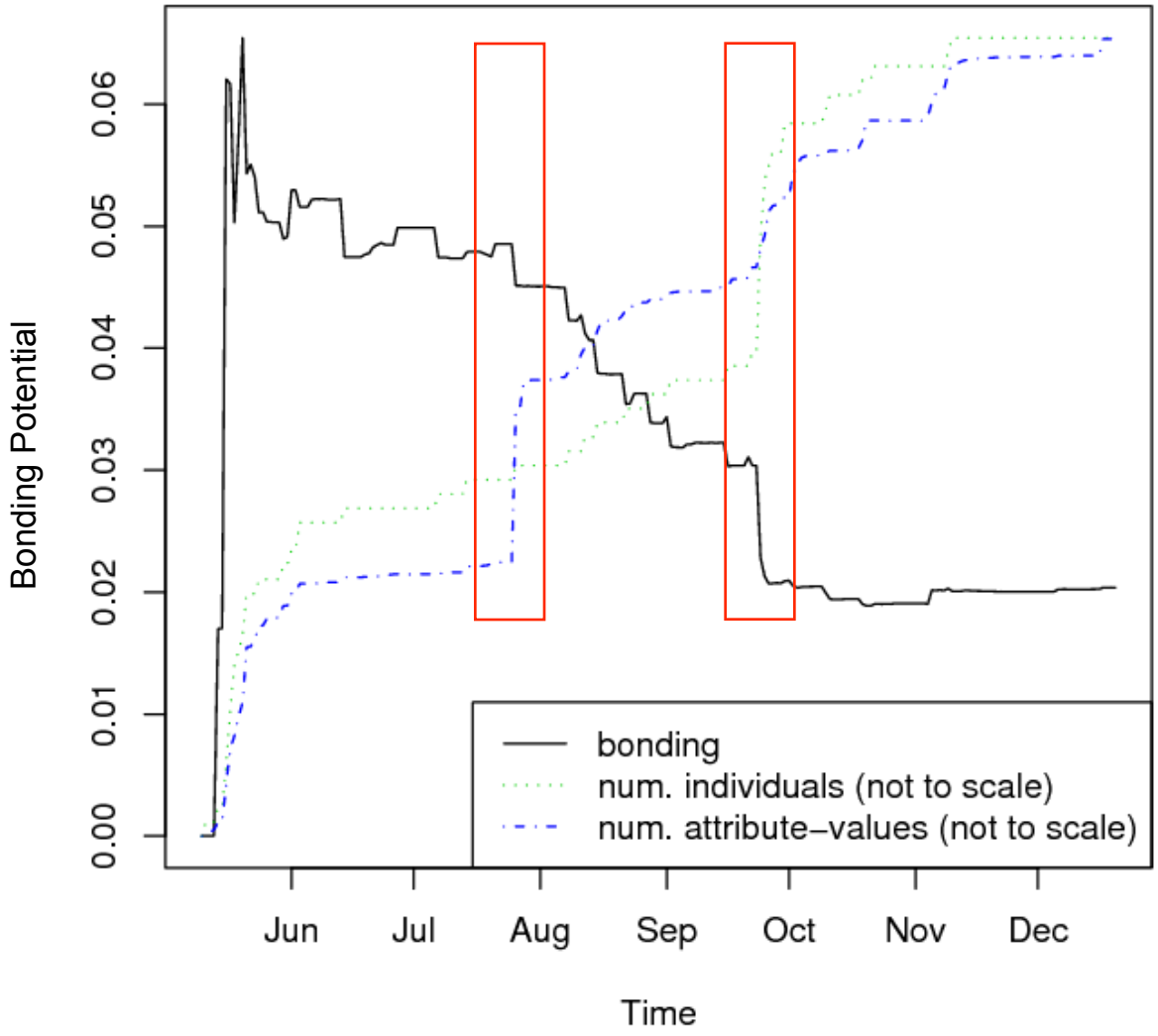
ALL AFFINITIES SHOWN



ONLY STRONGEST AFFINITIES SHOWN

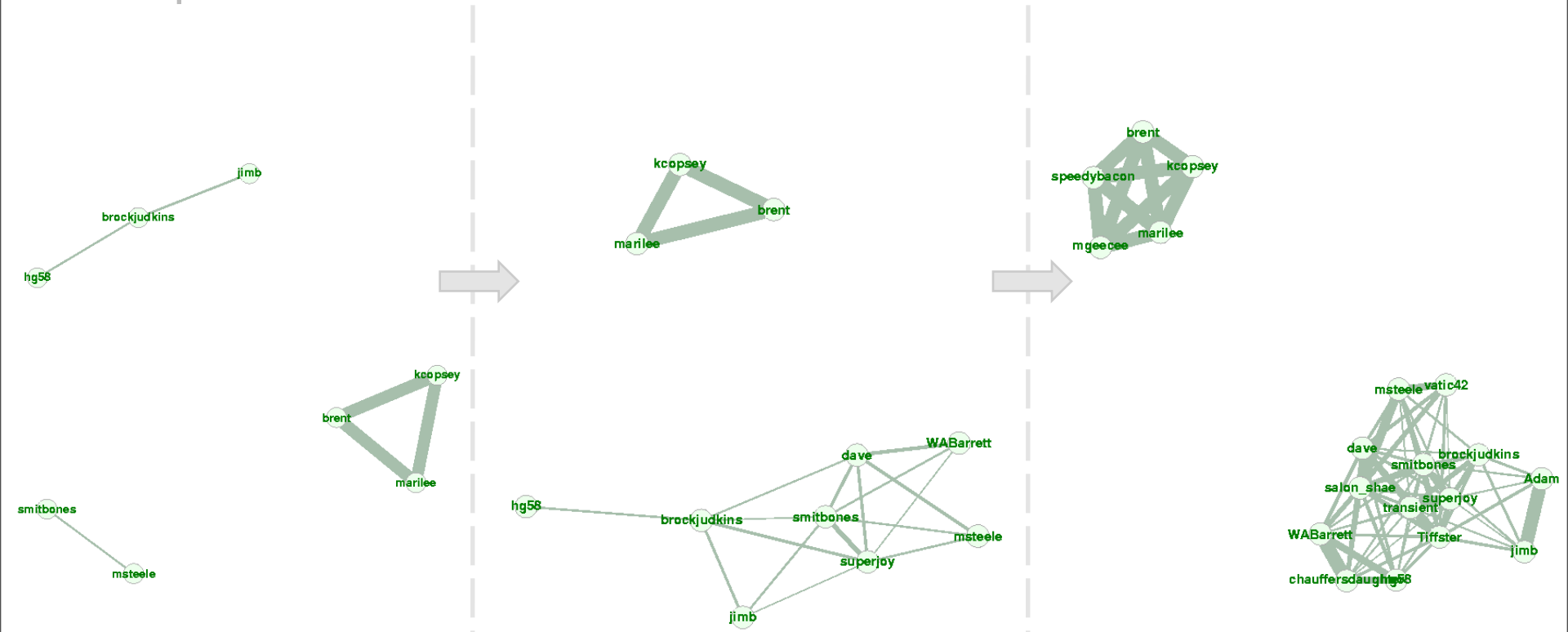
(IAN Snapshots taken: December 7, 2006)

