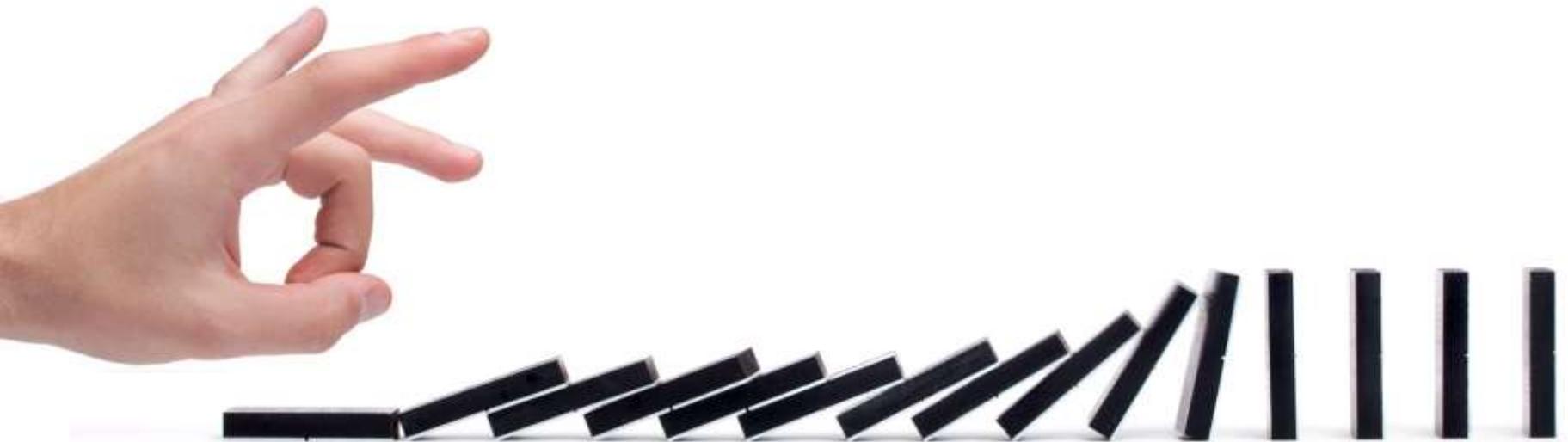


Part I → Part II → Part III → Part IV → Part V

Data and Software



Part II: Outline

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

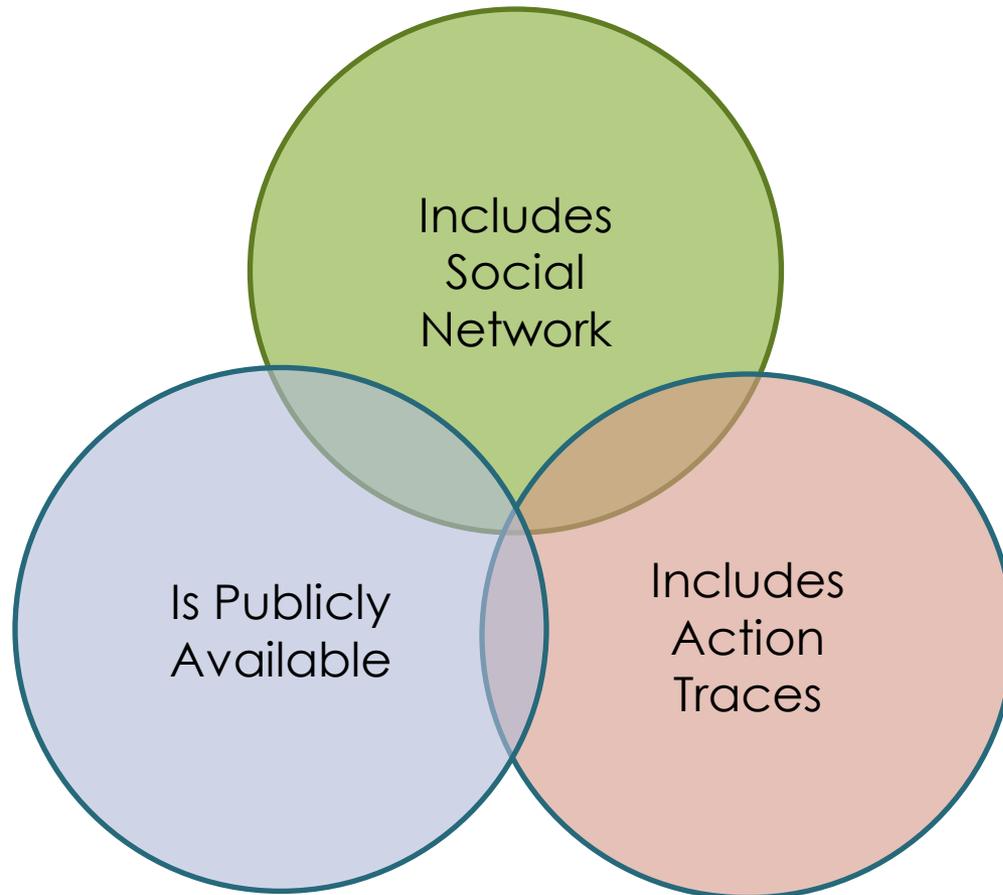


Contents of a dataset

- 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



