Dynamic harvesting of social graphs

J. I. Orlicki[1]

[1]ITBA

WDCS 2009, Buenos Aires, Argentina

Mini Sci Bio

- Little background in theoretical computer science. Reversible Turing machines (?).
- Little background in applied security research: information gathering and attack simulation.
- Working with Phd advisors J. I. Alvarez-Hamelin and P. I. Fierens at ITBA on reputation and privacy in social networks.

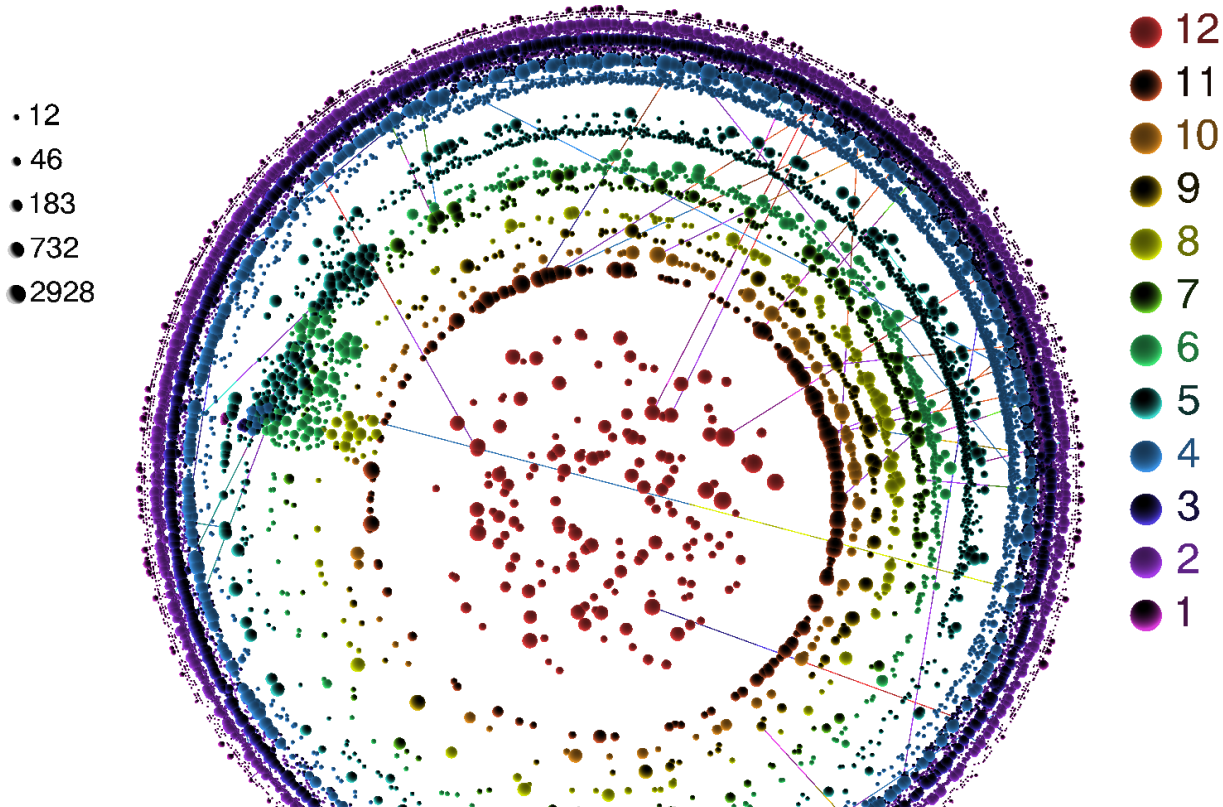
Agenda

- What?
  - Answer: studying social graphs/networks dynamics
- How?
  - Answer: collecting dynamic social data
- When?
  - Answer: 2009-2010, very interesting questions arise.
- Dangers?
  - "*There are fields, endless fields, where human beings are no longer born, we are grown.*" Morpheus, Matrix (1999)

The Social Graph (What?)