IAN Community Evolution



Data Mining Lab

Musical Talents Evolution



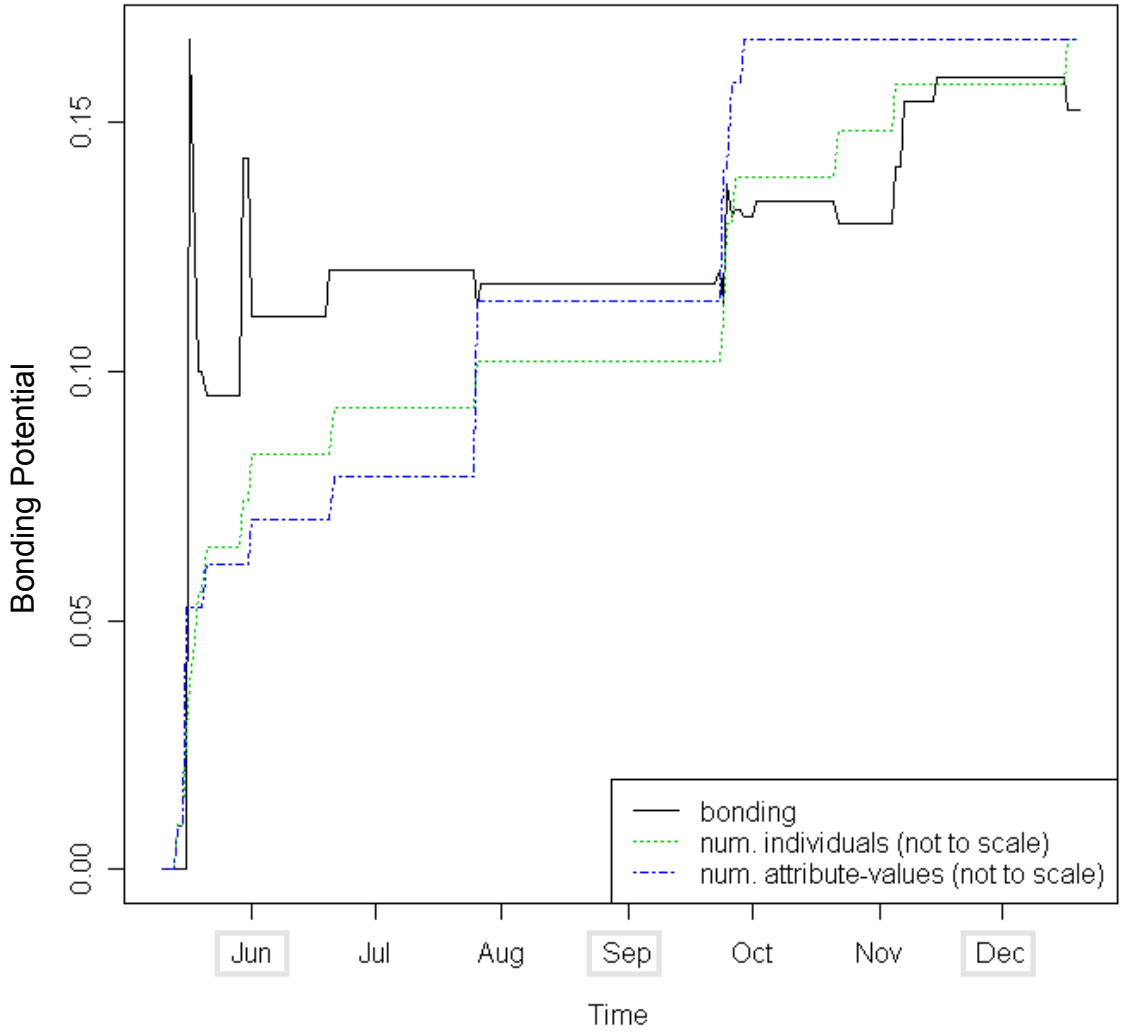
JUNE

SEPTEMBER

DECEMBER

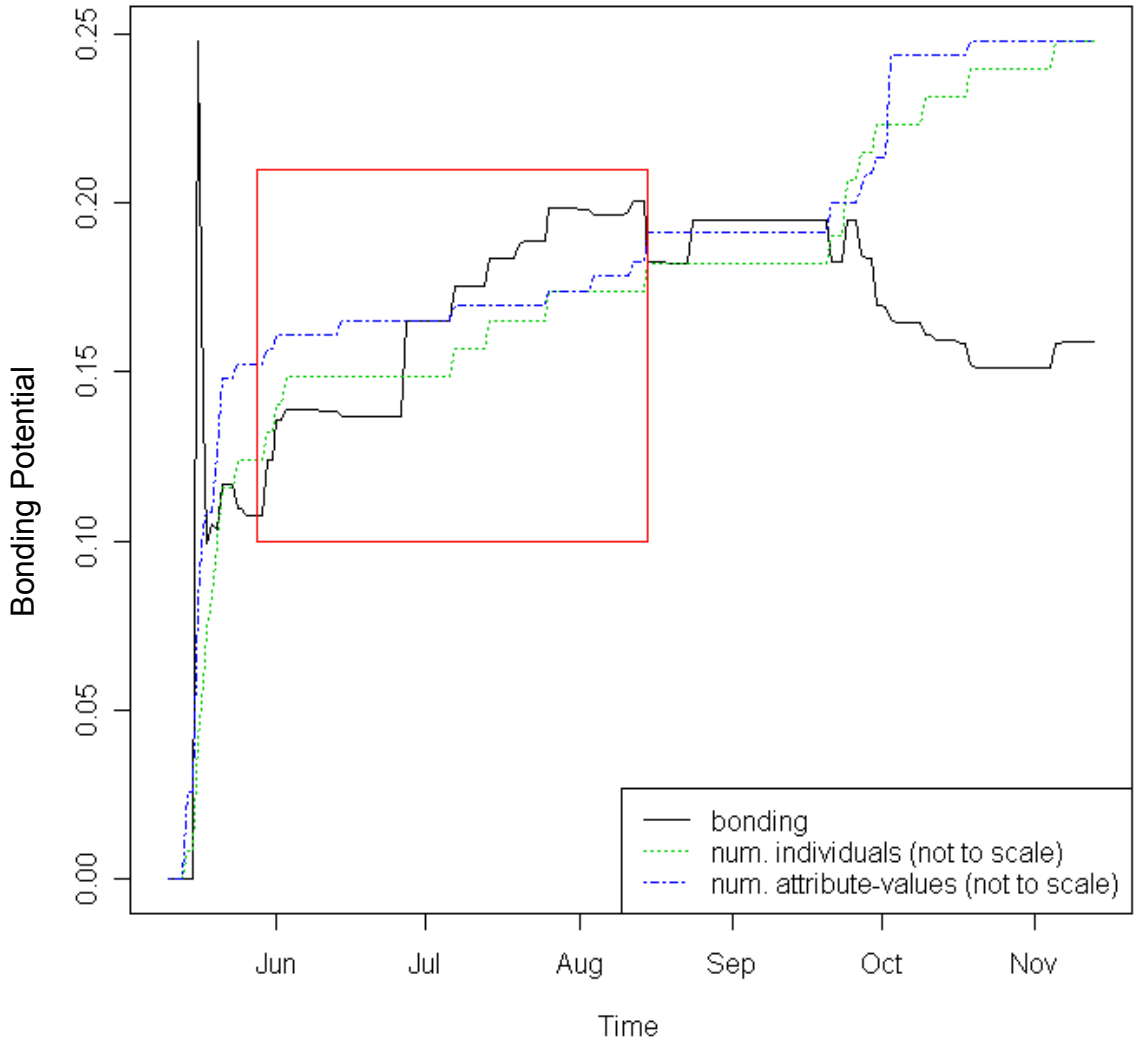
Data Mining Lab

Musical Talents Evolution



Data Mining Lab

Food Evolution



Data Mining Lab

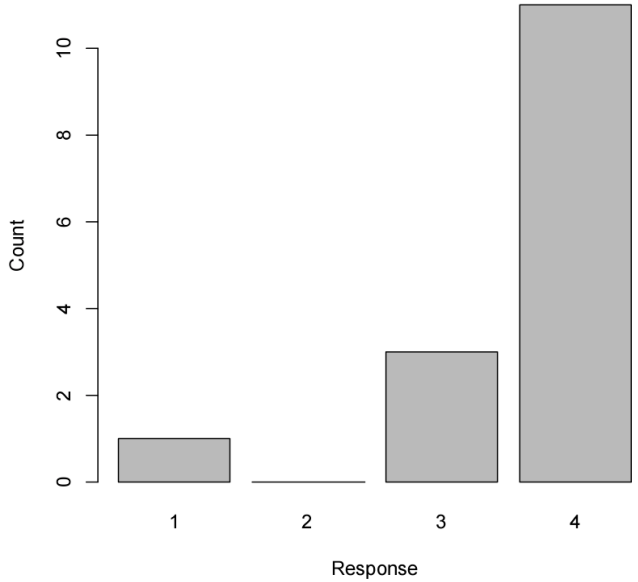
Qualitative Assessment - Survey

Q: What did you find most interesting about IAN?

It was interesting to see who was most related to me, and why

Q: What did you find most uninteresting about IAN?

Not knowing who some of the people in the community were made the affinities less interesting



Please choose one answer to complete the following sentence:

IAN _____

[ ] 1. did nothing for me

[ ] 2. showed me things I already knew

[ ] 3. highlighted things I suspected but was unsure about

[ ] 4. helped me discover new things

~23% responded (15 of 65)