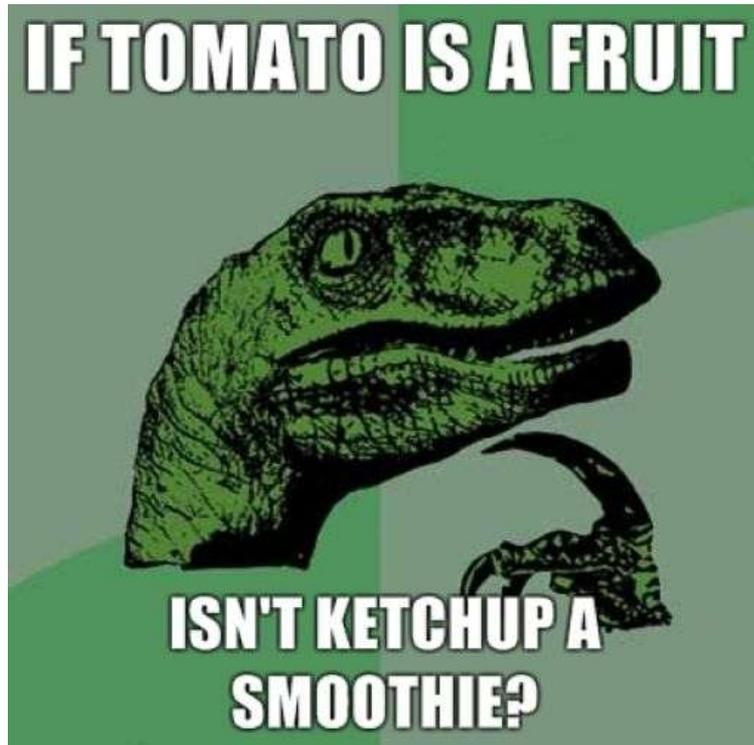
Classification: according to availability

- 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”



Mememes and “Internet Mememes”



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”

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 ...