- Is a real object! We use graphs as abstractions.
- (notation) Social graph: the social interactions between human beings.
- (notation) Social network (service): a communication system designed for using and expanding the social graph.

http://bit.ly/orkut-k-cores-big



Social Network Services (What?)

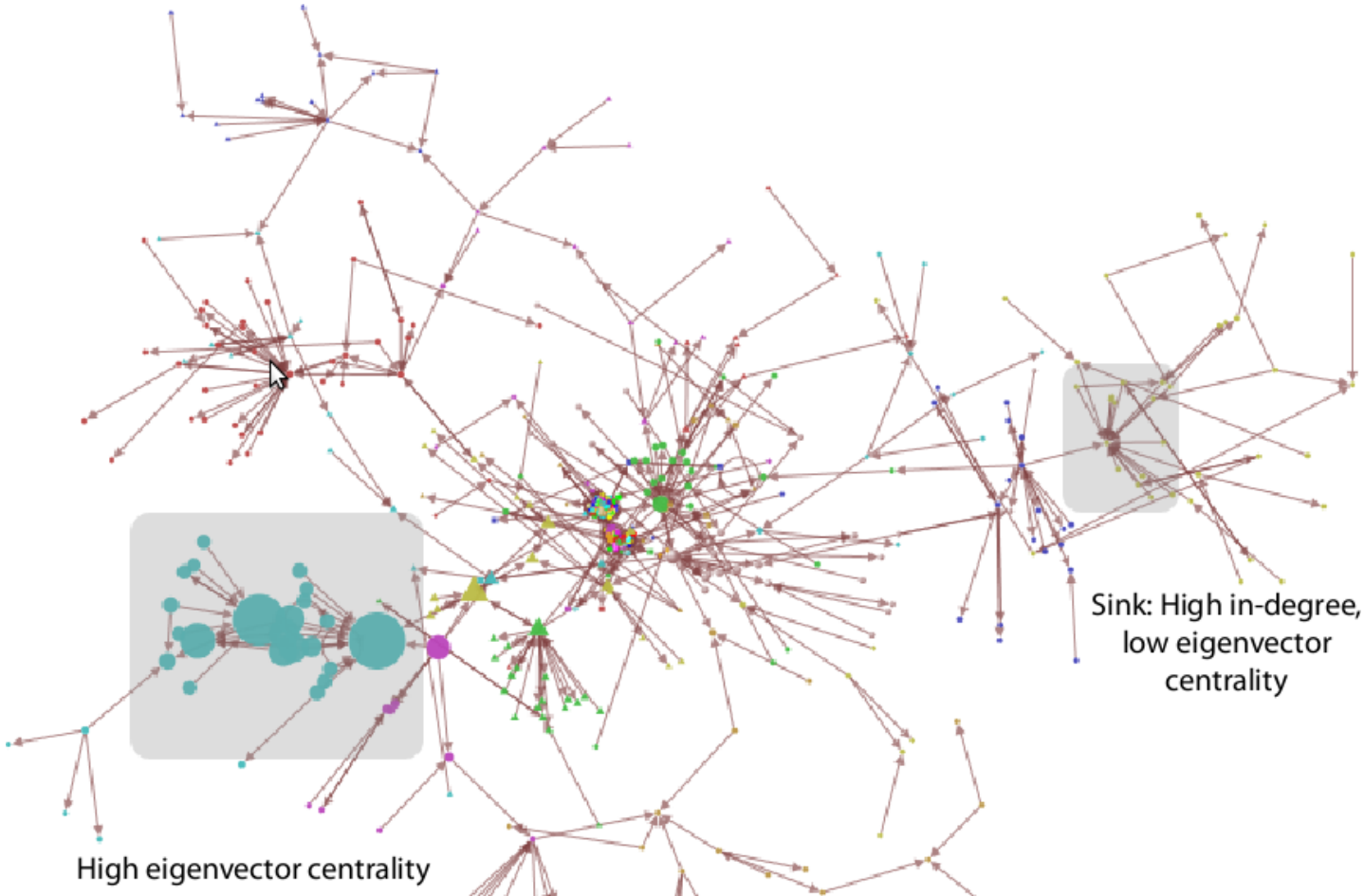
1. Provide abstraction of user profile and social contacts.
2. Provide communication with social contacts.
3. Provide ways to stay in touch with contacts.
4. Provide navigation and recomm. of new contacts.