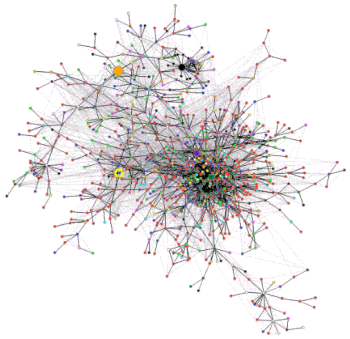
Data Mining Lab

Experiments & Observations

1. IAN Community Experiment

   Online at: http://dml.cs.byu.edu/IAN

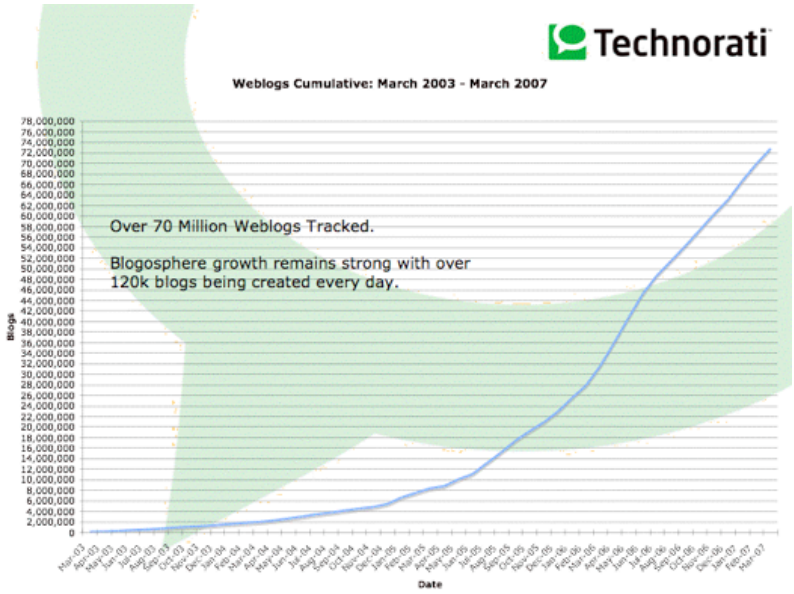
2. Blogosphere Experiment

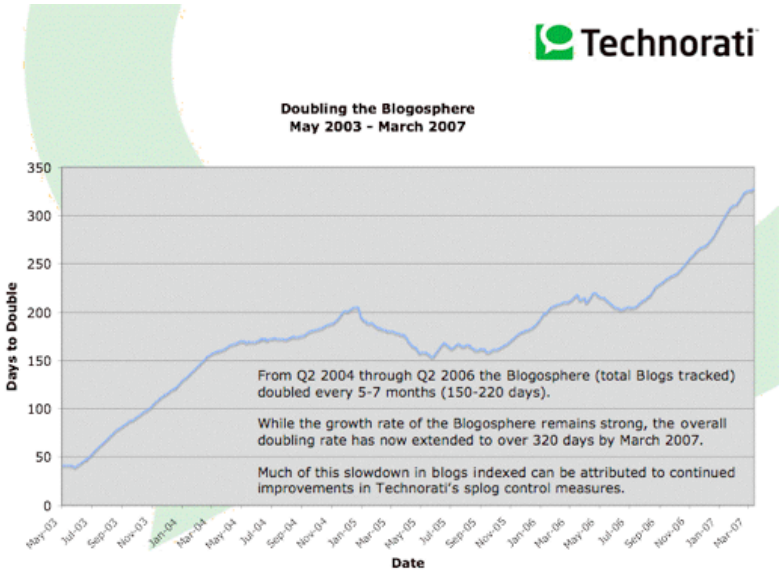


Data Mining Lab

Blogosphere

- The Blogosphere refers to the growing social network of people writing blogs, or web logs



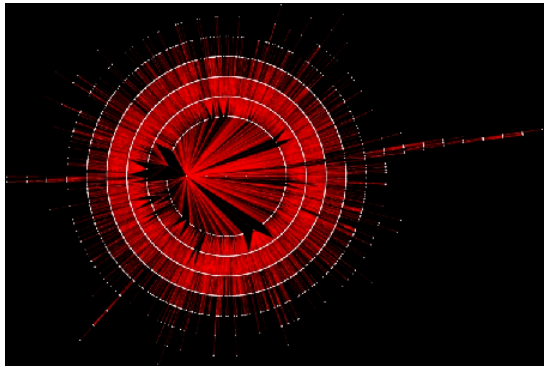


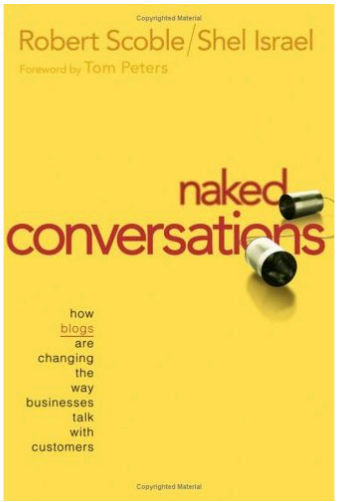
From "David Sifry's Alerts: State of the Live Web, April 2007" at: http://www.sifry.com/alerts/archives/000493.html

Data Mining Lab

Scobleizer's Blog List

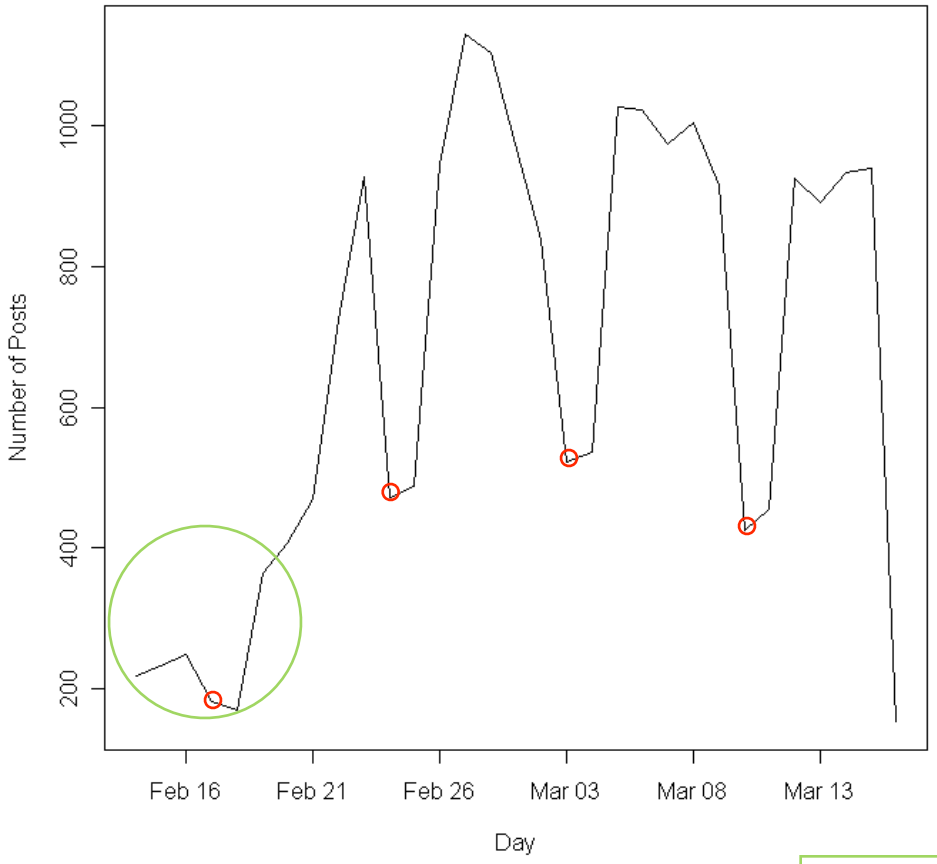
- Robert Scoble ("Scobleizer")
  - Blogger and book author
  - Technical evangelist (formerly with Microsoft)
- *Data Set Details:*
  - Scobleizer's reading list at Bloglines.com
  - *570 blogs*
  - *2380 bloggers*