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

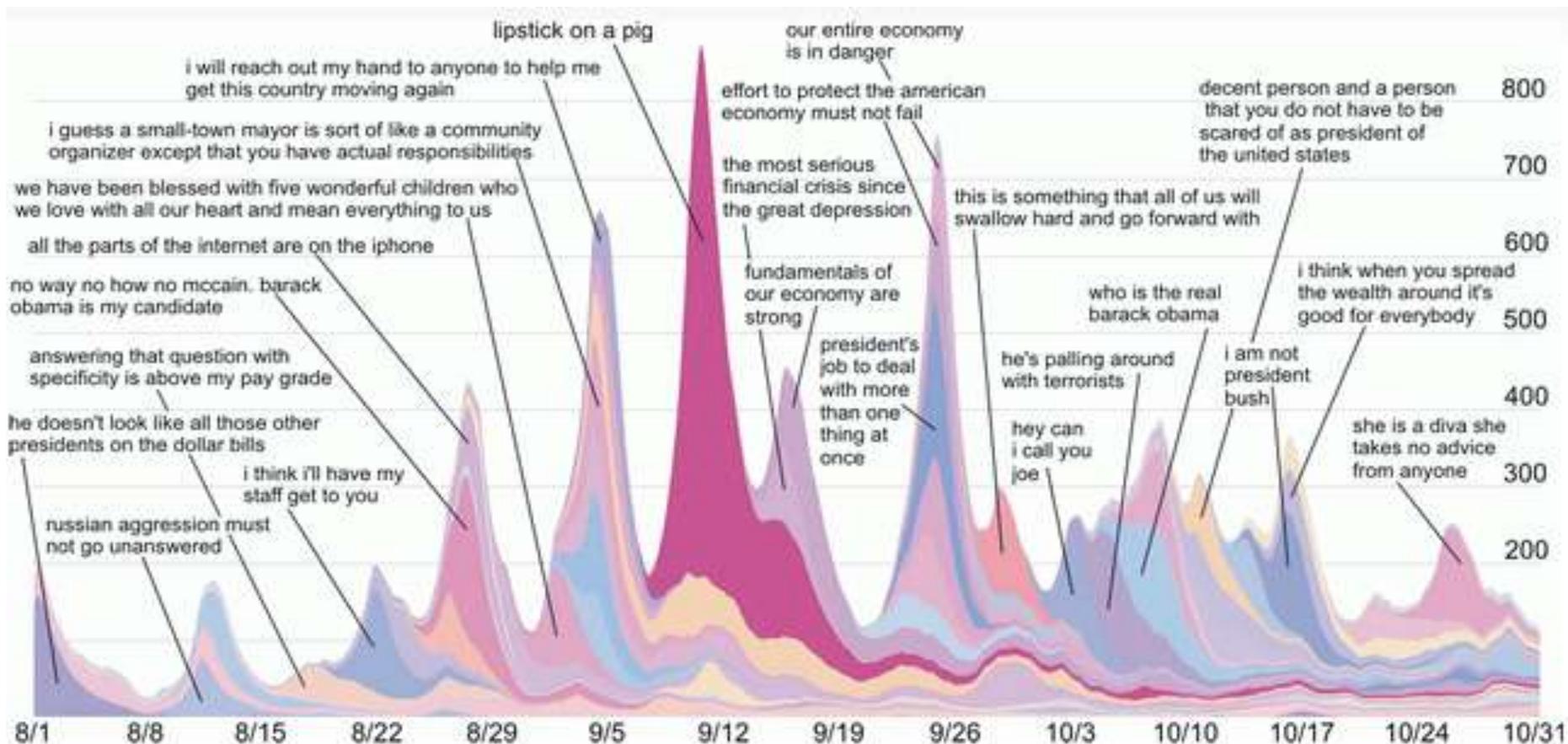
i promise you we will never put america in this position again we will clean up wall street

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.