parameter	*Cyworld*	*Flickr*	*LiveJournal*	*Orkut*
#nodes	*12,048,186*	*1,846,198*	*5,284,457*	*3,072,441*
%nodes	*100%*	*26.9%*	*95.4%*	*11.3%*
#links	*190,589,667*	*22,613,981*	*77,402,652*	*223,534,301*
%simmetric links	*100%*	*62.0%*	*73.5%*	*100.0%*
ave. deg $\langle k \rangle$	*31.6*	*12.24*	*16.97*	*106.1*
in/out α	-	*1.74 / 1.78*	*1.59 / 1.65*	*1.50 / 1.50*
clustering C	*0.16*	*0.31*	*0.33*	*0.17*
assortativity r	*-0.13*	*0.20*	*0.18*	*0.07*
ave. distance	*3.2*	*5.67*	*5.88*	*4.25*
diameter	-	*27*	*6*	*9*

analysis from [Mislove et al., 2007] and [Ahn et al., 2007]

Social Network Services, cont. (What?)

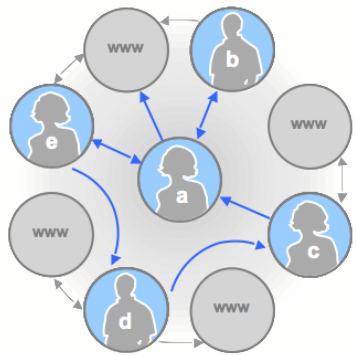
- Micro/nano-blogging: network simplicity, short messages, rapid diffusion.
- Dynamics: weighted mention networks [Shamma et al., 2009] and retweeting dynamics (message forwarding). Example: Twitter, Jaiku, etc.



Openness versus Privacy (What?)

