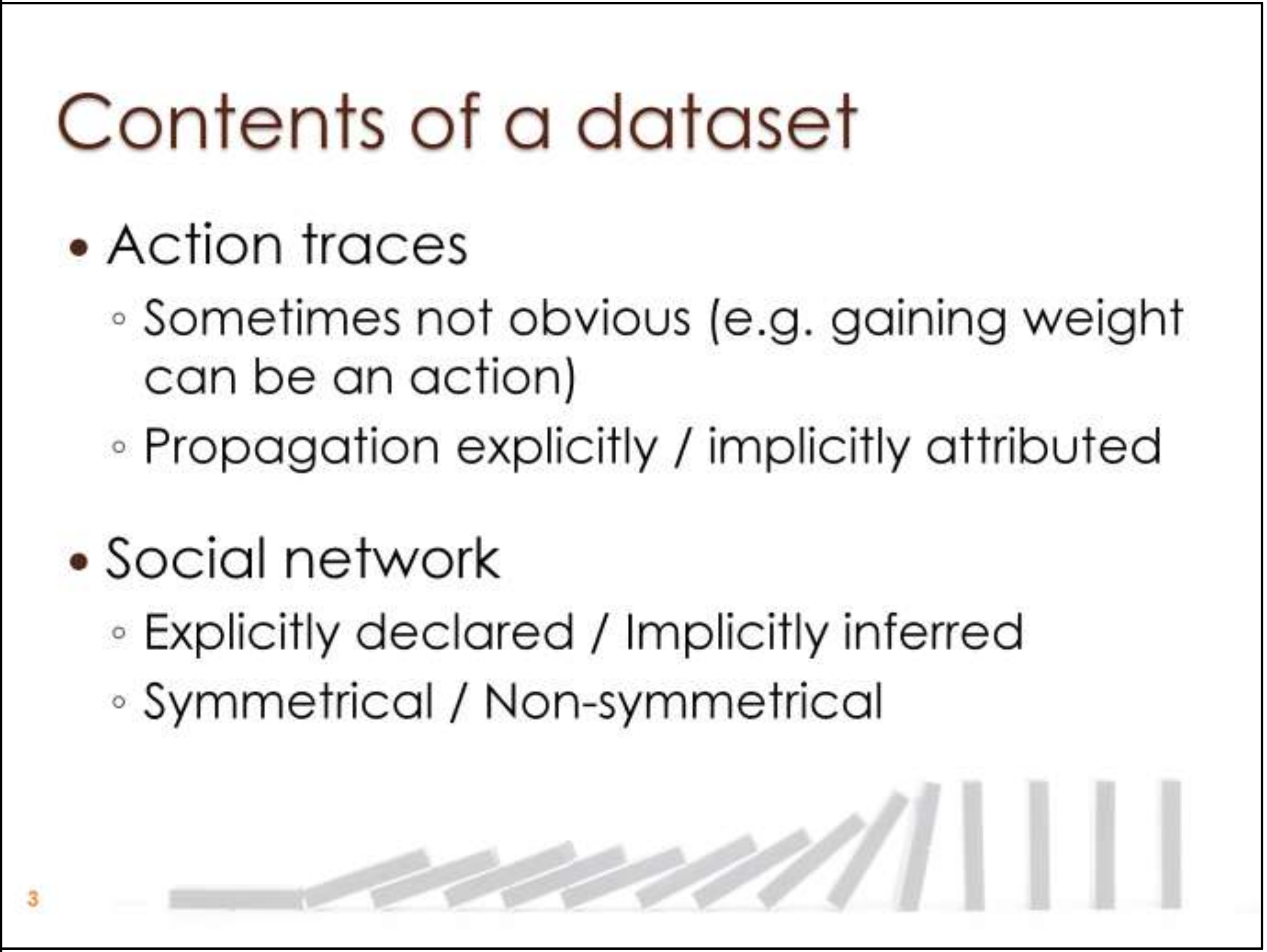
Part I → Part II → Part III → Part IV → Part V

Data and Software

Part II: Outline

- Types of datasets
- Propagation of information “memes”
- Propagation of other actions
- Synthetic datasets
- Software tools

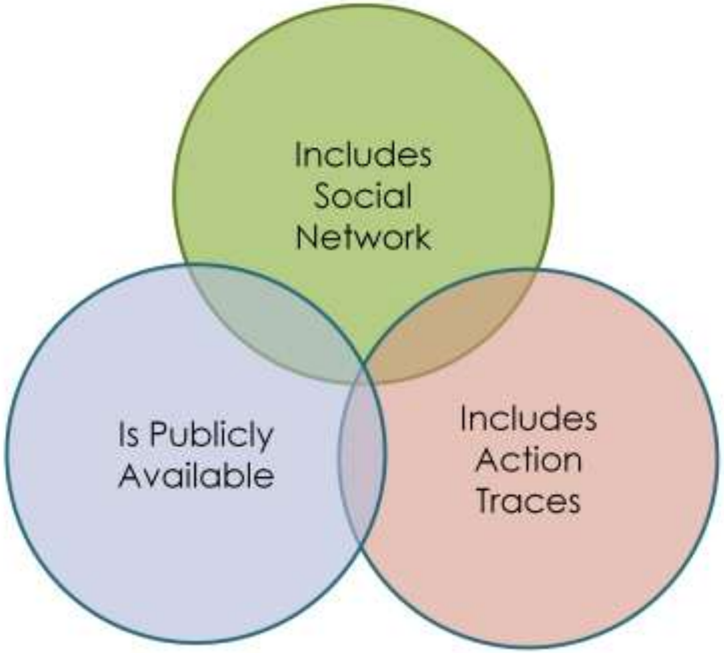


Contents of a dataset

- Action traces
  - Sometimes not obvious (e.g. gaining weight can be an action)
  - Propagation explicitly / implicitly attributed
- Social network
  - Explicitly declared / Implicitly inferred
  - Symmetrical / Non-symmetrical

Data availability limits research

- Often you have to pick two of these

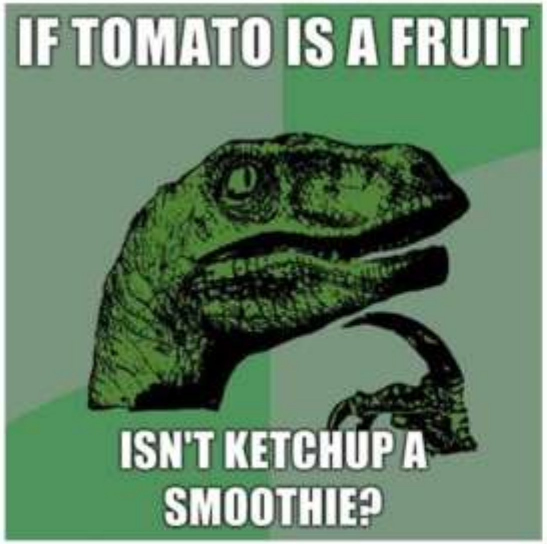


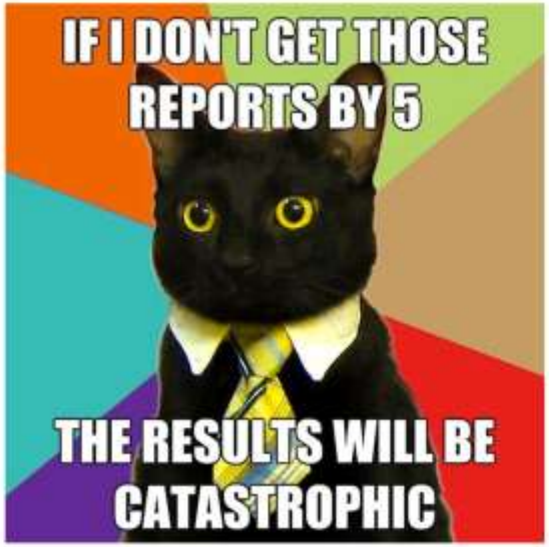
Classification: according to availability

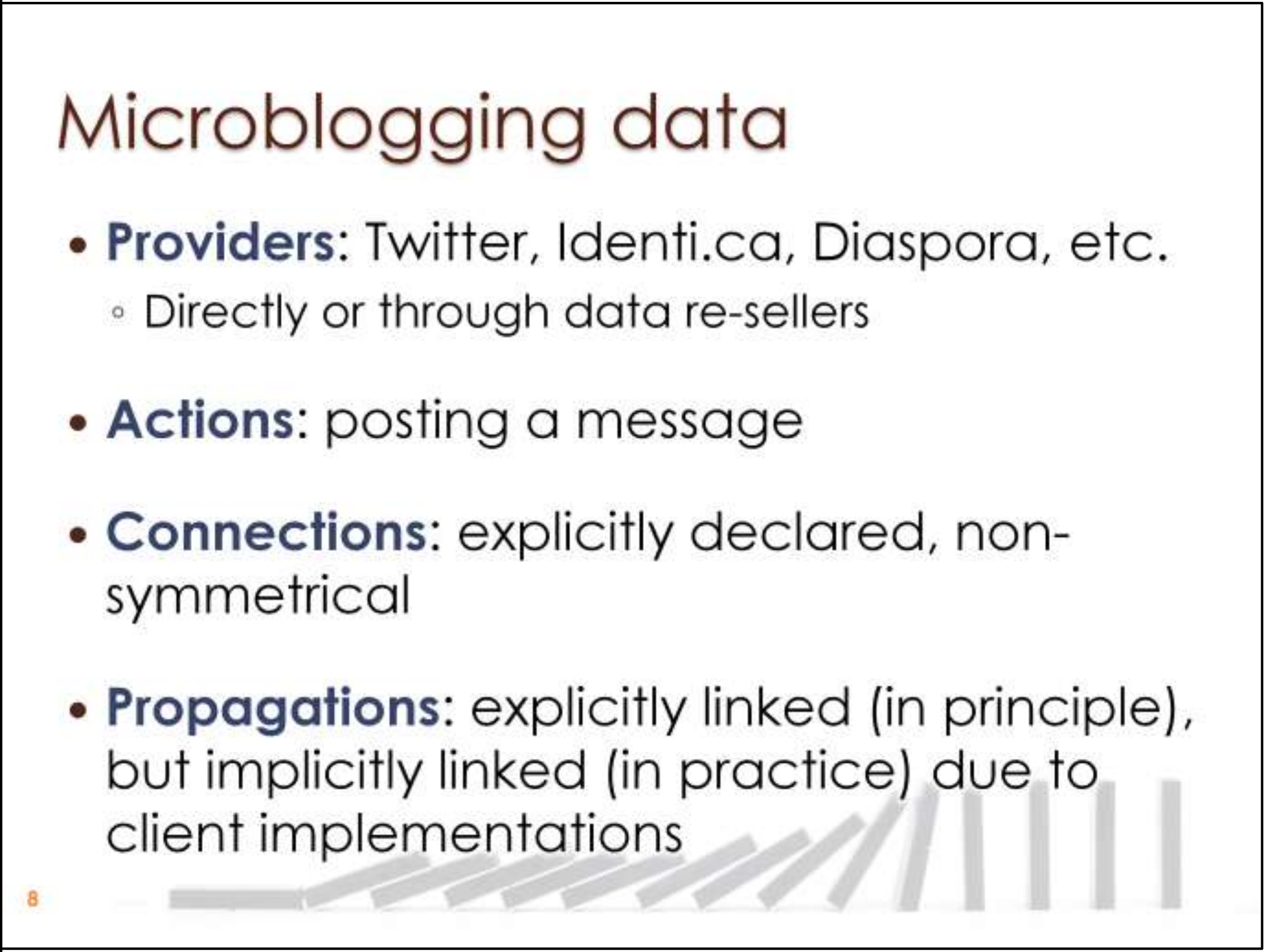
- Proprietary, impossible or very hard to reproduce (e.g. shopping history in e-commerce)
  - Increasingly being rejected in IR, DM communities
- Proprietary, reproducible (e.g. web crawl of a sub-set of public websites)
- Existing open dataset
- New open dataset

Propagation of Information
“Memes”

Memes and "Internet Memes"