Extracting info. propagations: Meme tracker

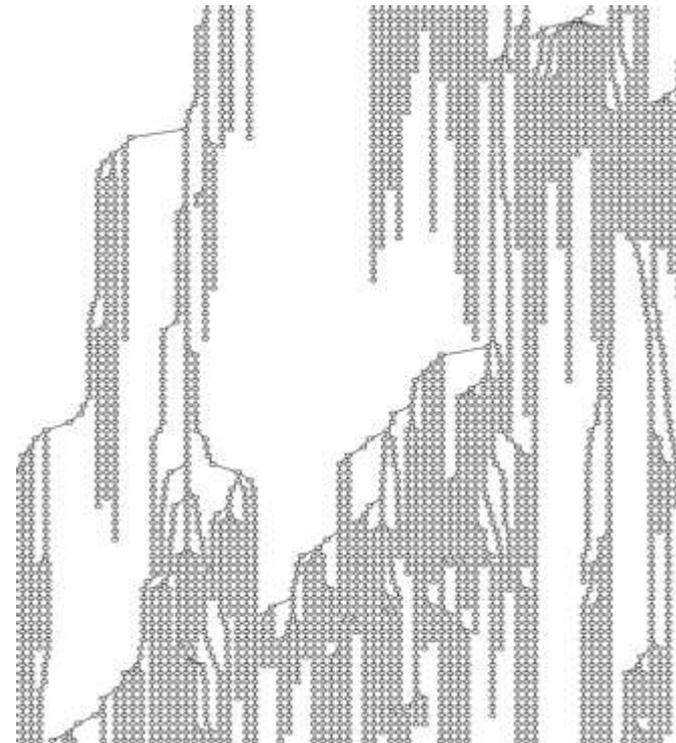
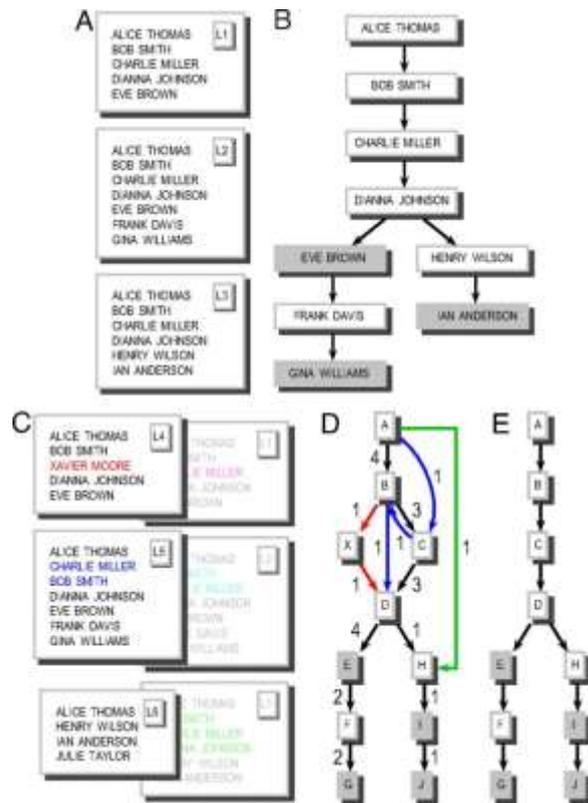


Extracting info. propagation: Trending topics

- 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”

Extracting information propagations: Other methods

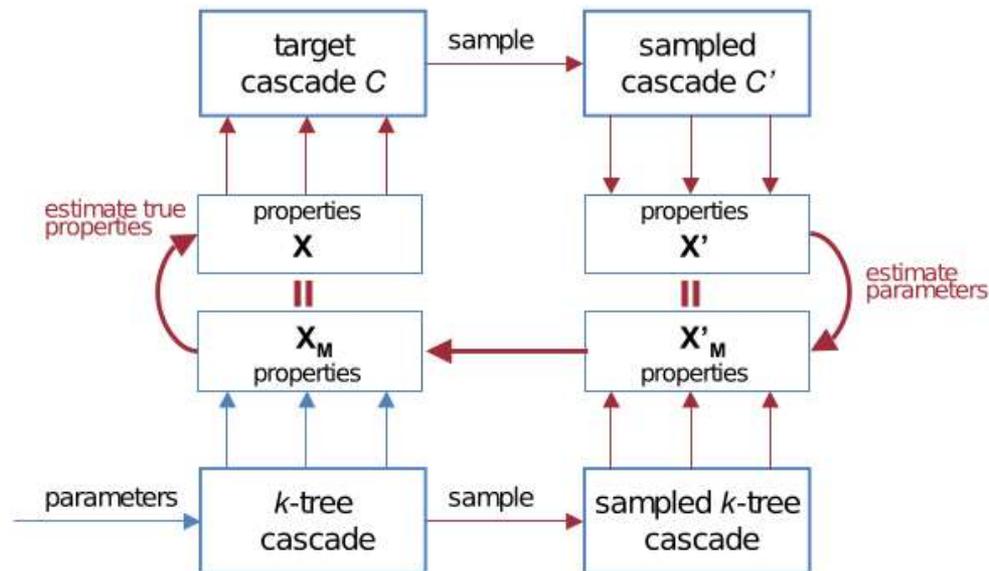
- Internet chain letters; look for copies online of petition letters