Data Mining Lab

Data Set Statistics – Blog posts per day



Lack of data for all bloggers during first few days

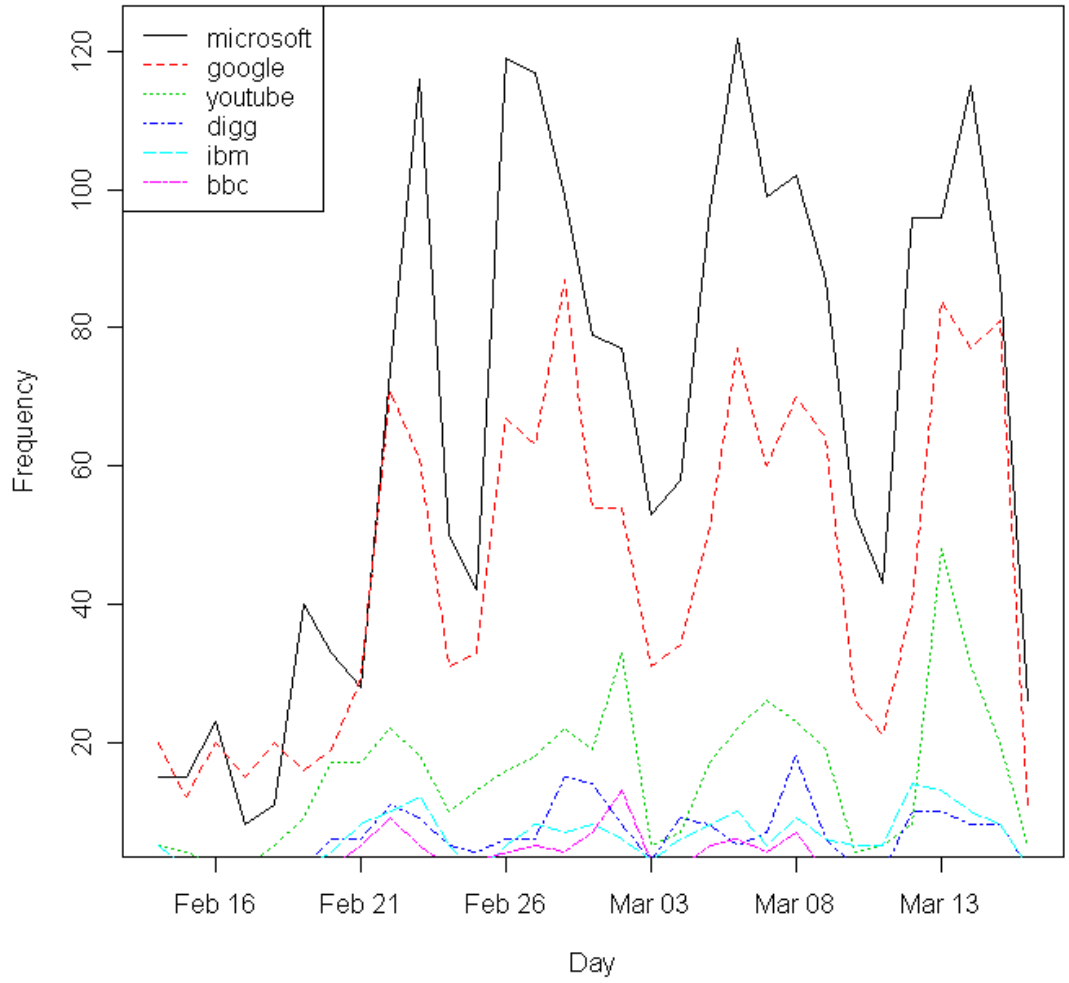
We observe fewer posts during the weekend (Friday & Saturday)

Data Mining Lab

Single Attribute: Companies

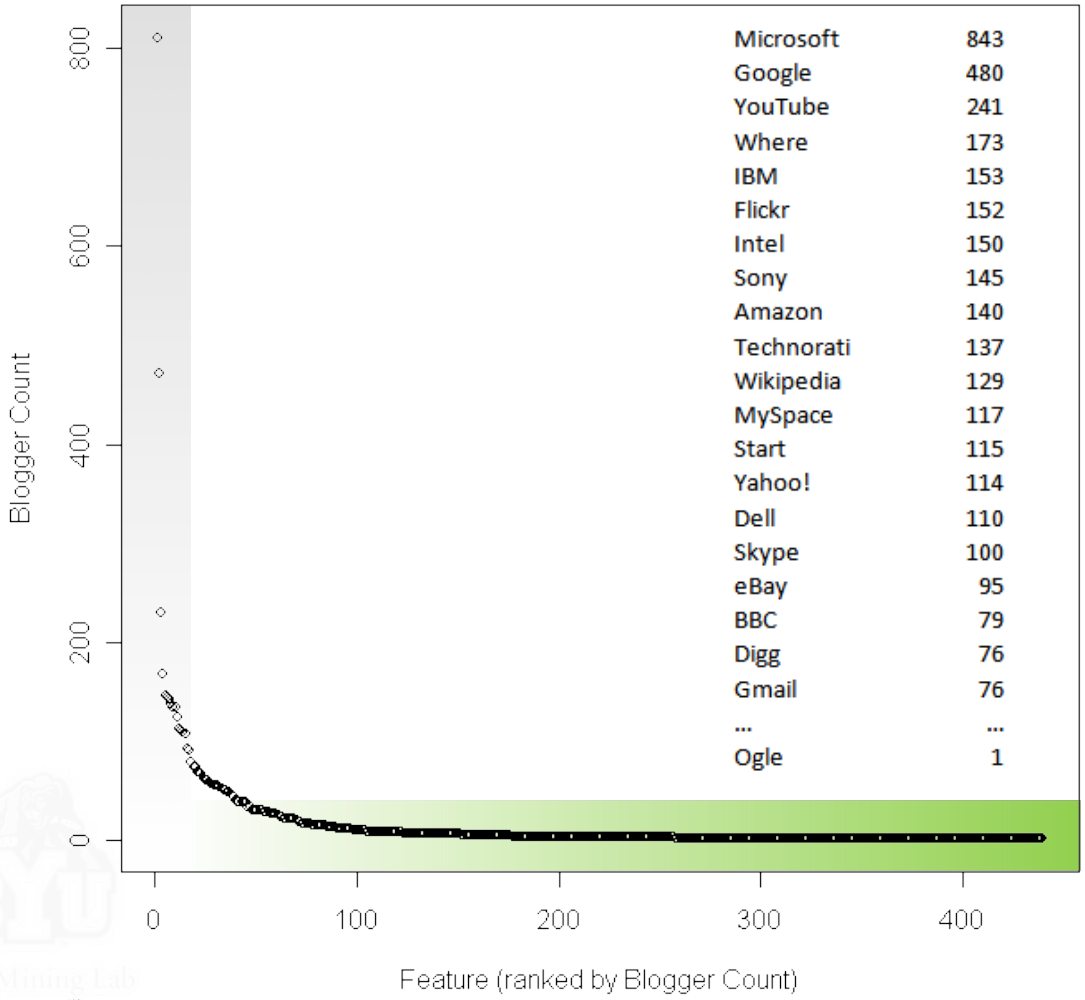
- Motivation
  - Many bloggers talk about various companies and what they are doing
- Methodology
  - Whenever a blogger mentions a company, it becomes a feature of the blogger
- Static company list used as attributes
  - 1,914 company names