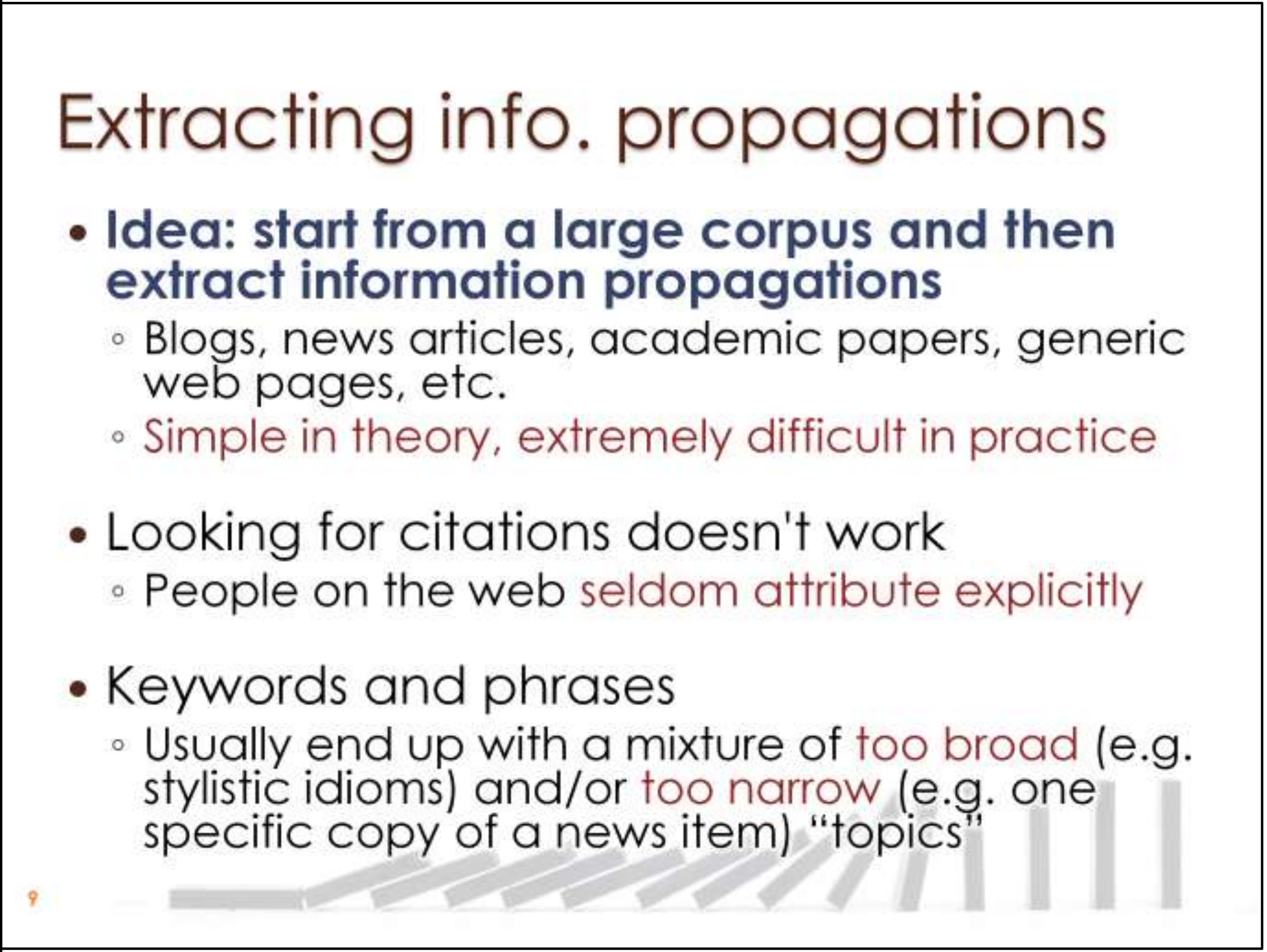



IF I DON'T GET THOSE REPORTS BY 5
THE RESULTS WILL BE CATASTROPHIC



Microblogging data

- **Providers**: Twitter, Identi.ca, Diaspora, etc.
  - Directly or through data re-sellers
- **Actions**: posting a message
- **Connections**: explicitly declared, non-symmetrical
- **Propagations**: explicitly linked (in principle), but implicitly linked (in practice) due to client implementations



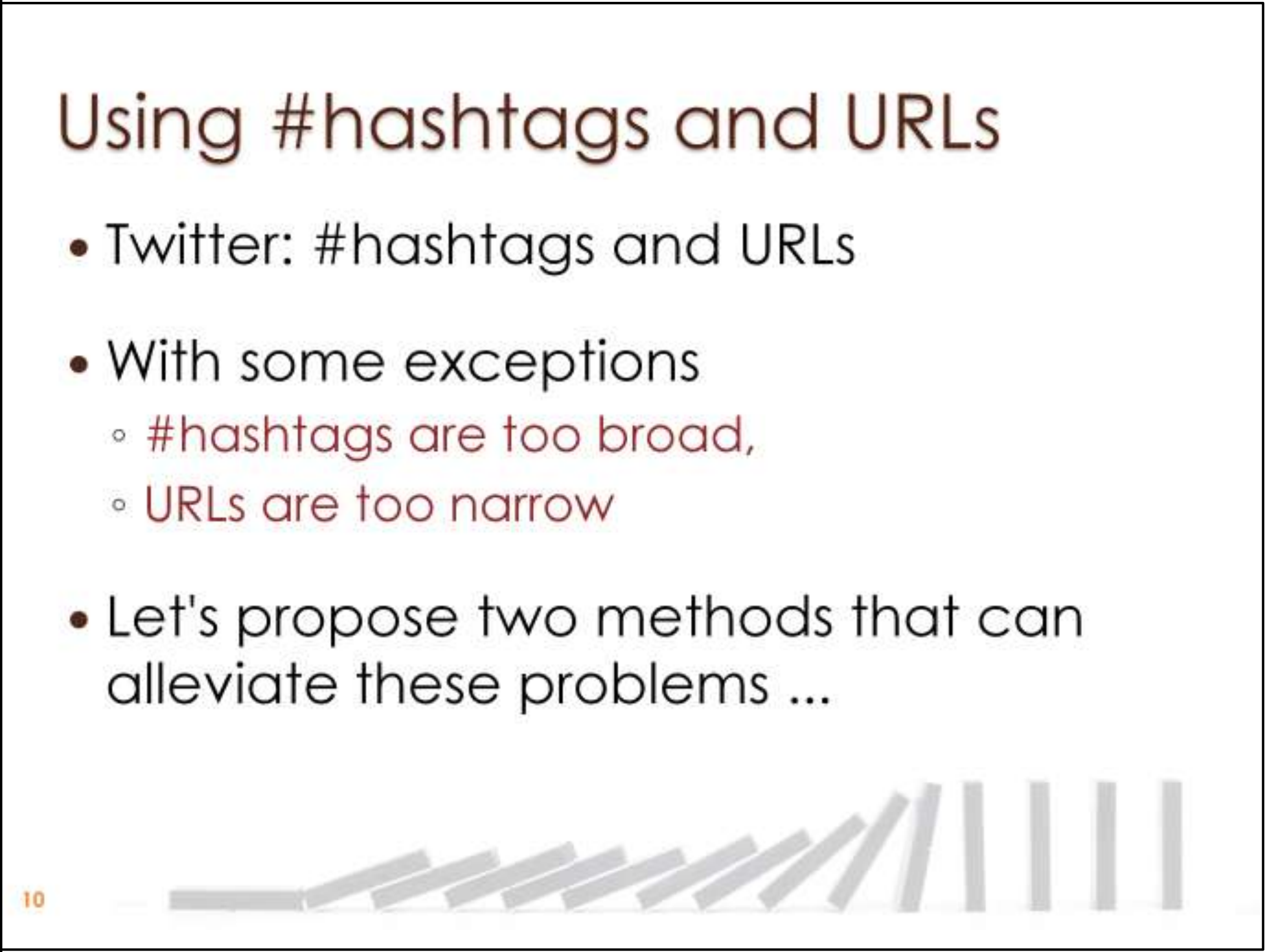
Extracting info. propagations

- **Idea: start from a large corpus and then extract information propagations**
  - Blogs, news articles, academic papers, generic web pages, etc.
  - Simple in theory, extremely difficult in practice
- Looking for citations doesn't work
  - People on the web seldom attribute explicitly
- Keywords and phrases
  - Usually end up with a mixture of too broad (e.g. stylistic idioms) and/or too narrow (e.g. one specific copy of a news item) "topics"

Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins:
Information diffusion through blogspace.
WWW 2004
http://doi.acm.org/10.1145/988672.988739

Eytan Adar and Lada Adamic:
Tracking information epidemics in blogspace.
Web Intelligence 2005
http://dx.doi.org/10.1109/WI.2005.151

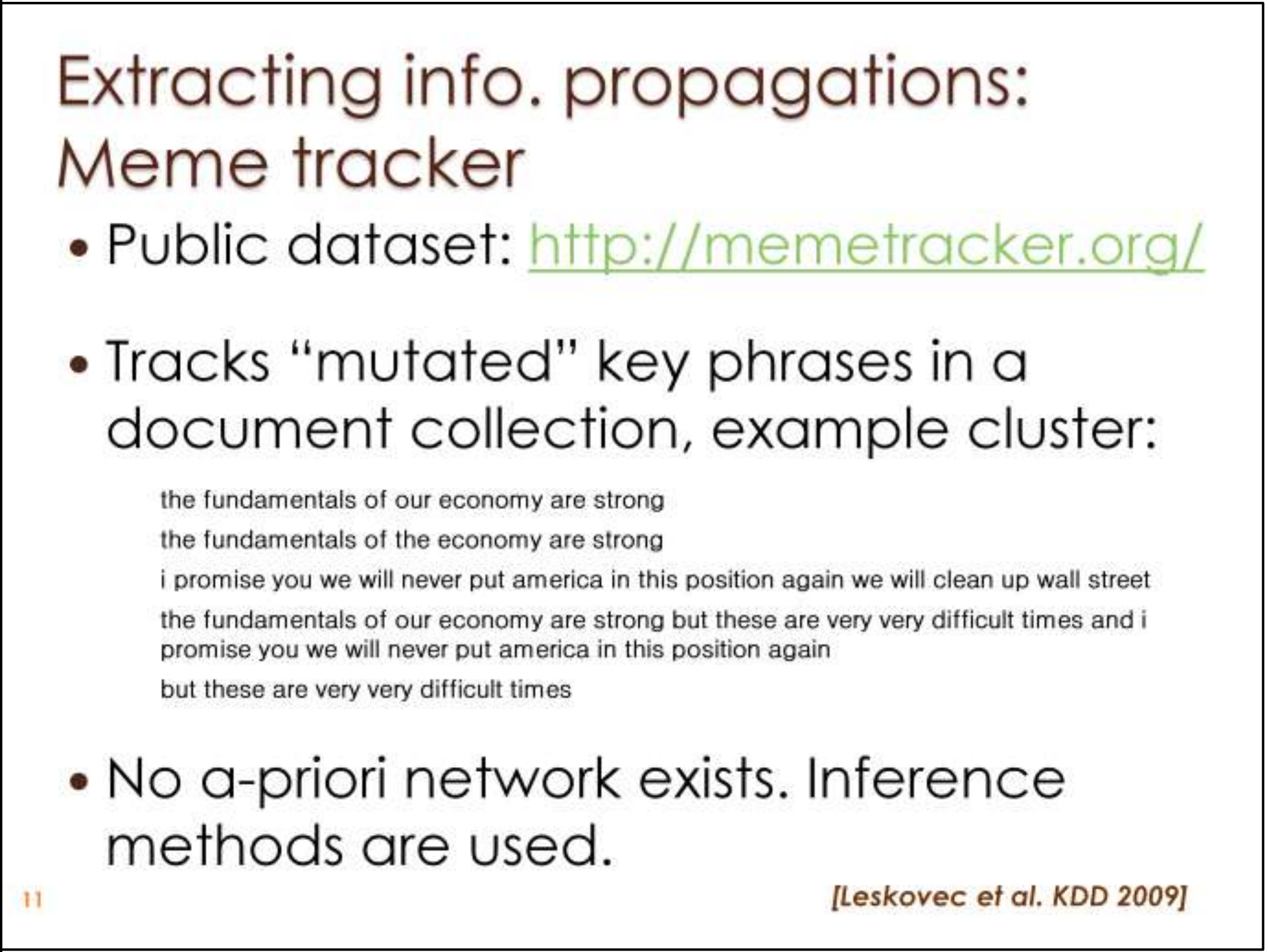
Ramesh Maruthi Nallapati, Xiaolin Shi, Daniel McFarland, Jure Leskovec, Daniel Jurafsky:
LeadLag LDA: Estimating Topic Specific Leads and Lags of Information Outlets.
ICWSM 2005
http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2746



Using #hashtags and URLs

- Twitter: #hashtags and URLs
- With some exceptions
  - #hashtags are too broad,
  - URLs are too narrow
- Let's propose two methods that can alleviate these problems ...