[Liben-Nowell and Kleinberg, PNAS 2008]

Sampling issues

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



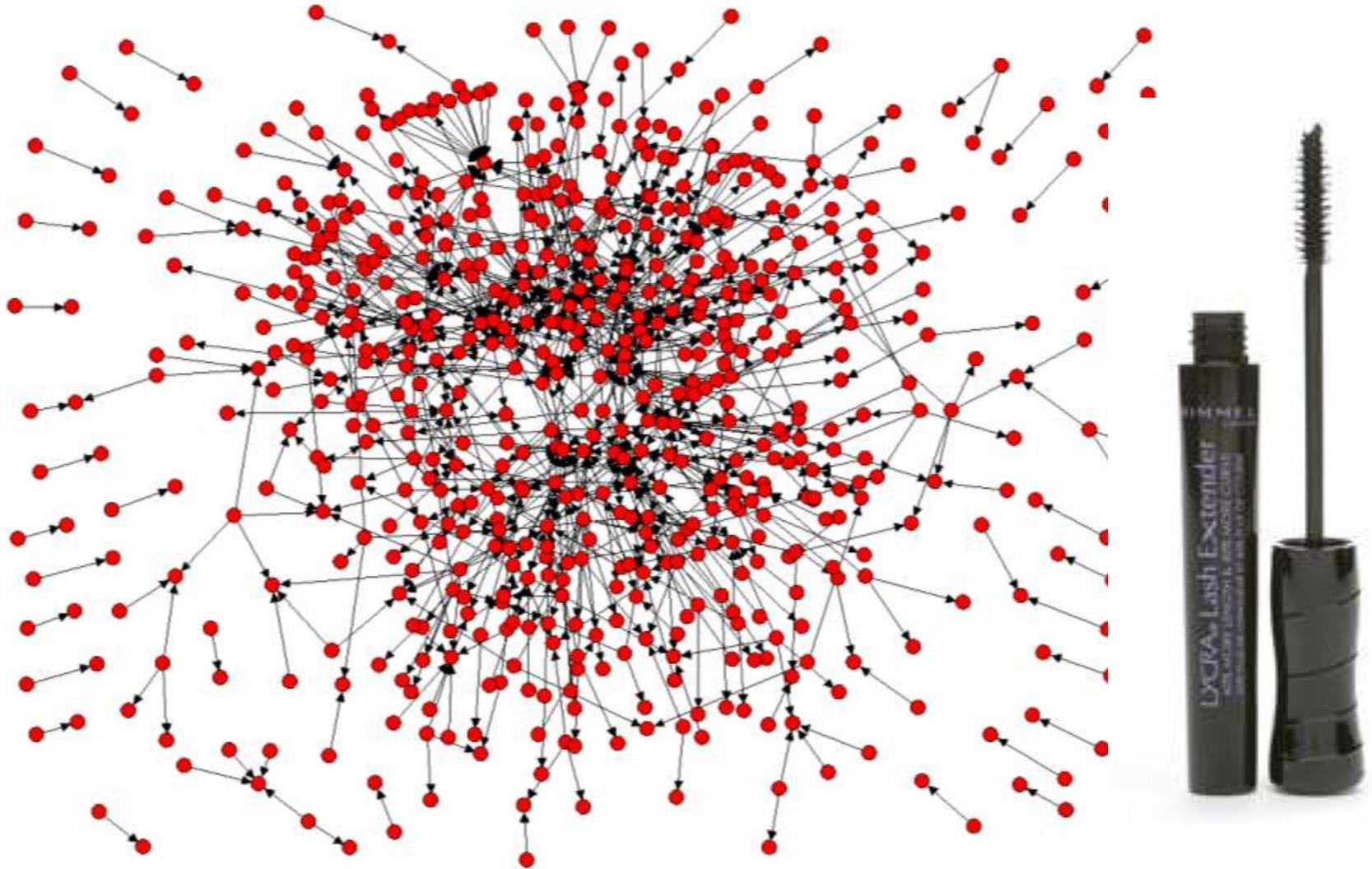
Propagation of Other Actions



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

@cosme recommendations



Cross-provider data

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