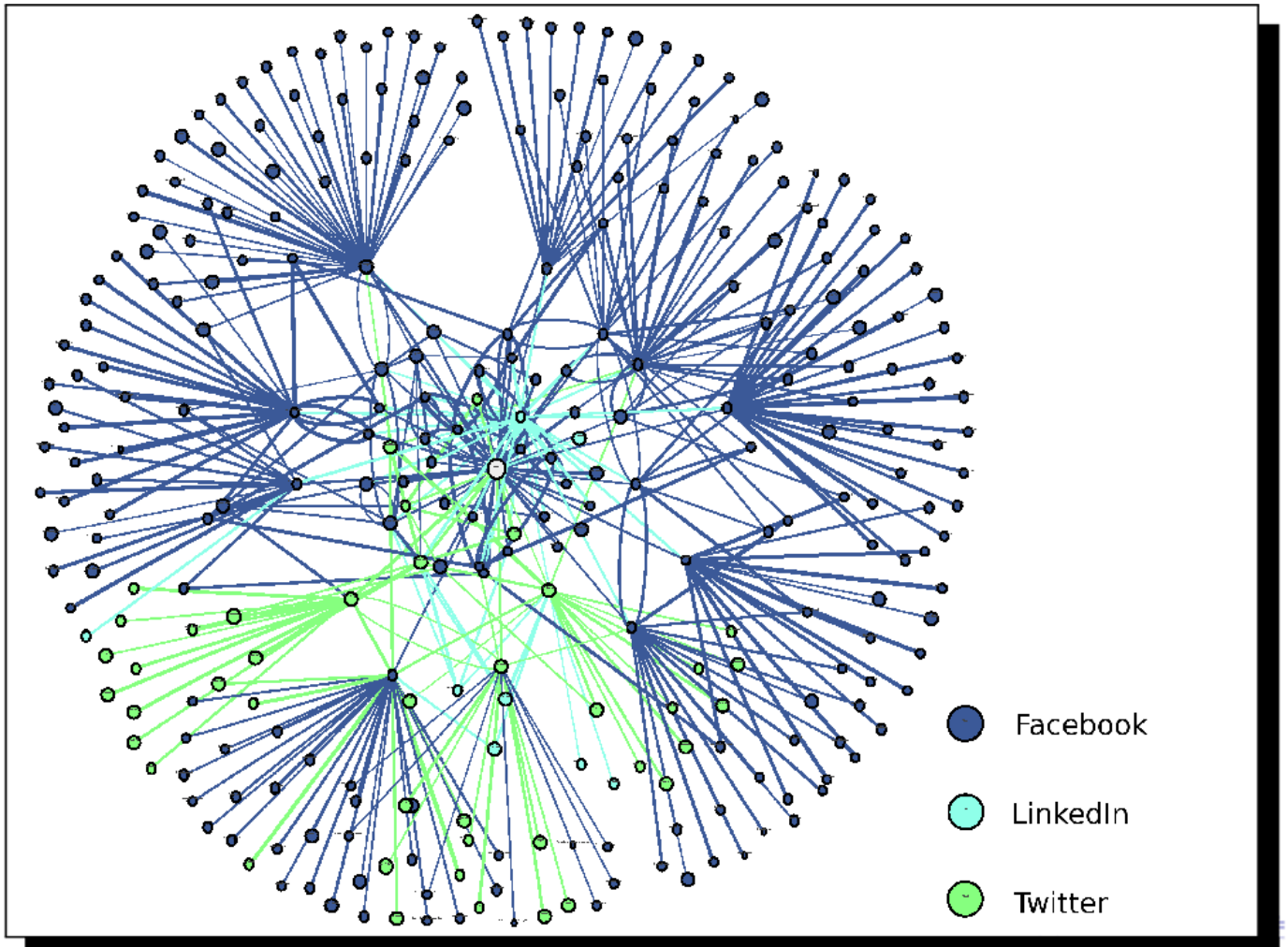
- Social networks are means of communication, so are useful.
- Tension between those who like privacy and those who like openness.
- Limited visibility usually via limited lookahead (the infamous Only Friends vs. Friends of Friends features in Facebook).
- Example of privacy: LinkedIn website, private data and limited lookahead (distance 3).

- Example of openness: Google SocialGraph [Fitzpatrick, 2007], public interface.



Tools for dynamic harvesting (How?)

- *Exomind* [JIO, 2008] (`http://bit.ly/exomind`): supported by *Core ST*, mostly focused in static OSINT and social infiltration. Includes dynamic updates Normalized Google Distance (NGD) and other search engine mining information.