Shaomei Wu, Chenhao Tan, Jon Kleinberg and Michael Macy:
Does Bad News Go Away Faster?
ICWSM 2011
http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2877



Extracting info. propagations: Meme tracker

- Public dataset: http://memetracker.org/
- Tracks "mutated" key phrases in a document collection, example cluster:

  the fundamentals of our economy are strong

  the fundamentals of the economy are strong

  i promise you we will never put america in this position again we will clean up wall street

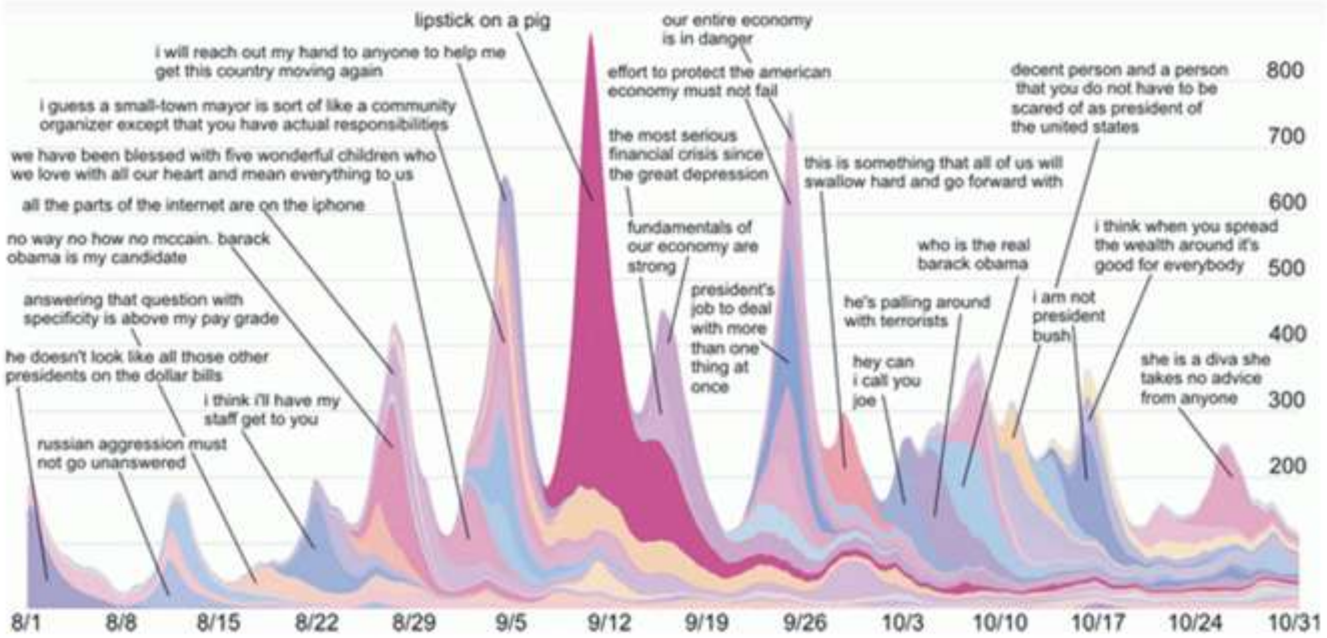
  the fundamentals of our economy are strong but these are very very difficult times and i promise you we will never put america in this position again

  but these are very very difficult times
- No a-priori network exists. Inference methods are used.

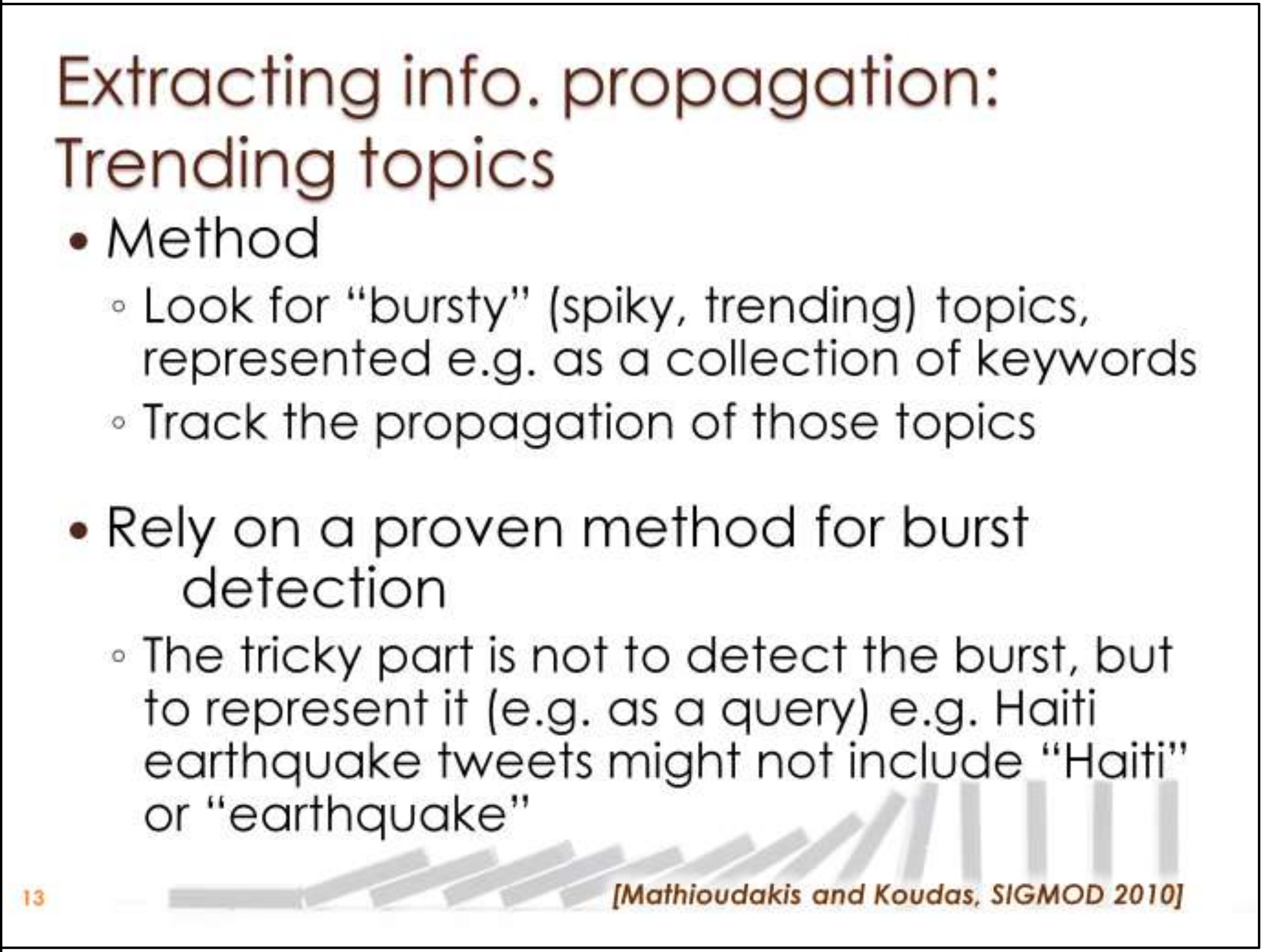
[Leskovec et al. KDD 2009]

Jure Leskovec, Lars Backstrom and Jon Kleinberg:
Meme-tracking and the Dynamics of the News Cycle.
KDD 2009
http://memetracker.org/

Extracting info. propagations: Meme tracker



[Leskovec et al. KDD 2009]

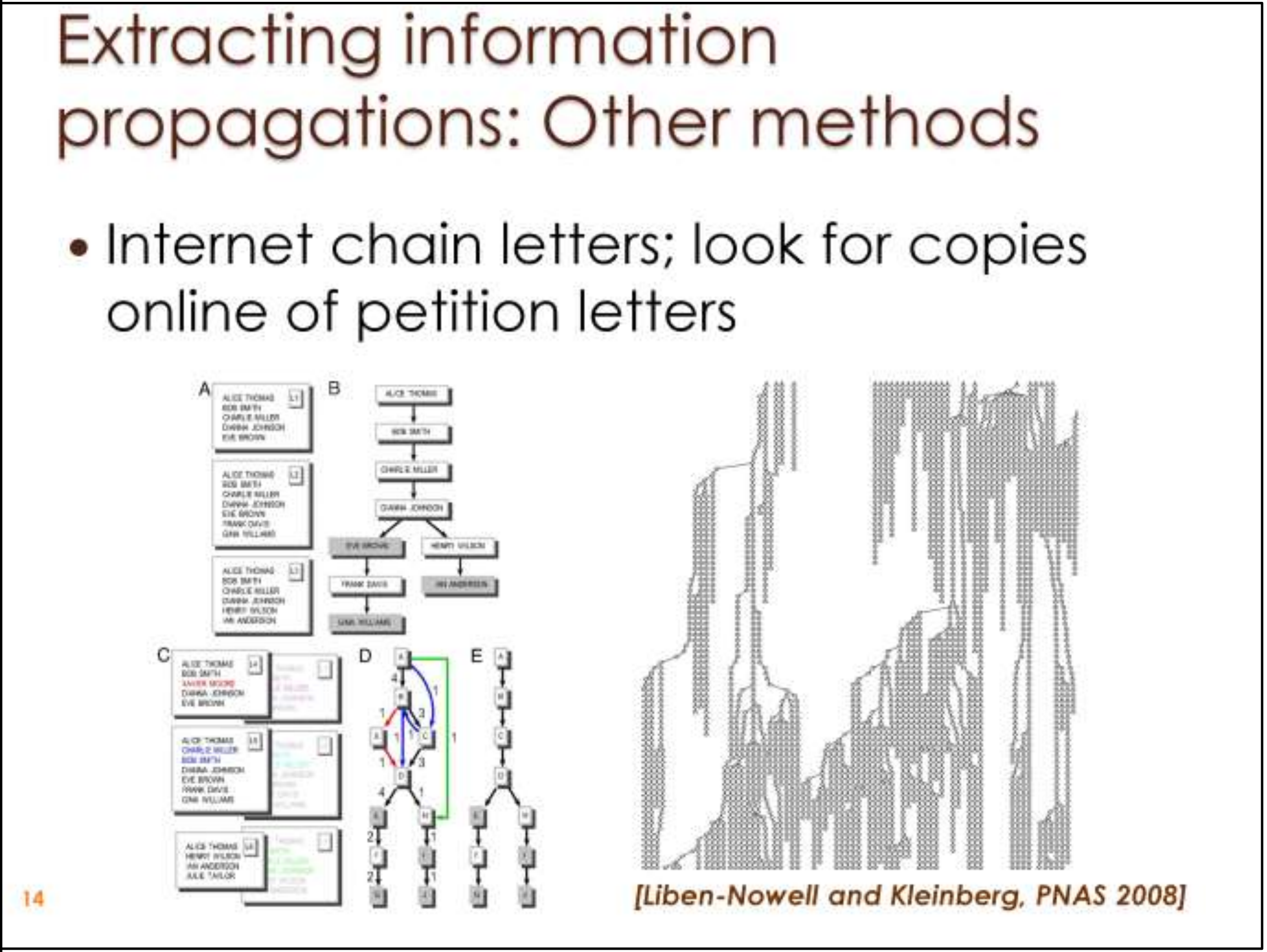


Extracting info. propagation: Trending topics

- Method
  - Look for "bursty" (spiky, trending) topics, represented e.g. as a collection of keywords
  - Track the propagation of those topics
- Rely on a proven method for burst detection
  - The tricky part is not to detect the burst, but to represent it (e.g. as a query) e.g. Haiti earthquake tweets might not include "Haiti" or "earthquake"

[Mathioudakis and Koudas, SIGMOD 2010]

Michael Mathioudakis and Nick Koudas:
TwitterMonitor: trend detection over the twitter stream.
SIGMOD 2010
http://doi.acm.org/10.1145/1807167.1807306

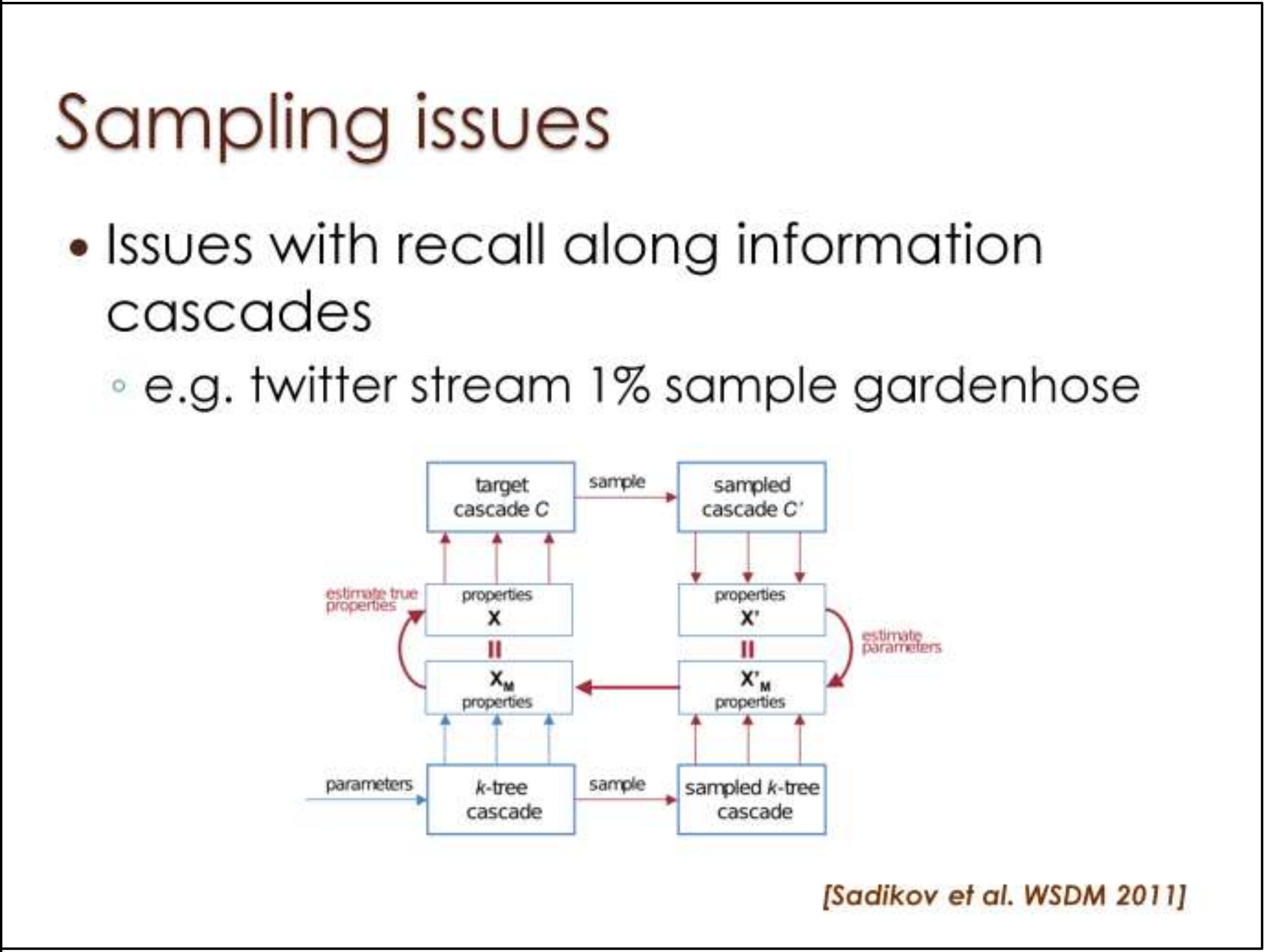


Extracting information propagations: Other methods

- Internet chain letters; look for copies online of petition letters

[Liben-Nowell and Kleinberg, PNAS 2008]