Phone calls

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

Phone calls



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

Other datasets

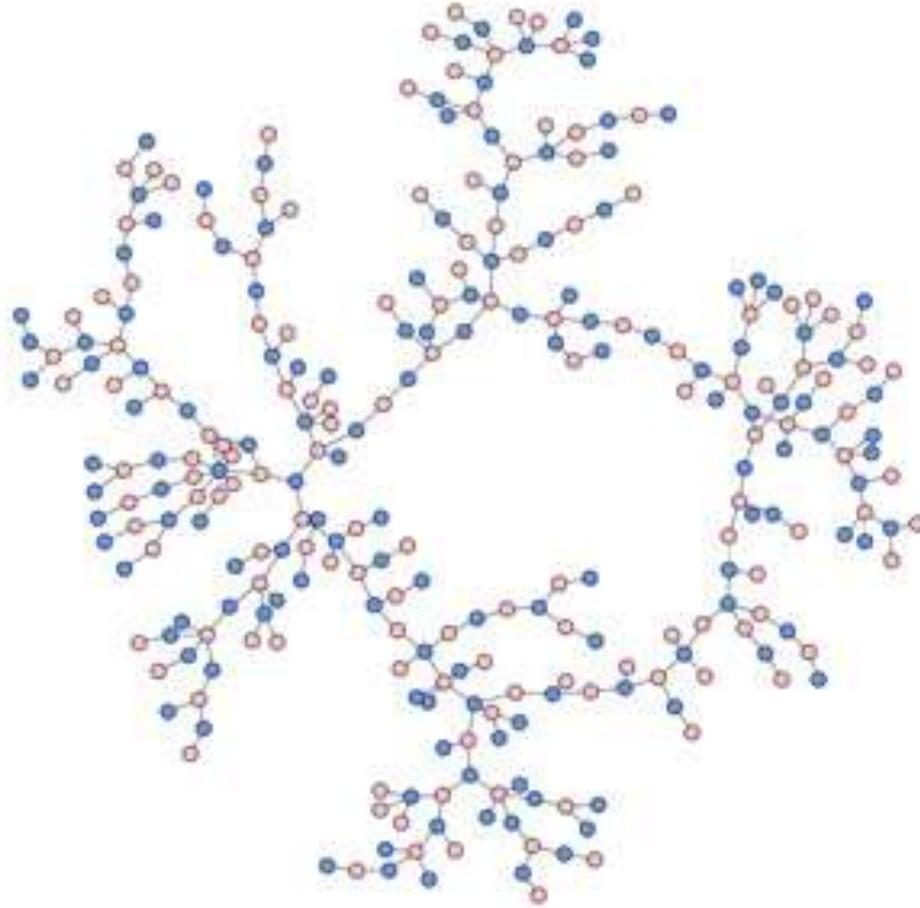
- Flickr
 - Explicit friendship, action is (1) favoring a photo or (2) using a tag
- Digg/Reddit votes
 - Explicit friendship, action is vote-up