Cyclic Feature Usage



Data Mining Lab

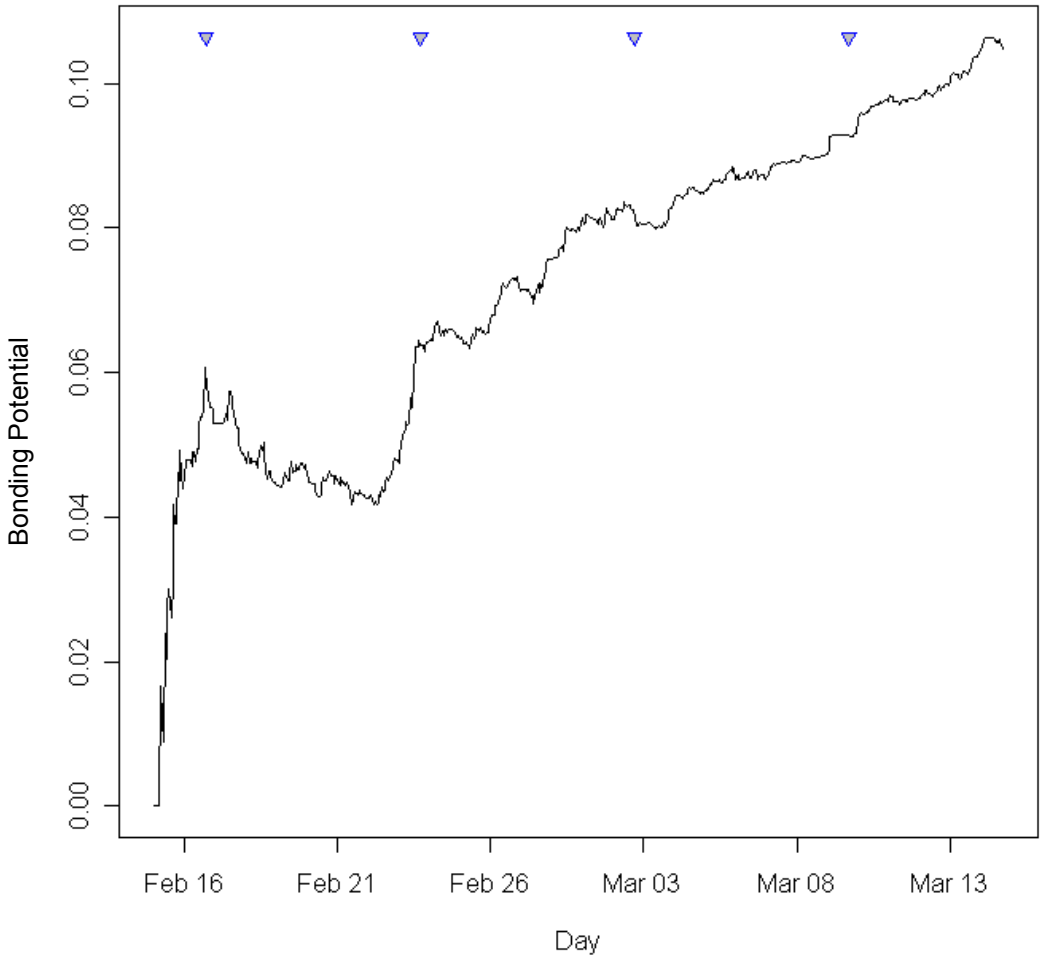
Power-law Behavior – Features



Observations

- Few companies
  - mentioned by many
- Many companies
  - mentioned by few

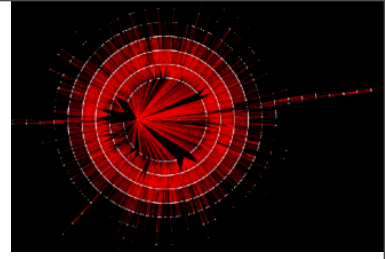
Blog Community Evolution

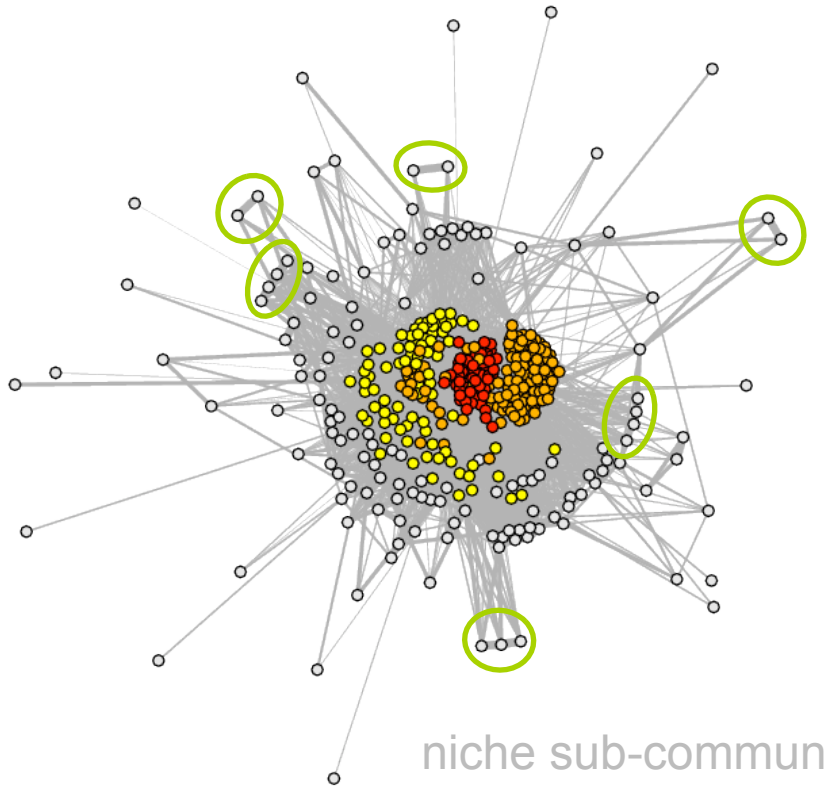


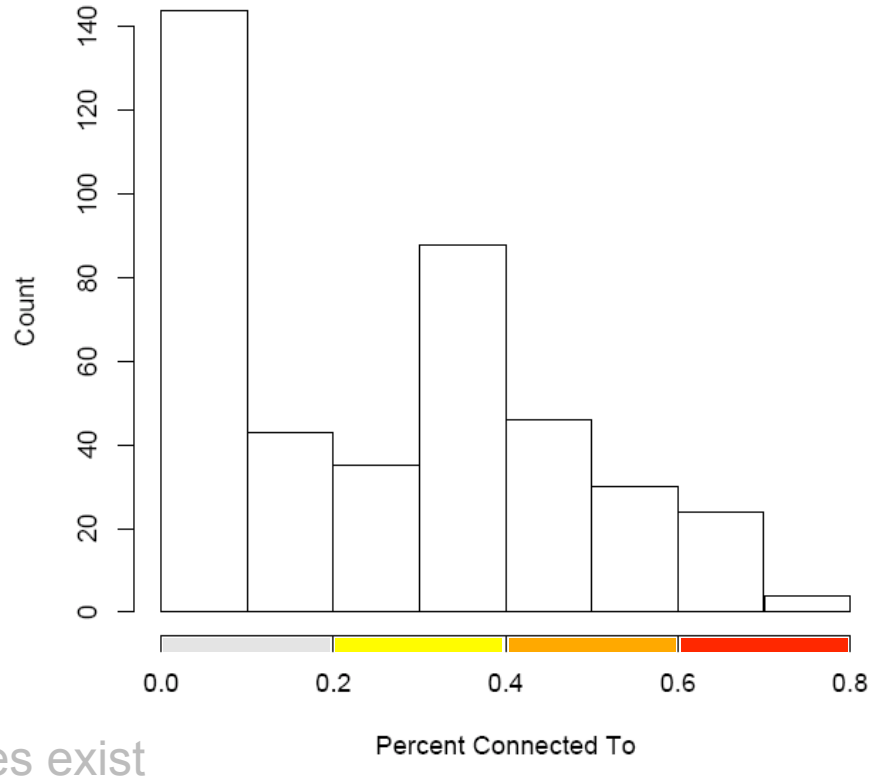
Observations

- Weekend bonding?
- Bridging indicates
  - newly used features
  - new bloggers
- Overall bonding (expected)
  - static set of features
  - no decay
- Blogosphere is full of buzz

Blog-based IAN – Feb. 24







niche sub-communities exist

Observations – Blog-based IAN

- Blog posts were cyclic within this community
  - Posted more during the week and less during the weekends
  - Interestingly, bonding occurs during the weekends
- Companies were mentioned in a power-law way
  - Few companies are mentioned often
  - Most companies are mentioned rarely
- Niche sub-communities
  - Bloggers focusing on long-tail companies were identified
- Blog-based IAN
  - Appears to follow power-law connectivity like ESNs

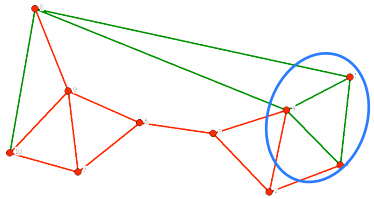
Conclusion

- Introduced Implicit Affinity Networks (IANs)
  - Individual-centered social networks
  - Capture dynamic, multi-faceted relationships
- Provided examples
  - Interests-based community
  - Blog-based community
- Introduced the notion of Social Capital
  - Measure evolving potential of community
  - Built around interests and blog content

Data Mining Lab

Future Work (In Progress)

- Hybrid Networks
  - IAN and ESN of the same community
  - Compute actual Social Capital
  - Analyze evolution (social capital vs. density)
- Refine implicit attribute extraction
  - Allow for dynamic feature extraction
  - Allow features to naturally decay with time
  - Use LDA to automatically extract “concepts”
- Experiment on domain-specific communities
  - Medical – patient communities
  - Political – jump start grass-roots campaigns



Data Mining Lab

Questions & Comments

Ask me now:

?