David Liben-Nowell and Jon Kleinberg:
Tracing information flow on a global scale using Internet chain-letter data.
PNAS 2008
http://www.pnas.org/content/105/12/4633.abstract



Sampling issues

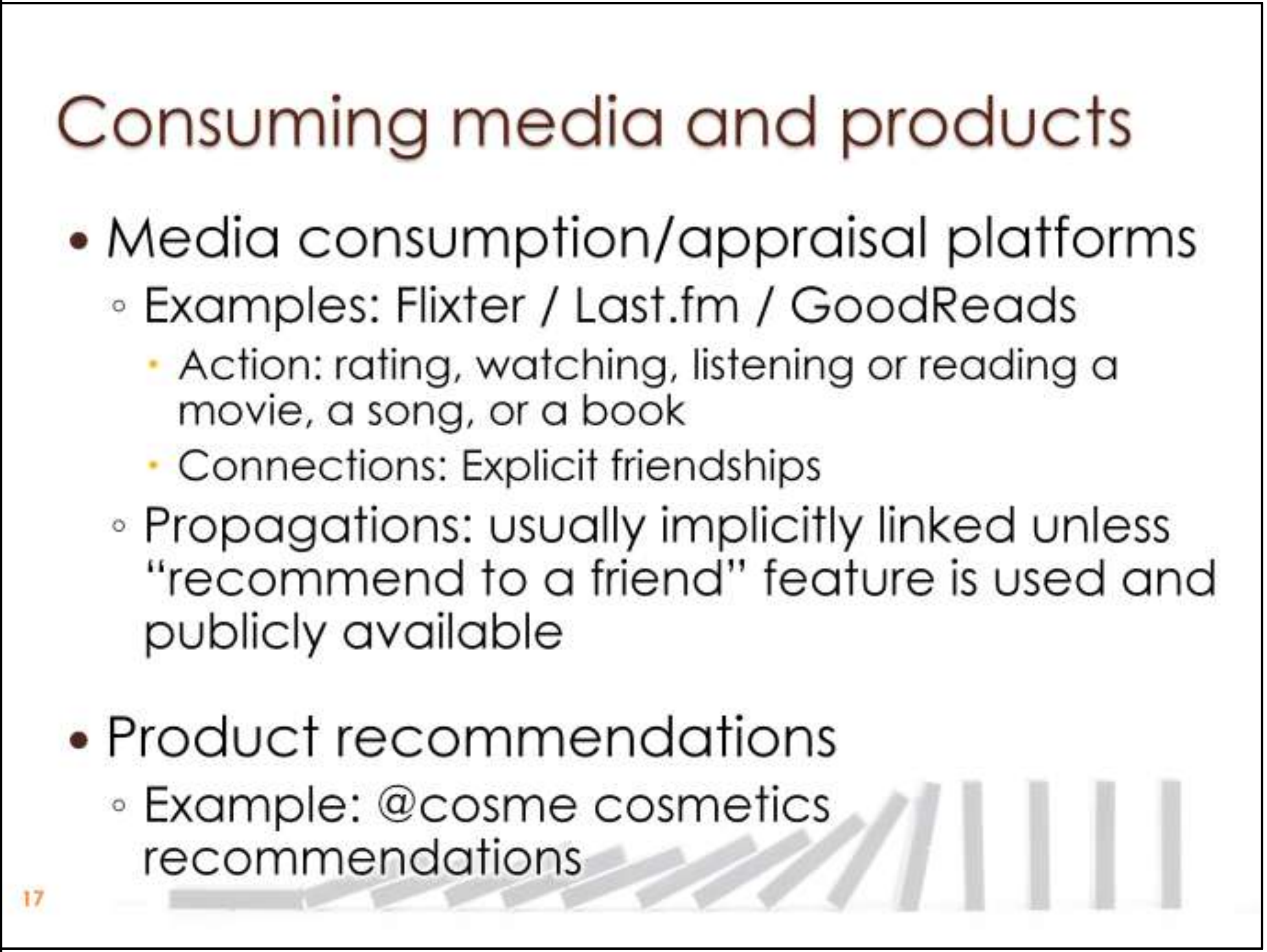
- Issues with recall along information cascades
  - e.g. twitter stream 1% sample gardenhose

target cascade *C*
sample
sampled cascade *C'*
estimate true properties
properties X
properties X'
estimate parameters
X_M properties
X'_M properties
parameters
k-tree cascade
sample
sampled *k*-tree cascade

[Sadikov et al. WSDM 2011]

Eldar Sadikov, Montserrat Medina, Jure Leskovec, and Hector Garcia-Molina:
Correcting for missing data in information cascades.
WSDM 2011
http://doi.acm.org/10.1145/1935826.1935844

Propagation of Other Actions



Consuming media and products

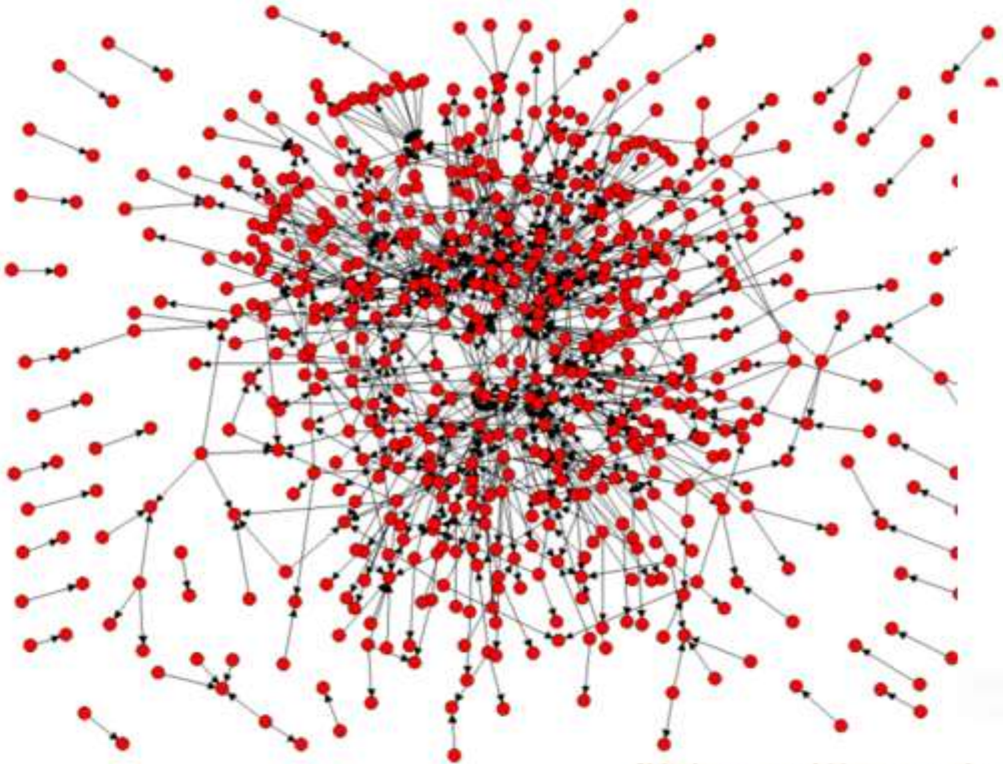
- Media consumption/appraisal platforms
  - Examples: Flixter / Last.fm / GoodReads
    - Action: rating, watching, listening or reading a movie, a song, or a book
    - Connections: Explicit friendships
  - Propagations: usually implicitly linked unless “recommend to a friend” feature is used and publicly available
- Product recommendations
  - Example: @cosme cosmetics recommendations

Smriti Bhagat, Amit Goyal, and Laks V.S. Lakshmanan:
Maximizing product adoption in social networks.
WSDM 2012
http://doi.acm.org/10.1145/2124295.212436
(Flixter and Last.fm)

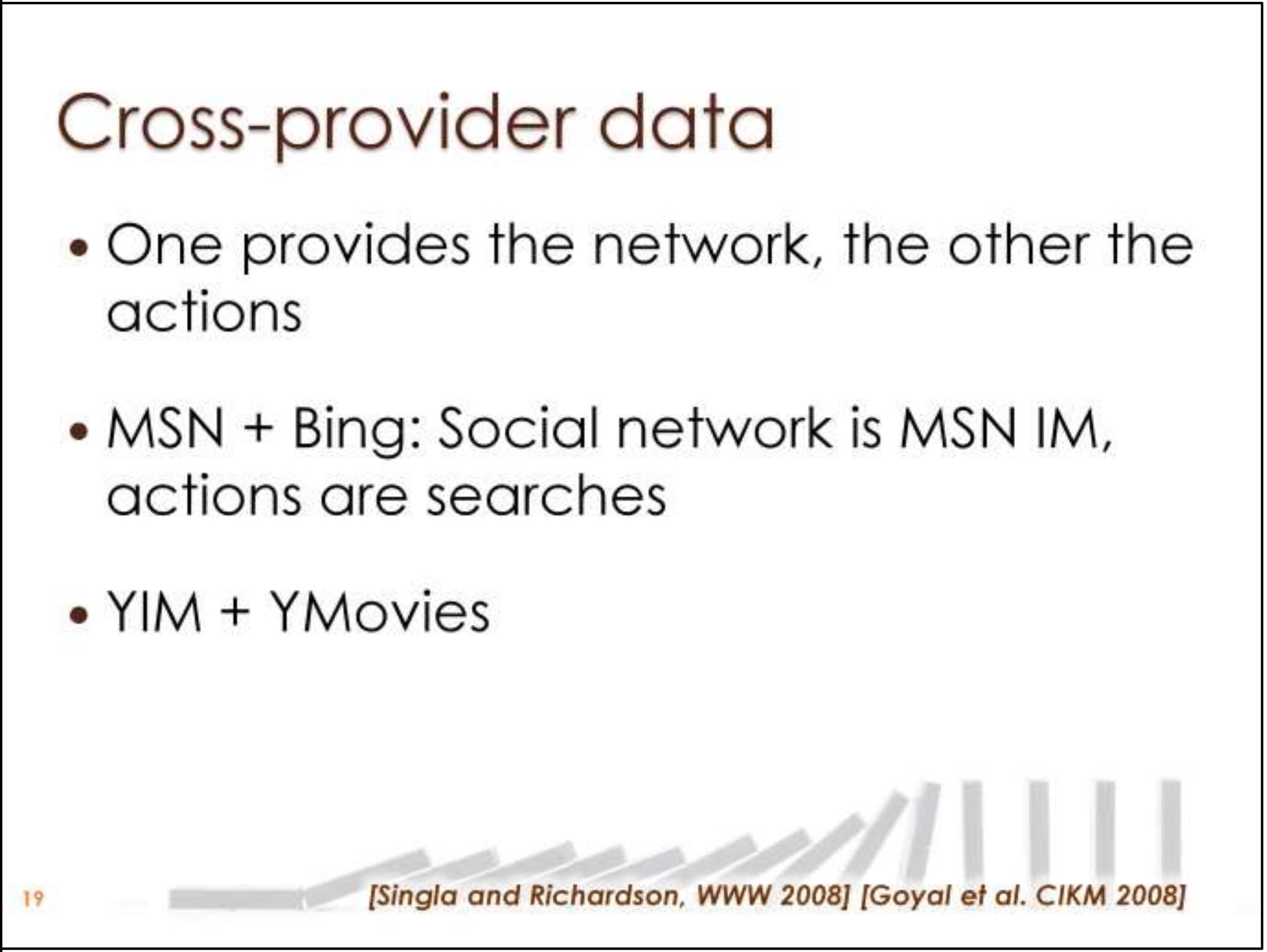
Junming Huang, Xue-Qi Cheng, Hua-Wei Shen, Tao Zhou, and Xiaolong Jin:
Exploring social influence via posterior effect of word-of-mouth recommendations.
WSDM 2012
http://doi.acm.org/10.1145/2124295.2124365
(Douban, GoodReads)

Yutaka Matsuo and Hikaru Yamamoto:
Community gravity: measuring bidirectional effects by trust and rating on online social networks.
WWW 2009
http://doi.acm.org/10.1145/1526709.1526810
(Cosme)

@cosme recommendations



[Matsuo and Yamamoto, WWW 2009]