Off-line datasets

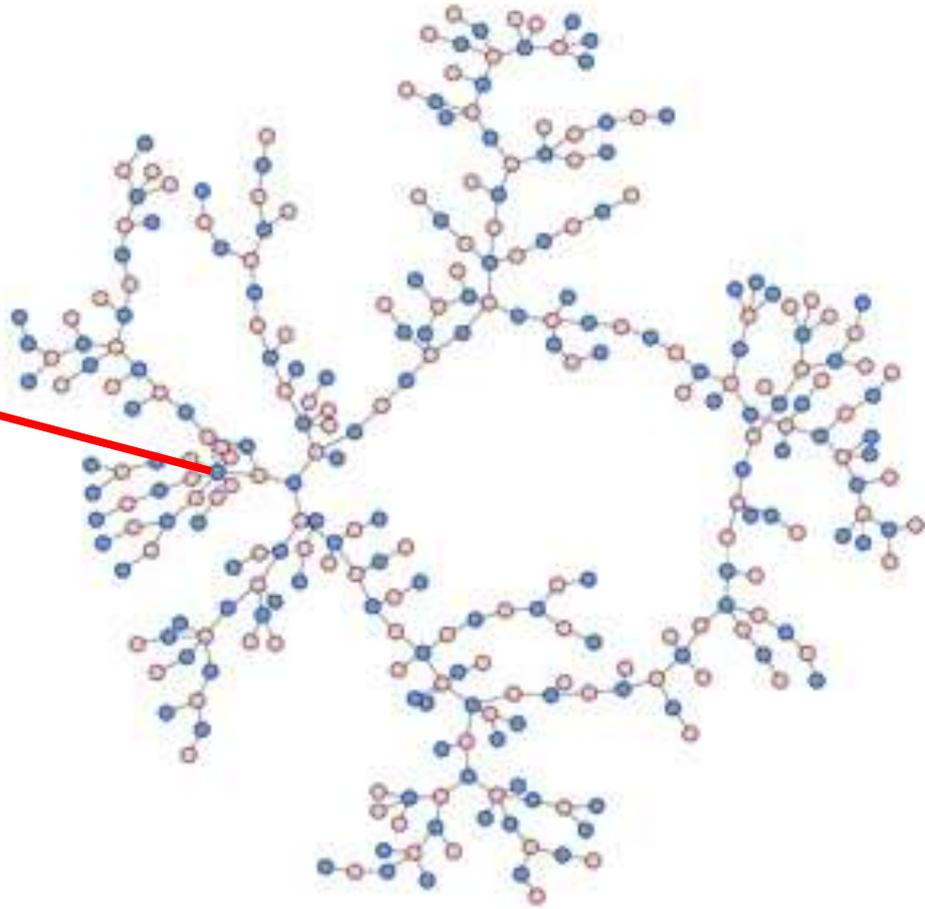
- Participation of women in 14 social activities over 9 months in US south (n=18)
- Romantic network in a high school (n=288)
- Medical records during 32 years (n=12,067)
- Network only
 - Zachary's Karate club
 - Presumed acquaintances links between terrorist suspects (n=74, n=63 if main CC is used)