Email me:

Matthew Smith
smitty@byu.edu

Connect:

Web: **http://dml.cs.byu.edu/~smitty**
Blog: **http://dmine.blogspot.com**
LinkedIn: **http://linkedin.com/in/smitty**

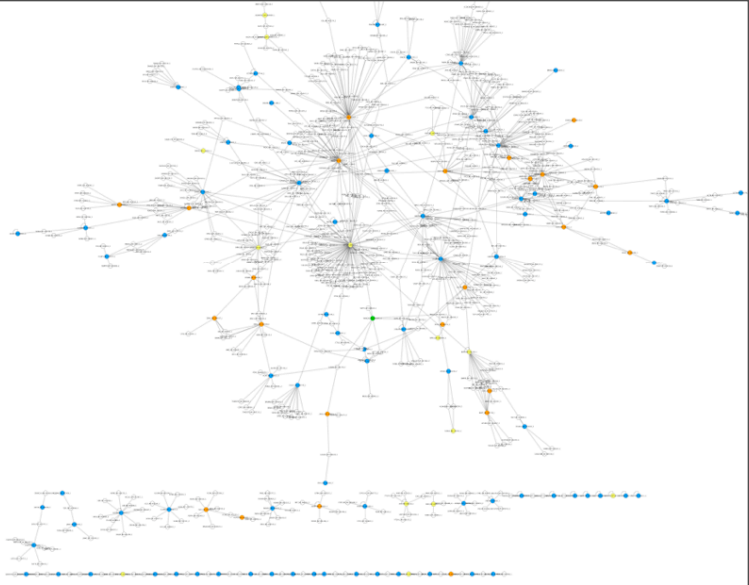
Data Mining Lab

Extra Slides

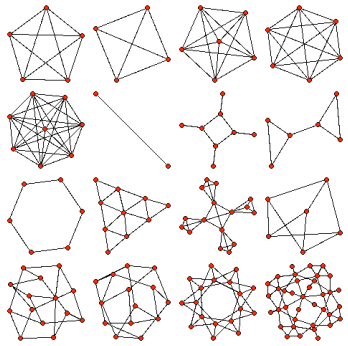
- Graphing Tools
- Hybrid Network

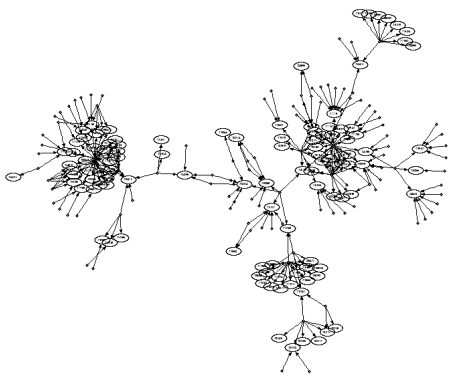
Data Mining Lab

Graphing Tools



- Cytoscape
  - http://www.cytoscape.org/
  - Very flexible, large graphs, many layouts
- R, igraph package
  - http://cneurocvs.rmki.kfki.hu/igraph/
  - Easily automated, less flexible, social networking algorithms
- GraphViz
  - http://www.graphviz.org/
  - Easily automated





Data Mining Lab

Hybrid Network

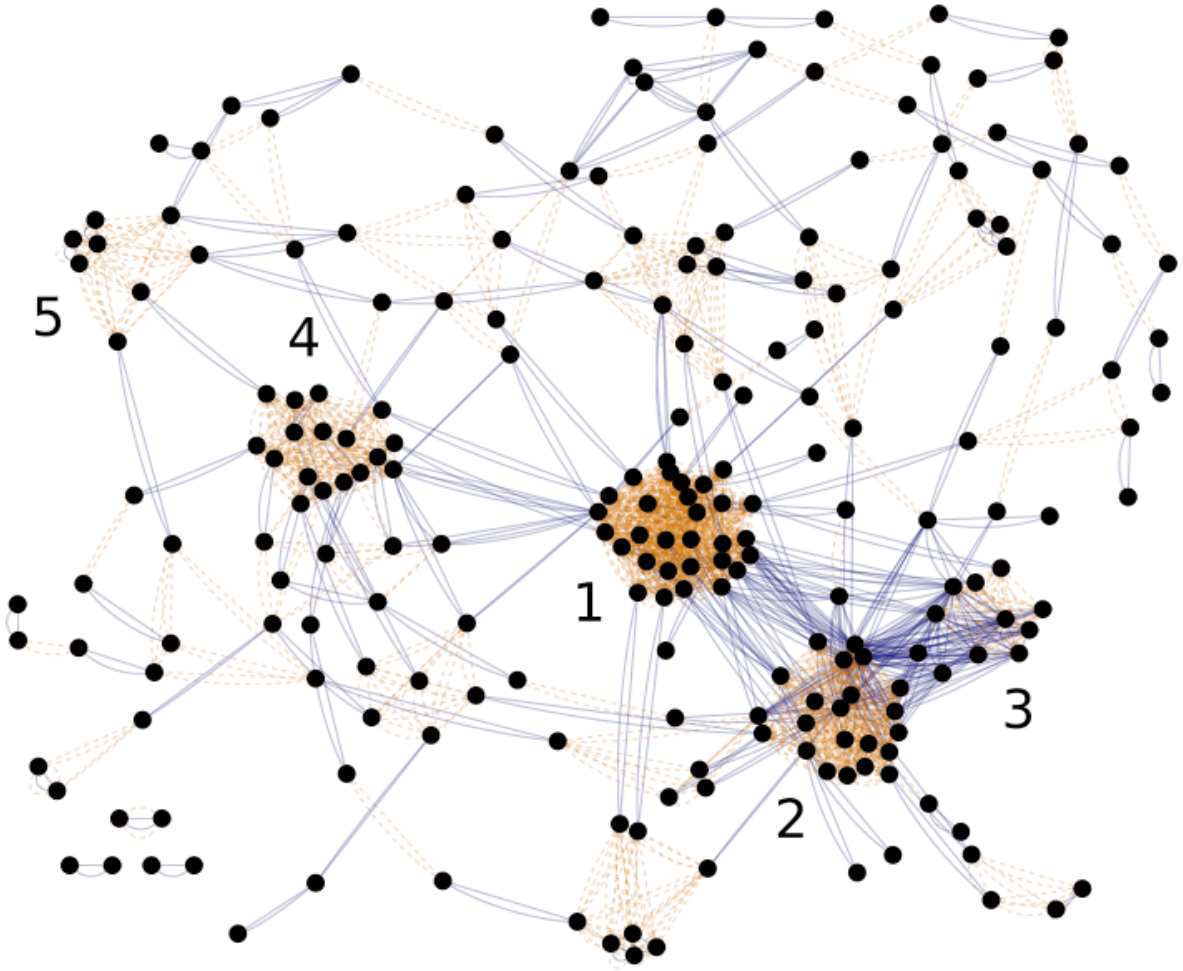


Figure 2: Hybrid Network for Blog Experiment (significant clusters are labeled 1-5)

Data Mining Lab