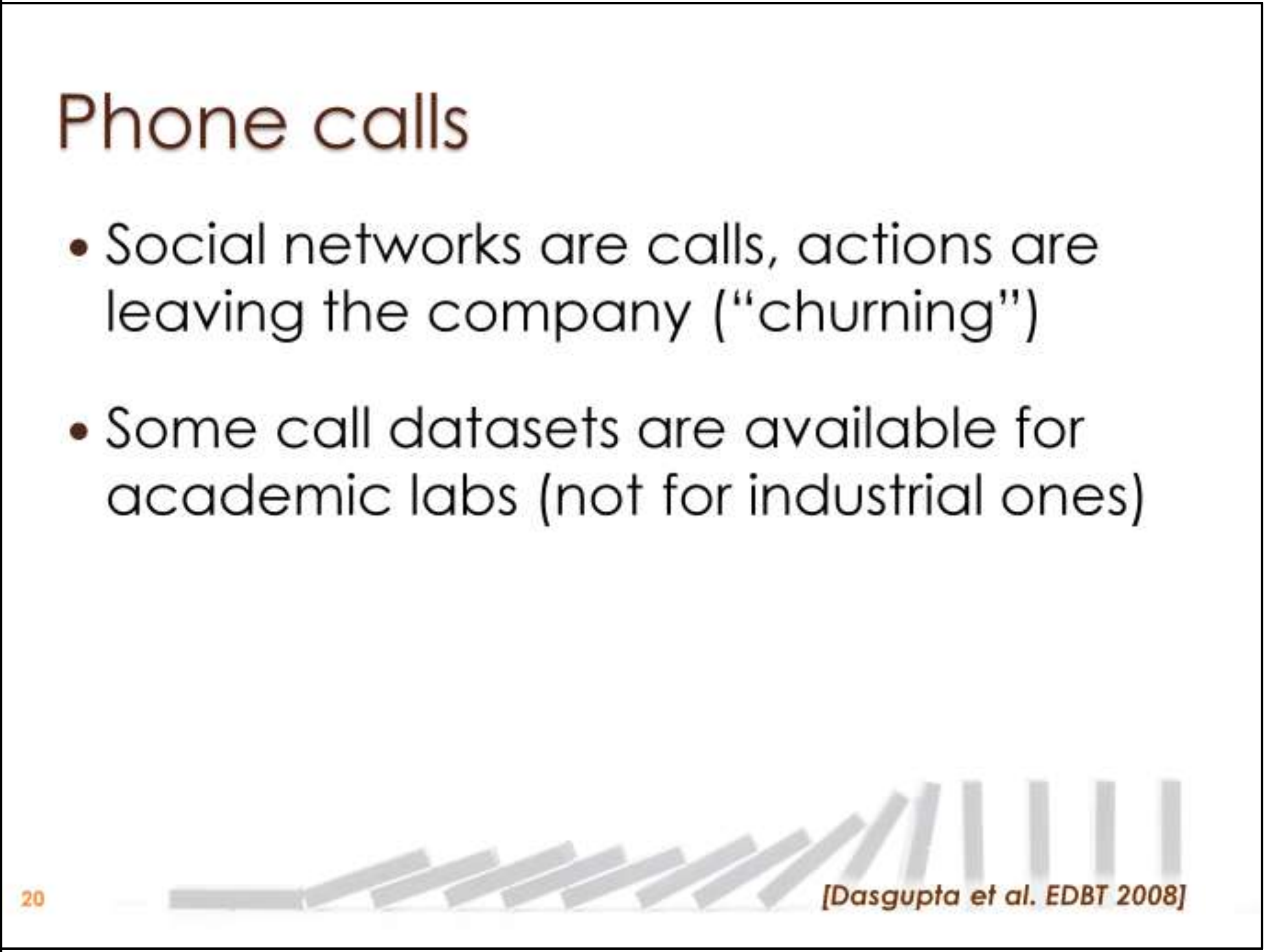
Cross-provider data

- One provides the network, the other the actions
- MSN + Bing: Social network is MSN IM, actions are searches
- YIM + YMovies

[Singla and Richardson, WWW 2008] [Goyal et al. CIKM 2008]

Parag Singla and Matthew Richardson:
Yes, there is a correlation: from social networks to personal behavior on the web.
WWW 2008
http://doi.acm.org/10.1145/1367497.1367586

Amit Goyal, Francesco Bonchi, and Laks V.S. Lakshmanan:
Discovering leaders from community actions.
CIKM 2008
http://doi.acm.org/10.1145/1458082.1458149



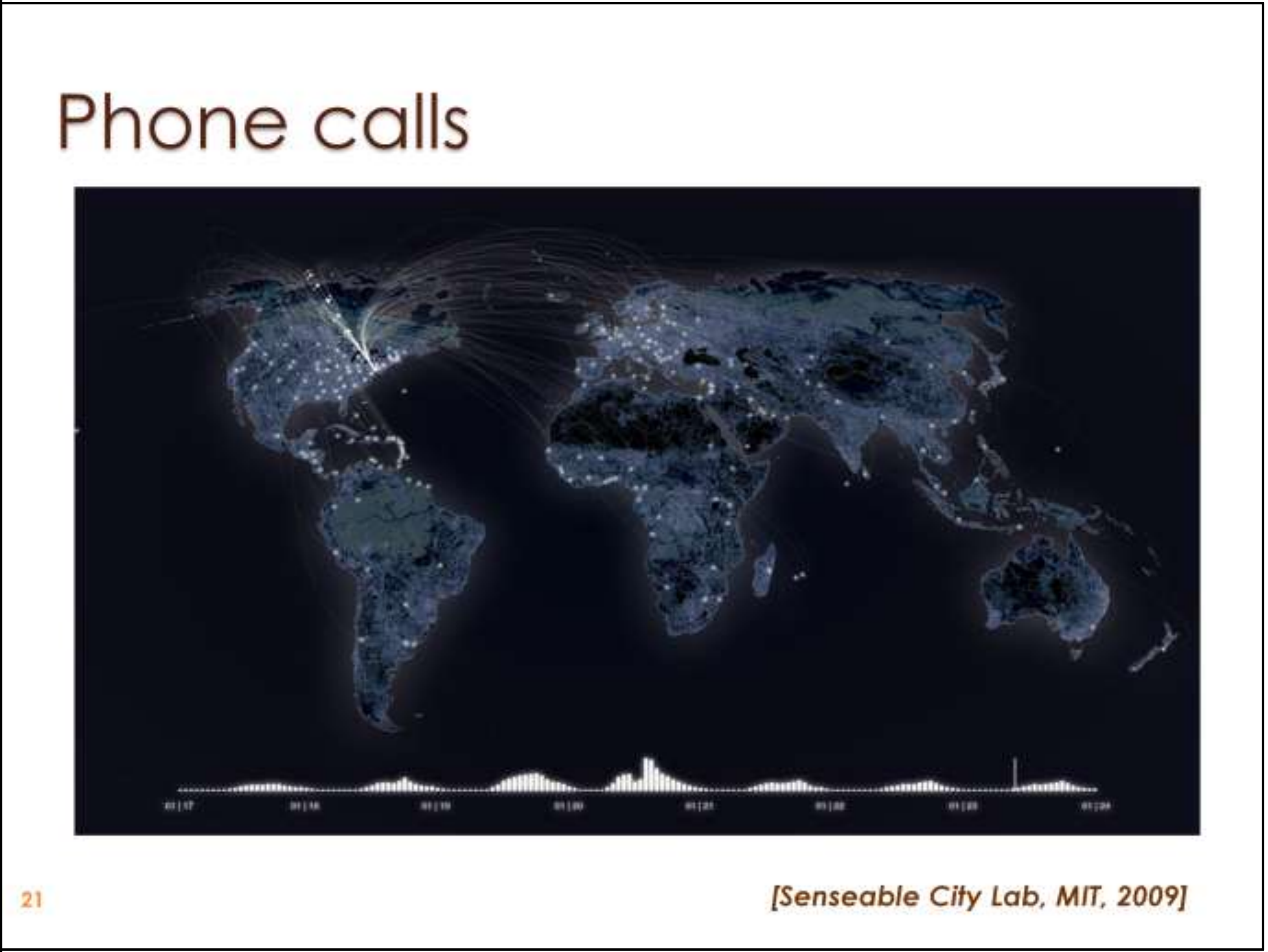
Phone calls

- Social networks are calls, actions are leaving the company (“churning”)
- Some call datasets are available for academic labs (not for industrial ones)

[Dasgupta et al. EDBT 2008]

Koustuv Dasgupta, Rahul Singh, Balaji Viswanathan, Dipanjan Chakraborty, Sougata Mukherjea, Amit A. Nanavati, and Anupam Joshi: Social ties and their relevance to churn in mobile telecom networks. EDBT 2008
http://doi.acm.org/10.1145/1353343.1353424

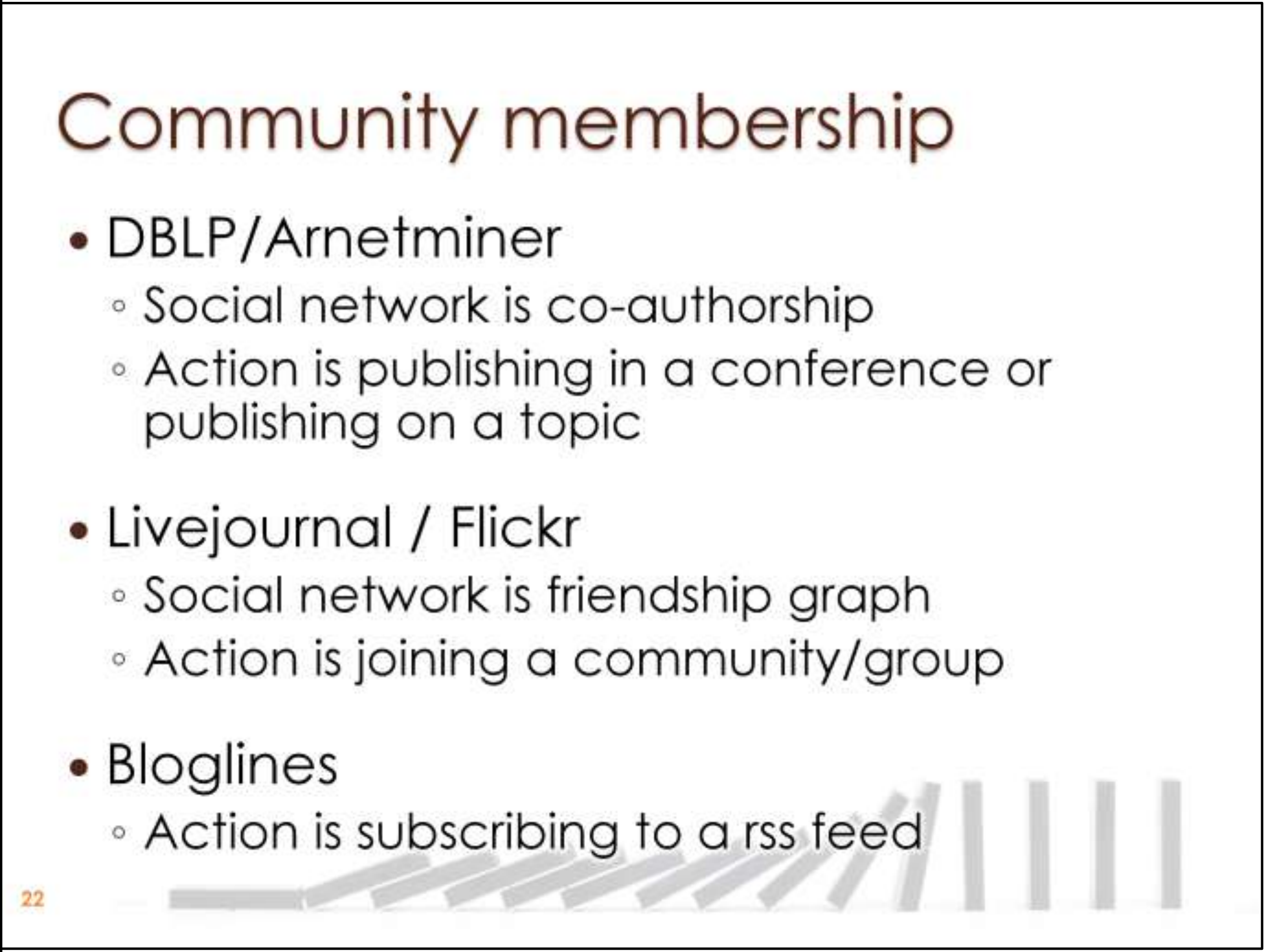
Nokia datasets:
http://research.nokia.com/page/12000



Phone calls

[Senseable City Lab, MIT, 2009]