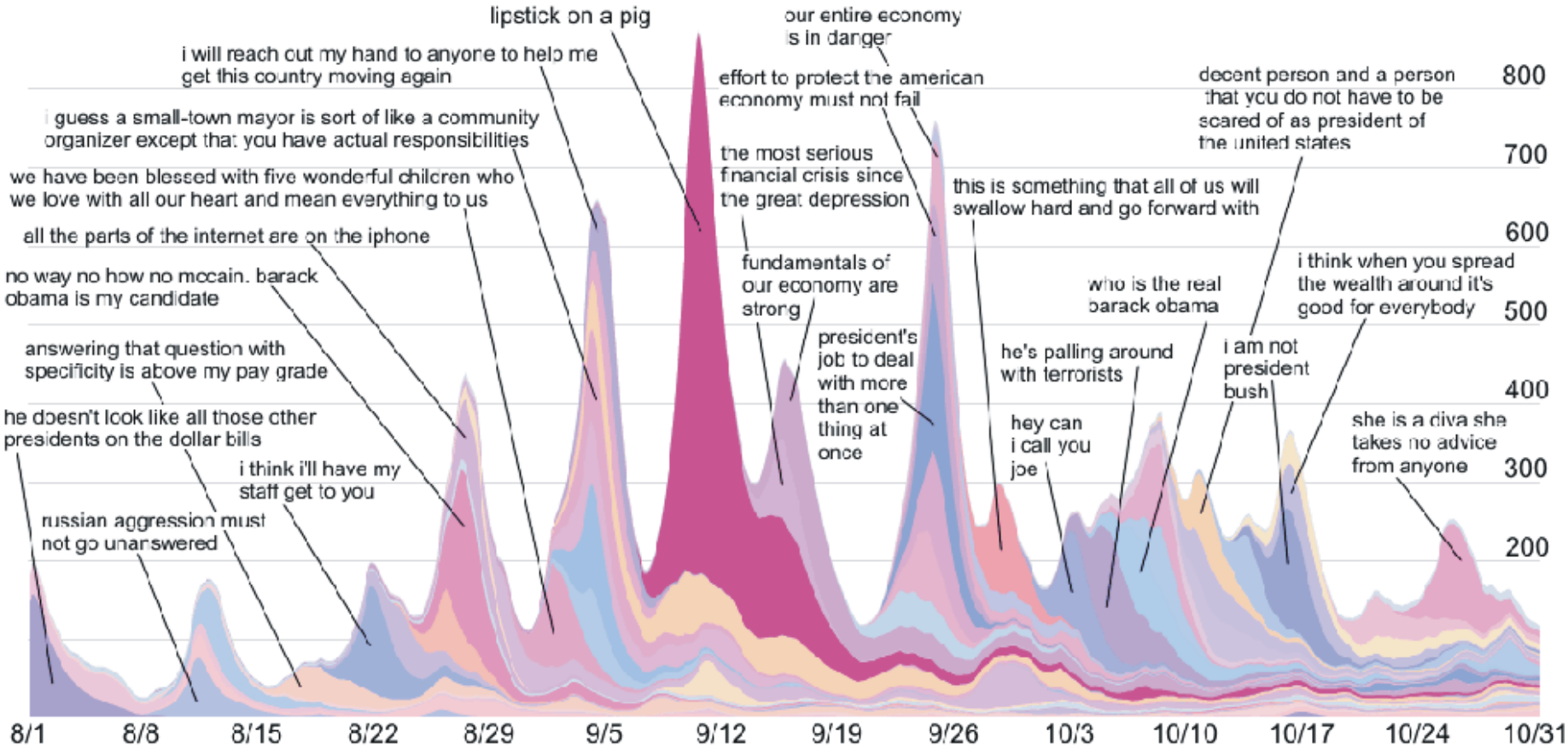
Tools for dynamic harvesting, cont. (How?)

- *Tuplets* [JIO, 2009] (`http://tuplets.appspot.com`): an experimental cloud shell for social network crawling and interaction (Web 4.0?).
- Nano-blogging weighted mention links:

```
>>>> twitter @asdfasdf X ?
yes
twitter @asdfasdf @qwerqwer1 (33.9 %)
twitter @asdfasdf @qwerqwer2 (10.7 %)
twitter @asdfasdf @qwerqwer3 (8.9 %)
twitter @asdfasdf @qwerqwer4 (7.1 %)
twitter @asdfasdf @qwerqwer5 (7.1 %)
twitter @asdfasdf @qwerqwer6 (5.4 %)
twitter @asdfasdf @qwerqwer7 (5.4 %)
...
>>>> socialgraph asdfasdf.blogspot.com X ?
...
>>>> alpha “argentina external debt” X ?
...
>>>> alpha “argentina patents” X ?
...
>>>> alpha "the meaning of life the universe and everything"
```