January 20, 2009 – Obama’ s inauguration day
http://senseable.mit.edu/obama



Community membership

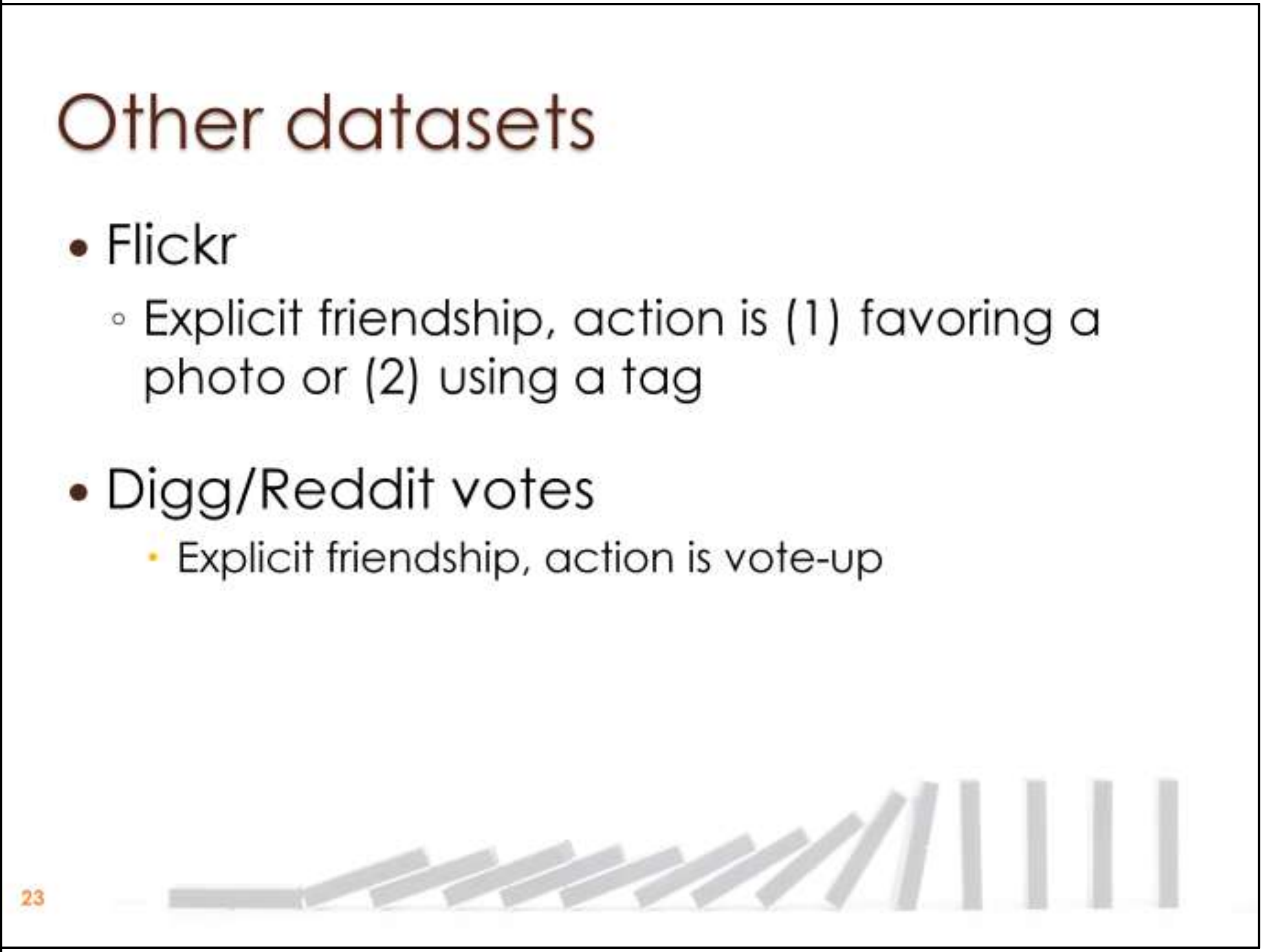
- DBLP/Arnetminer
  - Social network is co-authorship
  - Action is publishing in a conference or publishing on a topic
- Livejournal / Flickr
  - Social network is friendship graph
  - Action is joining a community/group
- Bloglines
  - Action is subscribing to a rss feed

Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan:
Group formation in large social networks: membership, growth, and evolution.
KDD 2006
http://doi.acm.org/10.1145/1150402.1150412
(uses DBLP)

Chenhao Tan, Jie Tang, Jimeng Sun, Quan Lin, and Fengjiao Wang:
Social action tracking via noise tolerant time-varying factor graphs.
KDD 2010
http://doi.acm.org/10.1145/1835804.1835936
(uses ArnetMiner)

Amit Goyal, Francesco Bonchi, and Laks V. S. Lakshmanan:
A data-based approach to social influence maximization
VLDB 2011
http://www.vldb.org/pvldb/vol5/p073_amitgoyal_vldb2012.pdf

Akshay Java, Pranam Kolari, Tim Finin, Anupam Joshi and Tim Oates:
Feeds That Matter: A Study of Bloglines Subscriptions.
ICWSM 2007
http://ebiquity.umbc.edu/get/a/publication/290.pdf



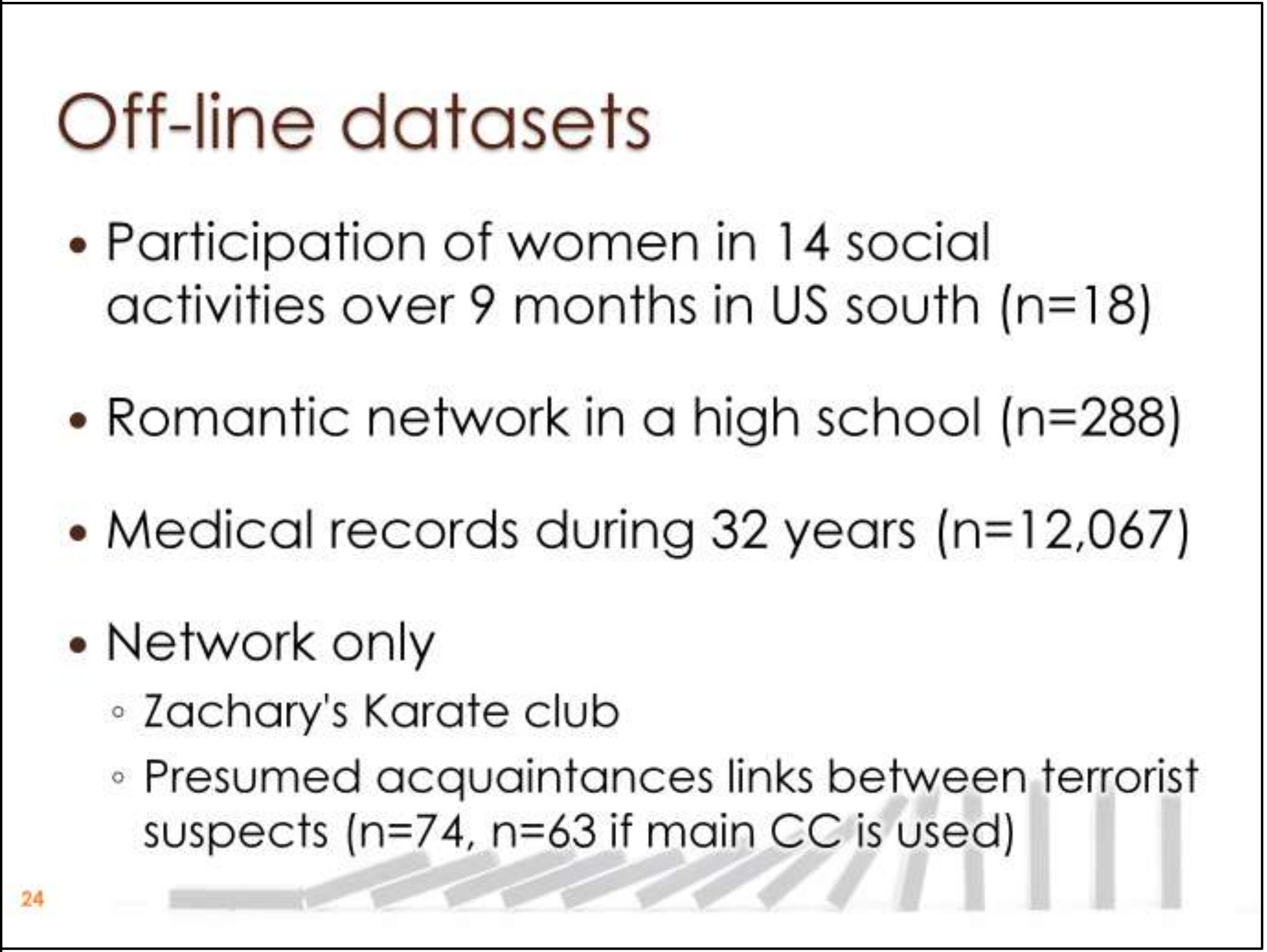
Other datasets

- Flickr
  - Explicit friendship, action is (1) favoring a photo or (2) using a tag
- Digg/Reddit votes
  - Explicit friendship, action is vote-up

Meeyoung Cha, Alan Mislove, and Krishna P. Gummadi:
A measurement-driven analysis of information propagation in the flickr social network.
WWW 2009
http://doi.acm.org/10.1145/1526709.1526806
(uses favorites)

Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian:
Influence and correlation in social networks.
KDD 2008
http://doi.acm.org/10.1145/1401890.1401897
(uses tags)

Kristina Lerman:
Social Information Processing in News Aggregation.
Internet Computing 2007
http://doi.ieeecomputersociety.org/10.1109/10.1109/MIC.2007.136



Off-line datasets

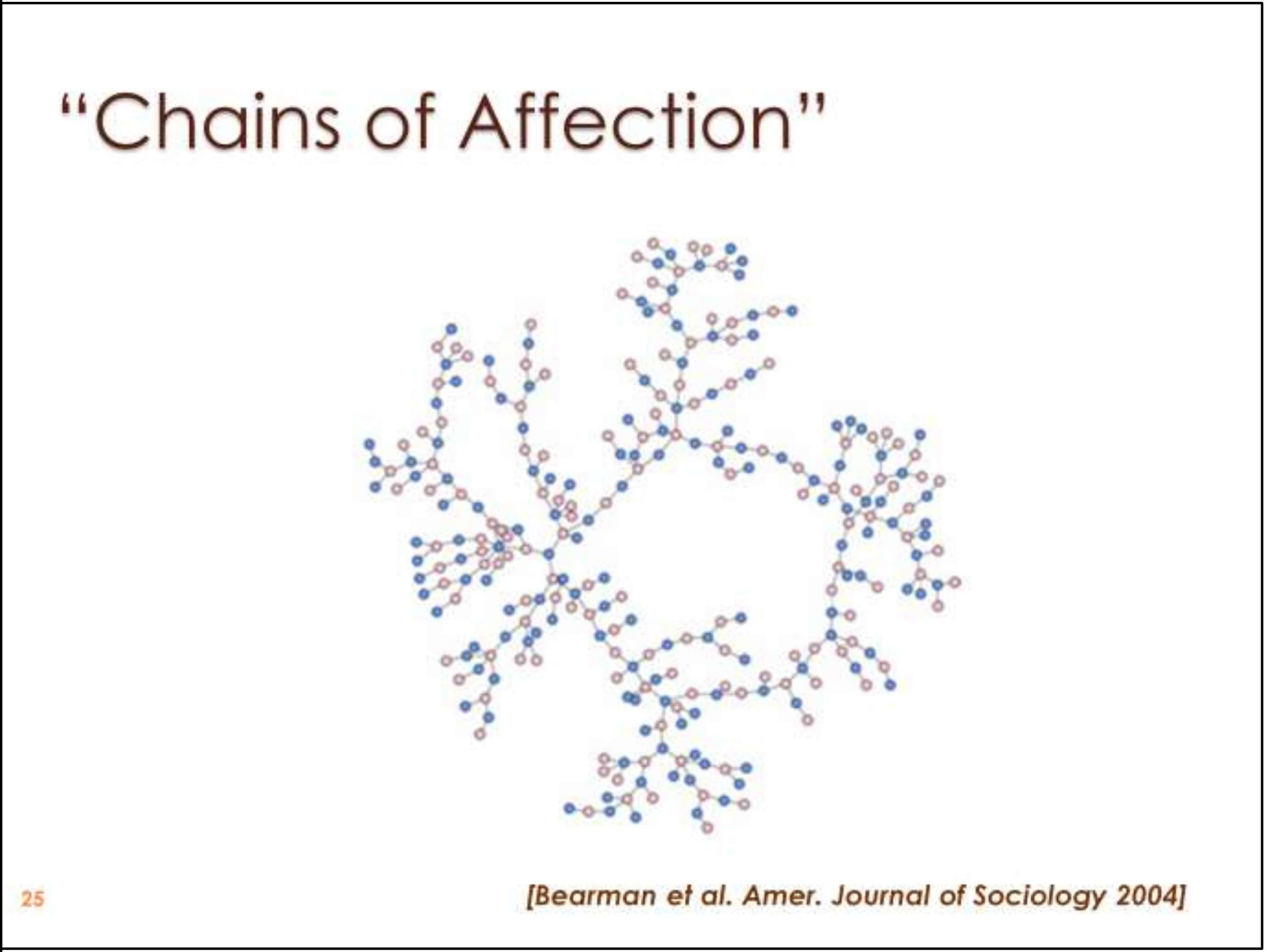
- Participation of women in 14 social activities over 9 months in US south (n=18)
- Romantic network in a high school (n=288)
- Medical records during 32 years (n=12,067)
- Network only
  - Zachary's Karate club
  - Presumed acquaintances links between terrorist suspects (n=74, n=63 if main CC is used)

A. Davis, B. B. Gardner, and M. R. Gardner:
Deep South.
1941 (The University of Chicago Press)

W. W. Zachary:
An information flow model for conflict and fission in small groups.
Journal of Anthropological Research
1977
http://networkdata.ics.uci.edu/data.php?id=105

Nicholas A. Christakis and James H. Fowler:
The Spread of Obesity in a Large Social Network over 32 Years.
The New England Journal of Medicine
2006
http://www.nejm.org/doi/full/10.1056/NEJMsa066082

Valdis Krebs:
Uncloaking Terrorist Networks.
First Monday 2002
http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/941/863/

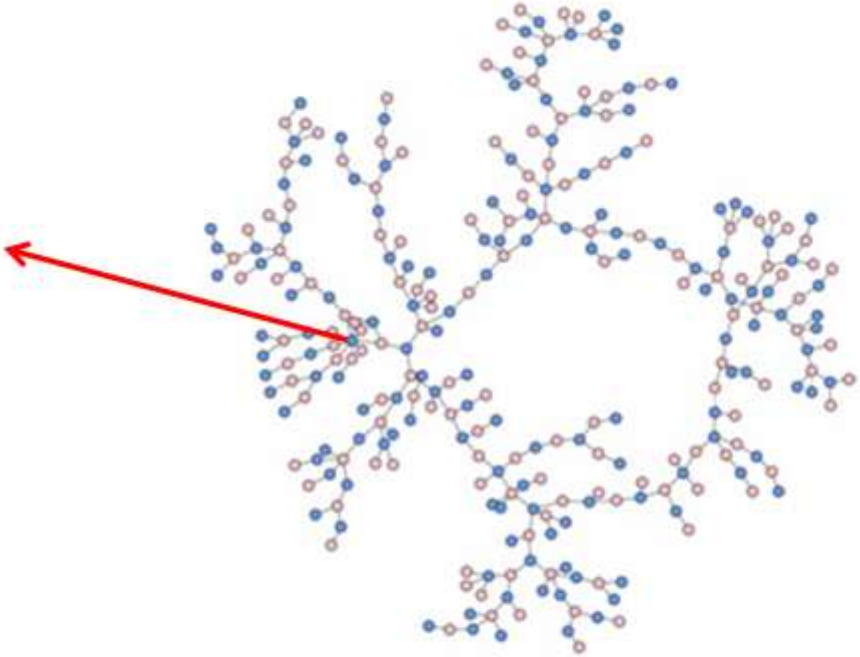


“Chains of Affection”

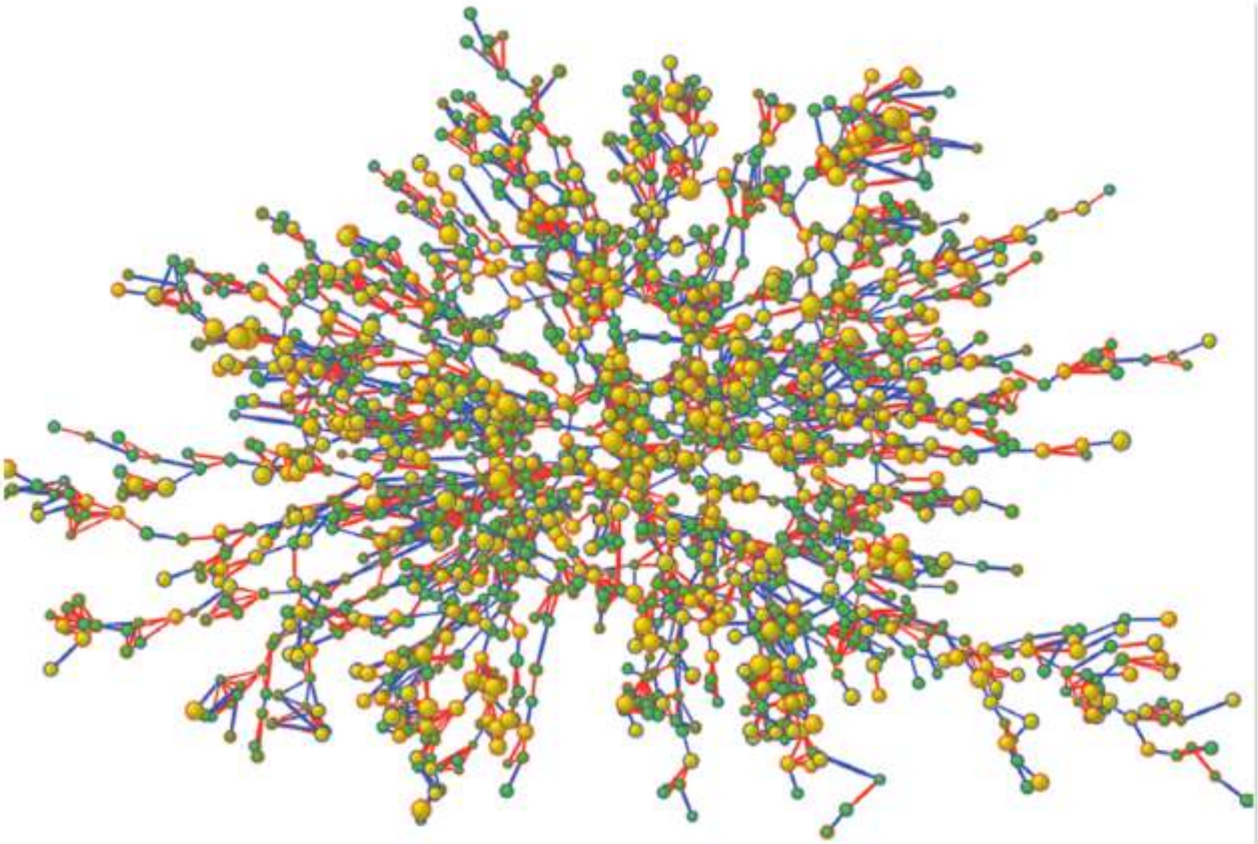
[Bearman et al. Amer. Journal of Sociology 2004]

Peter S. Bearman, James Moody and Katherine Stovel:
Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks.
American Journal of Sociology 2004
http://www.jstor.org/stable/10.1086/386272

“Chains of Affection”



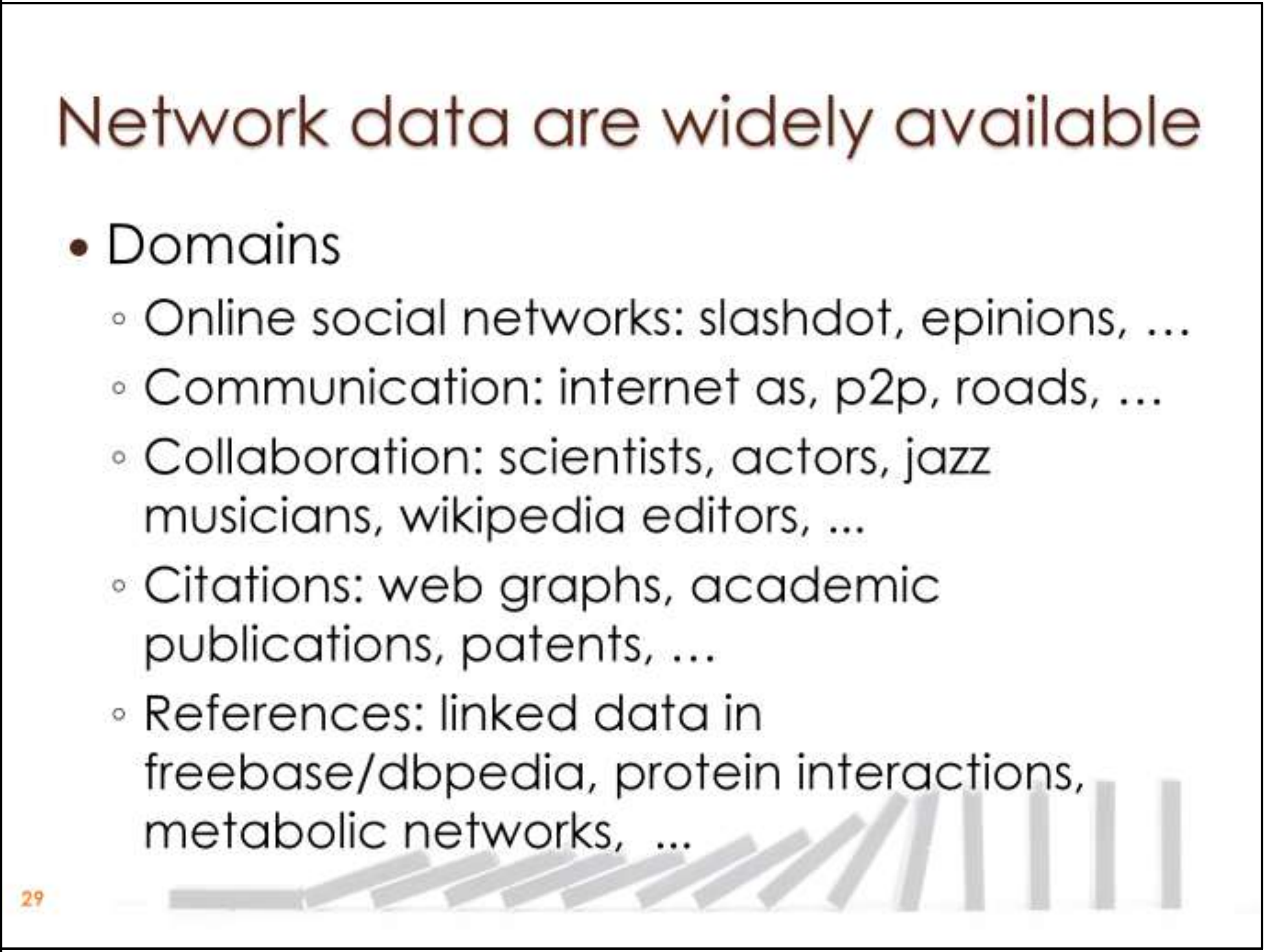
Probably not a future computer scientist ☺



Size proportional to BMI, yellow fill indicates obesity. Blue border=men, Red border=women

[Christakis and Fowler, New England Journal of Medicine 2007]

Synthetic Datasets



Network data are widely available

- Domains
  - Online social networks: slashdot, epinions, …
  - Communication: internet as, p2p, roads, …
  - Collaboration: scientists, actors, jazz musicians, wikipedia editors, ...
  - Citations: web graphs, academic publications, patents, …
  - References: linked data in freebase/dbpedia, protein interactions, metabolic networks, ...

http://snap.stanford.edu/data/

http://www-personal.umich.edu/~mejn/netdata/

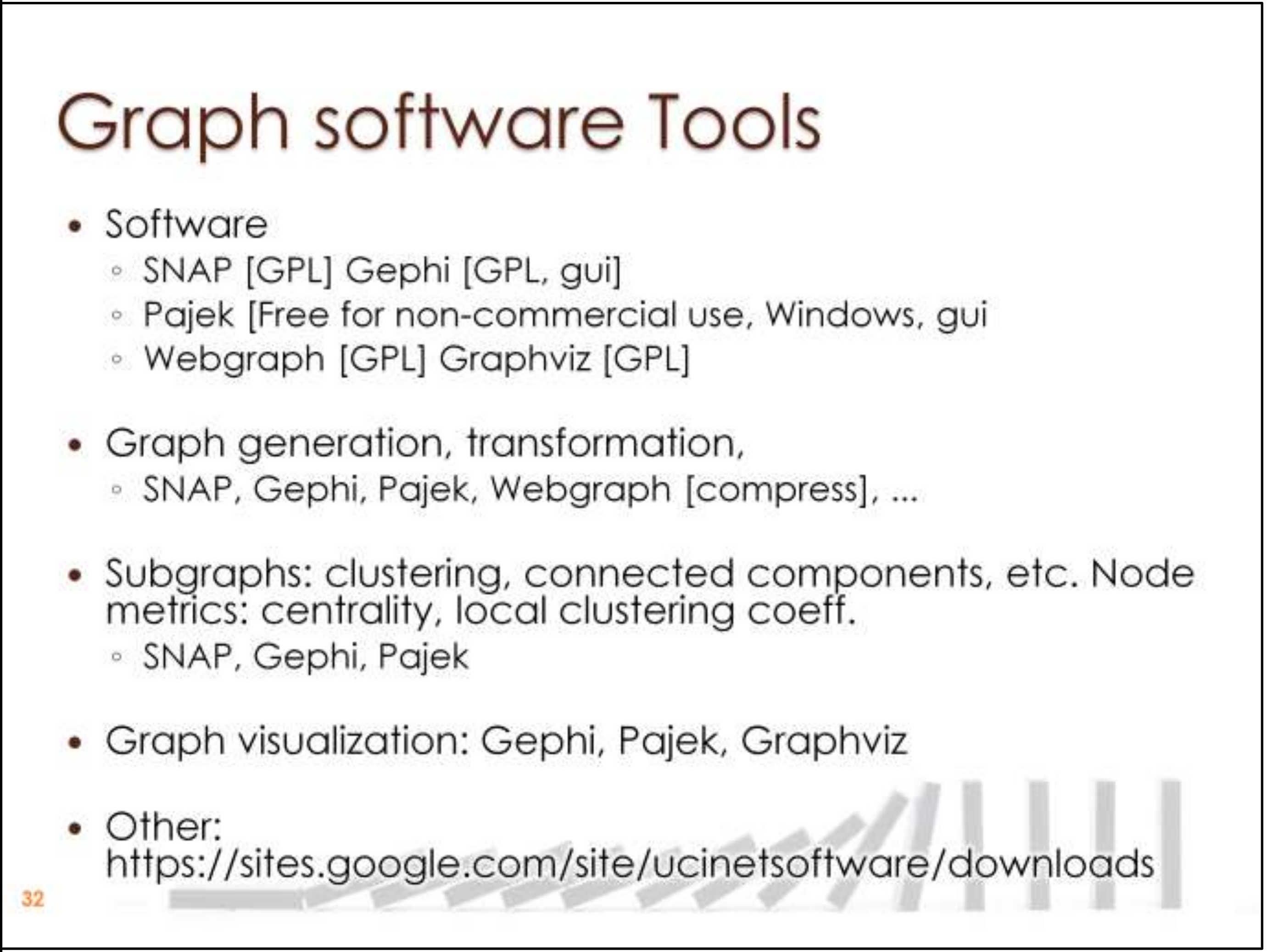
http://aws.amazon.com/datasets
(5x109 pages crawl)

http://networkdata.ics.uci.edu/

Publishing your own datasets

- Document every step of sampling, filtering, processing methodology
- CC0 (public domain) data releases
- Ad-hoc data releases: look at items in example agreements (duration, purpose, warranties, item deletion policies, etc.)
- Privacy concerns
- It may take some extra work, but remember that it is also in YOUR interest that your data is used by others