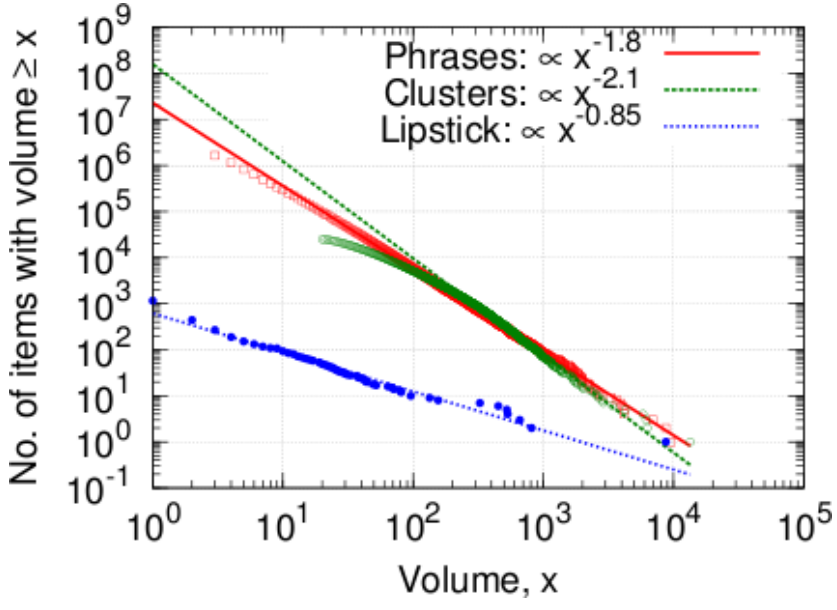
Topics: News Cycle (When?)

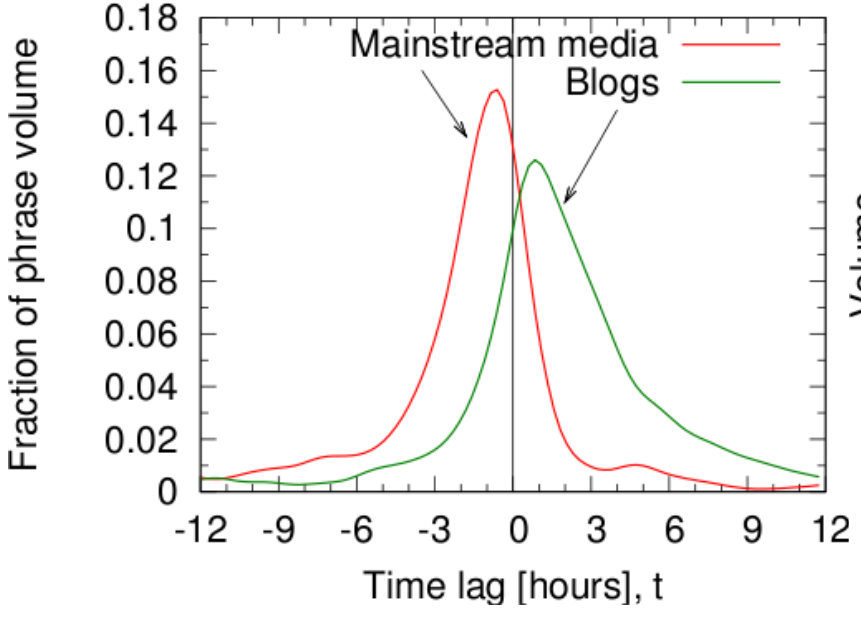
- News diffusion by news sources [Leskovec et al., 2009].
- Study of memes/phrases in blog and traditional news media.
- Analysis of 90 million articles to extract 112 million quotes and generate 35,800 phrase clusters. `http://memetracker.org/`