Software



Graph software Tools

- Software
  - SNAP [GPL] Gephi [GPL, gui]
  - Pajek [Free for non-commercial use, Windows, gui
  - Webgraph [GPL] Graphviz [GPL]
- Graph generation, transformation,
  - SNAP, Gephi, Pajek, Webgraph [compress], ...
- Subgraphs: clustering, connected components, etc. Node metrics: centrality, local clustering coeff.
  - SNAP, Gephi, Pajek
- Graph visualization: Gephi, Pajek, Graphviz
- Other: https://sites.google.com/site/ucinetsoftware/downloads

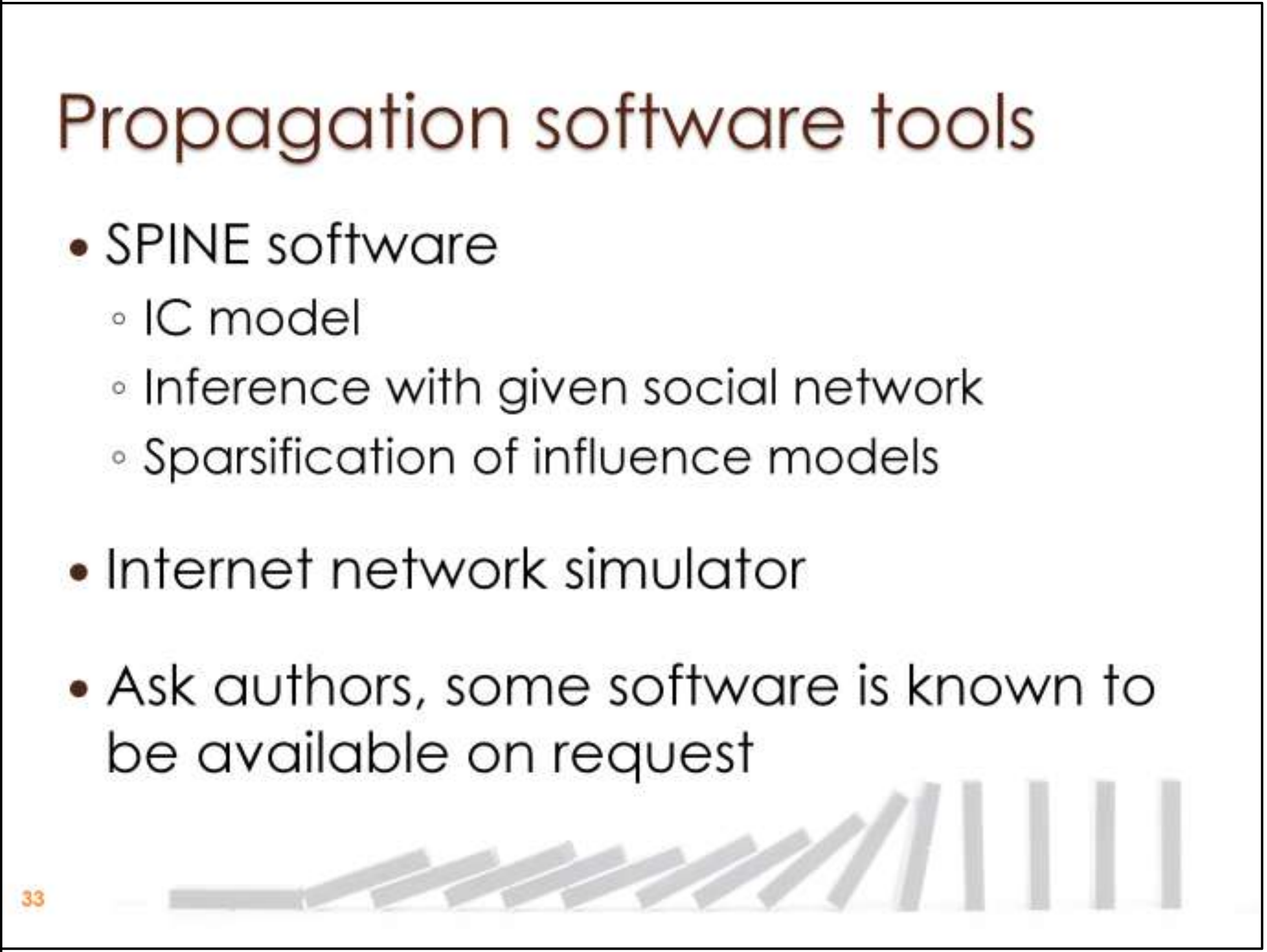
http://snap.stanford.edu/snap/

http://pajek.imfm.si/doku.php?id=pajek

http://gephi.org

http://graphviz.org/

http://webgraph.dsi.unimi.it/

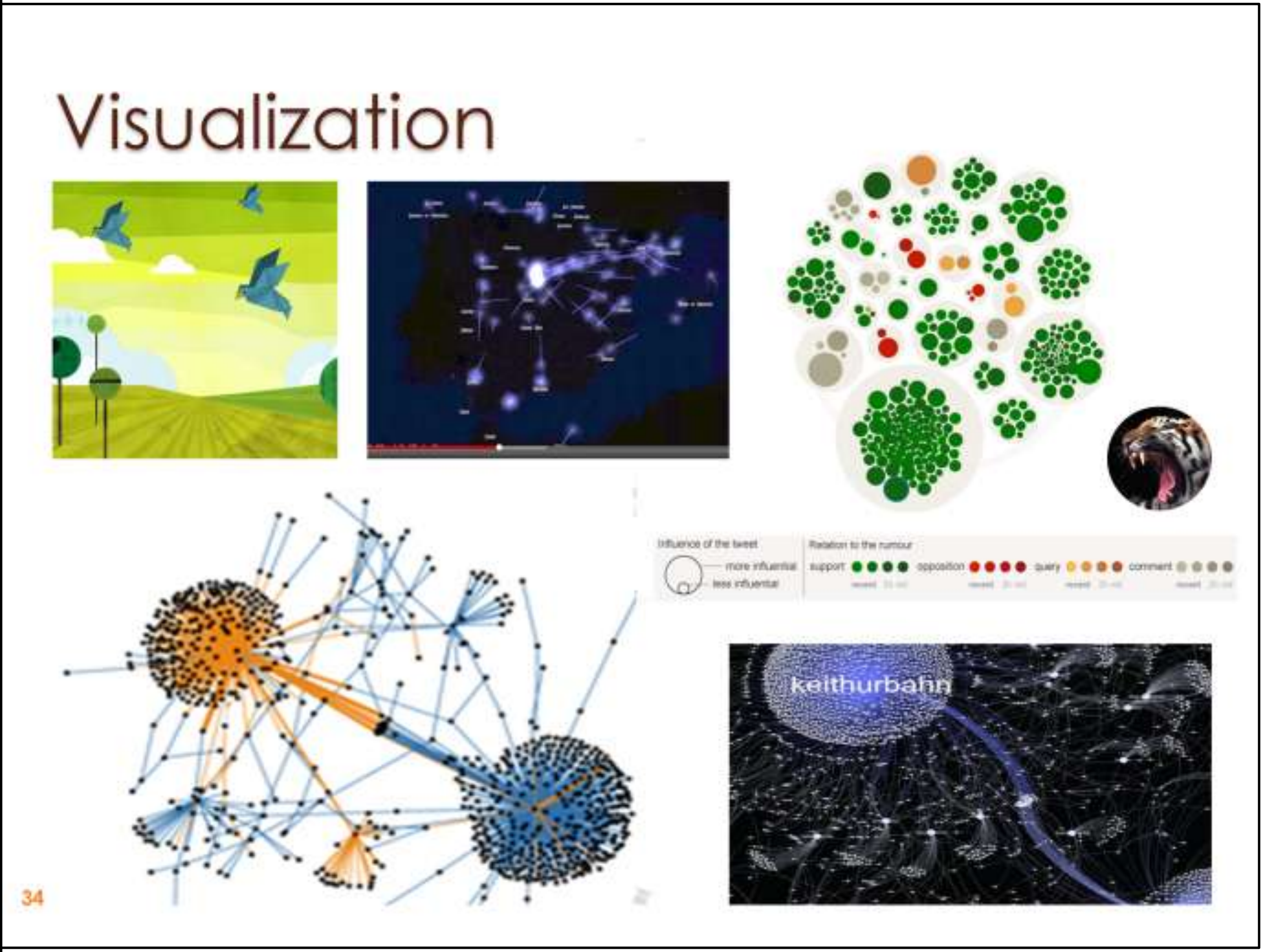


Propagation software tools

- SPINE software
  - IC model
  - Inference with given social network
  - Sparsification of influence models
- Internet network simulator
- Ask authors, some software is known to be available on request

SPINE [Mathioidakis et al. KDD 2011]
http://queens.db.toronto.edu/~mathiou/spine/

Internet network simulator.
http://isi.edu/nsnam/ns/doc/



Visualization

(From top-left, row-wise)

Tori’s Eye
http://toriseye.quodis.com/

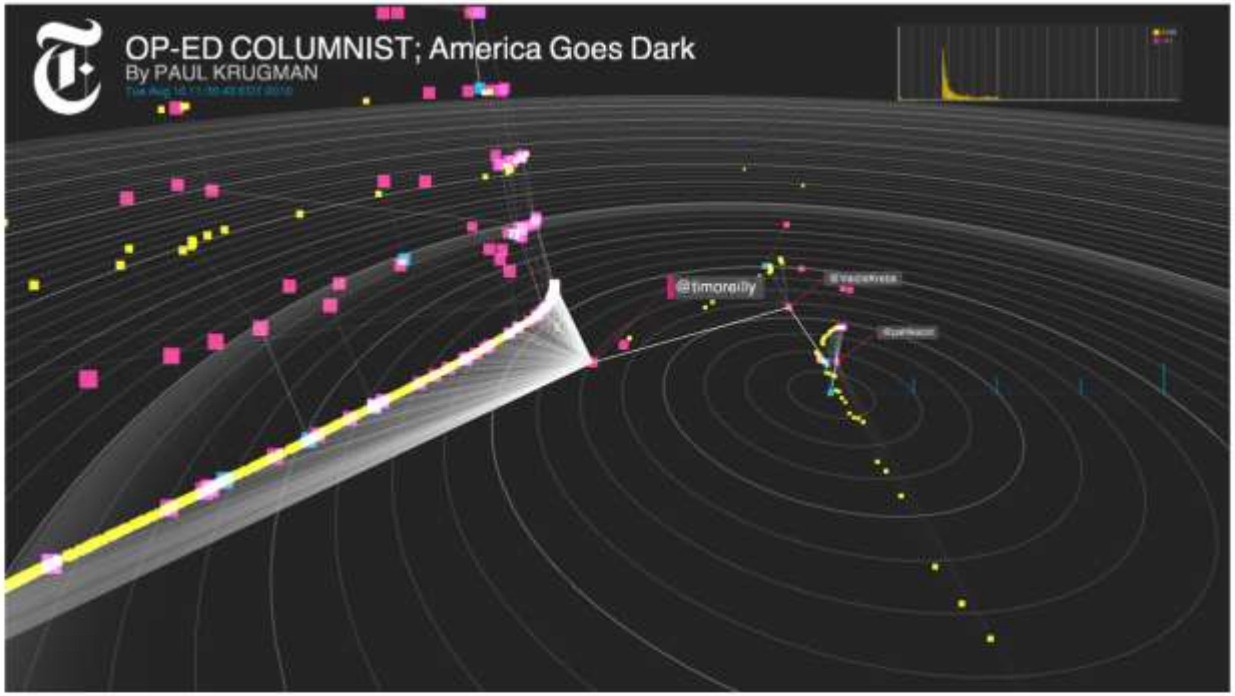
15M in Spain
http://www.youtube.com/watch?v=ECqzsom7axQ

Reading the Riots, by the Guardian
http://www.guardian.co.uk/uk/interactive/2011/dec/07/london-riots-twitter
[Rumour: “Rioters attack London Zoo and release animals”]

Truthy from Indiana University
http://truthy.indiana.edu/

Visualizing the Power of a Single Tweet
http://blog.socialflow.com/post/5246404319/breaking-bin-laden-visualizing-the-power-of-a-single

Visualization



[Project Cascade by NYT Labs 2011]

New York Times Labs:
Project Cascade.
http://nytlabs.com/projects/cascade.html

Key takeaways of part II

- Data availability affects our research

- Current alternatives are not good
  - Results on proprietary data sources are not reproducible
  - Synthetic information propagations might not be realistic

- Software is not readily available

- This is something to work on collectively!