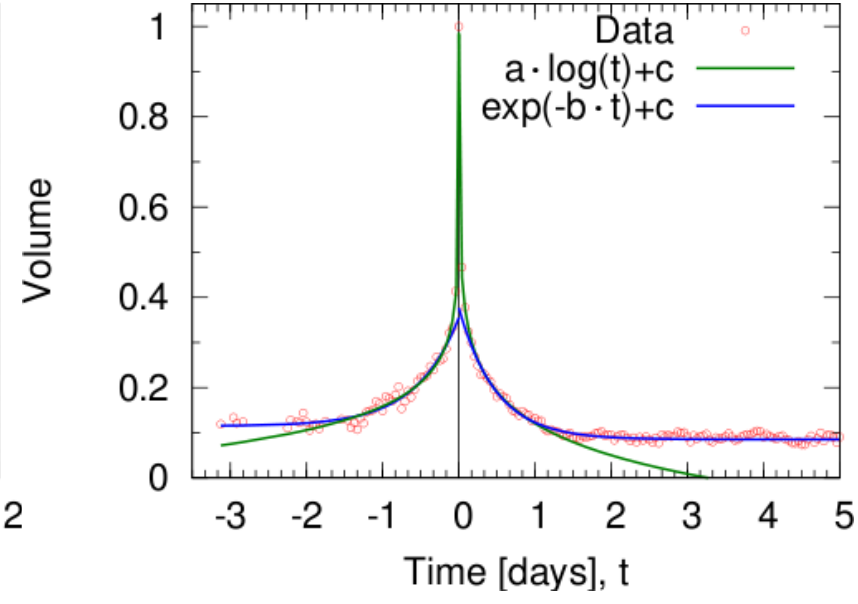


Topics: News Cycle, cont. (When?)

Some curious properties they found:





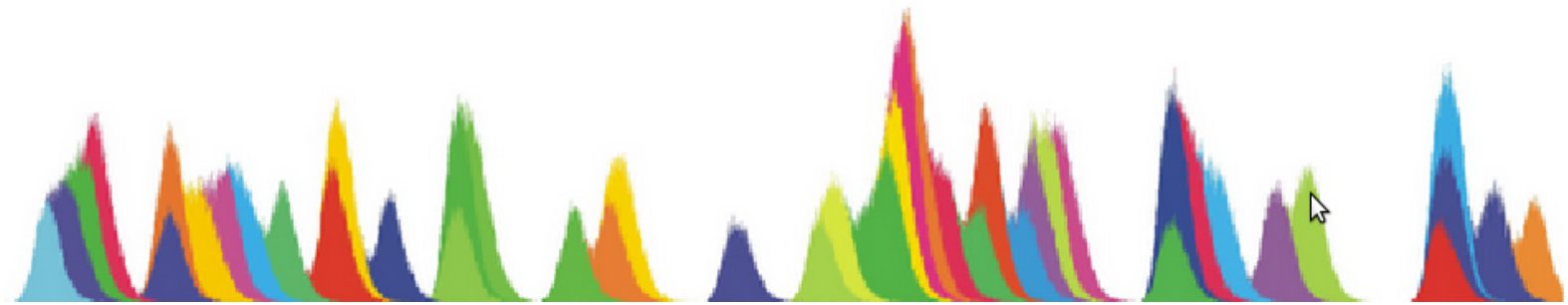


Topics: News Cycle, cont. (When?)

- Peak volume of memes/phrase and decay: $y(t) \sim a\log(t)$
- Model: periods $t = 1, 2, \ldots, T$, with N media sources reporting one thread per period, one new thread is generated each period and all source report different threads at $t = 0$.
- Simple pref. attachment:

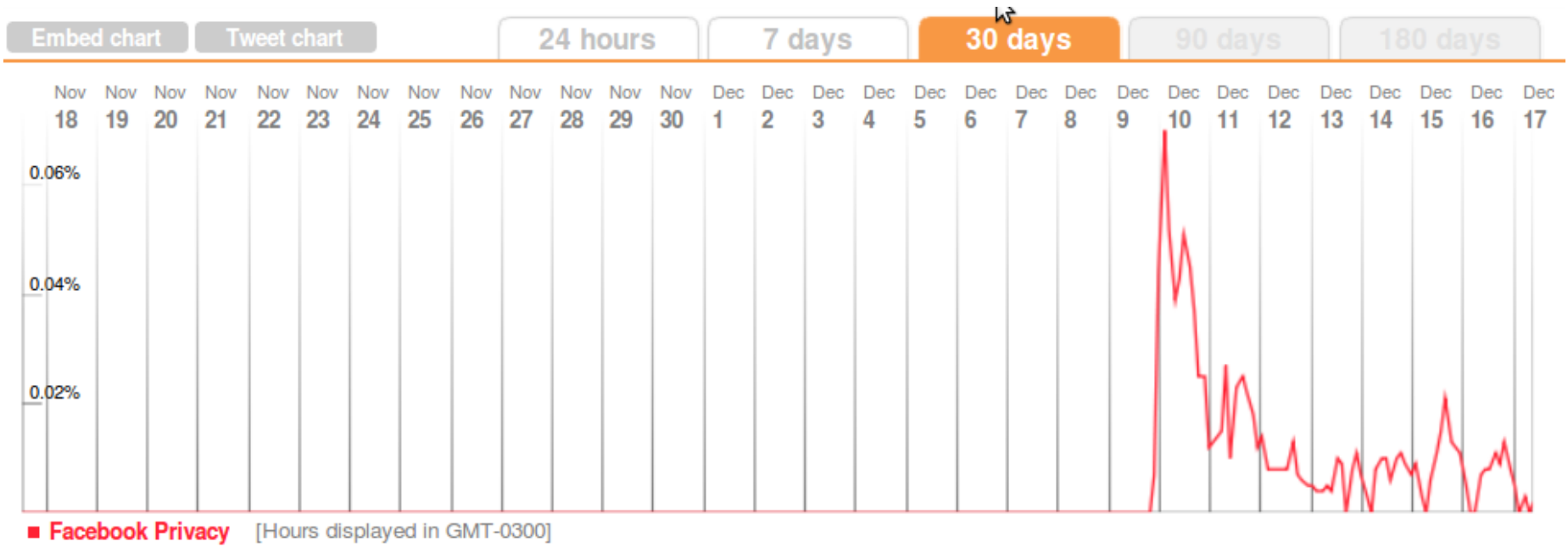
$$Pr(j,t) = f(n_j)\delta(t - t_j)$$

  where n_j is the total previous stories about thread j and t_j is when thread/history j was first produce .
- $f(\cdot)$ mono. increasing on n_j (imitation effect)
- $\delta(\cdot)$ mono. decreasing in $t - t_j$ (recency effect).



Topics: News Cycle, cont. (When?)

In a nano-blogging environment Trendistic (by Flaptor), collect and show similar meme/phrase peak behavior (`http://trendistic.com`).
Example: `facebook privacy` (last 30 days)



Topics: Nano-blogging economy, cont. (When?)

- WhuffieBank (http://whuffiebank.org) discussion: information replication = money = reputation (?).
- Each retweet is a replication of a message by another node/user. Local decision: +1 whuffie to @userA account if @userX retweets @userA.
- Criticism? Correlation with $\langle k_{in} \rangle$? A fake Deepak Chopra is #12 ?!
- In Egg-O-Matic prototype [JIO et al., 2009] small-scale experiments where PageRank with datasets of photographic and video social networks, *good* faceted results.

Conclusions

- Please analyze relation of information diffusion with epidemics.
- Can we harvest and interact with the social nano-blogging systems in a way to generate useful ad-hoc economies?
- Can we identify trendy people or people generating the trend?
- Can we discriminate ephemeral from persistent trends?
- Shall we prefer an implicit reputation systems including eigenvector centrality (such as PageRank) to more explicit/ad-hoc money economies?
- How privacy restrictions affect information diffusion?

Bibliography

[Mislove et al., 2007] Mislove, Marcon, Gummadi, Druschel and Bhattacharjee, "*Measurement and analysis of online social networks*".
`http://www.imconf.net/imc-2007/papers/imc170.pdf`

[Ahn et al., 2007] Ahn, Han, Kwak, Moon and Jeong, "*Analysis of topological characteristics of huge online social networking services*". `http://www2007.org/papers/paper676.pdf`

[Shamma et al., 2009] Shamma, Kennedy and Churchill, "*Tweet the Debates*".
`http://research.yahoo.com/files/wsm01a-shamma.pdf`

[Fitzpatrick, 2007] Fitzpatrick, "*Thoughts on the Social Graph*".
`http://www.bradfitz.com/social-graph-problem/`

[Leskovec et al., 2009] *Leskovec, Backstrom and Kleinberg,"Meme-tracking and the Dynamics of the News Cycle".*
`http://www.cs.cornell.edu/home/kleinber/kdd09-quotes.pdf`

[JIO et al., 2009] Orlicki, Alvarez-Hamelin and Fierens, "*Faceted Ranking of Egos in Collaborative Tagging Systems*".
`http://arxiv.org/pdf/0809.4668`