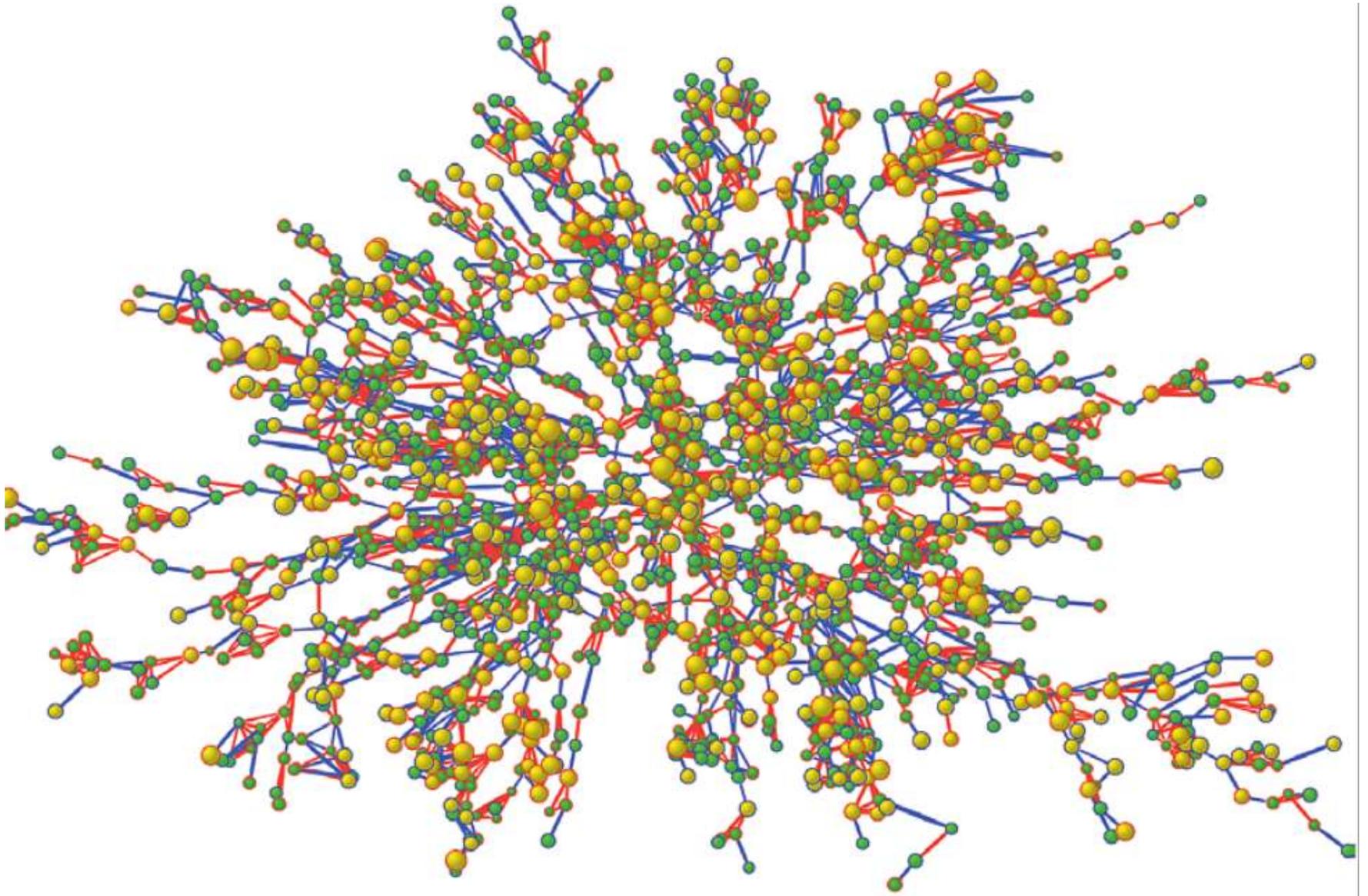
“Chains of Affection”



“Chains of Affection”

Probably
not a future
computer
scientist 😊





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

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, ...

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



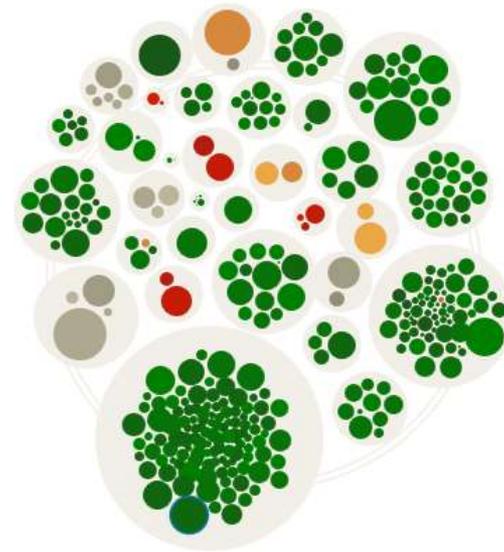
Graph software Tools

- 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>

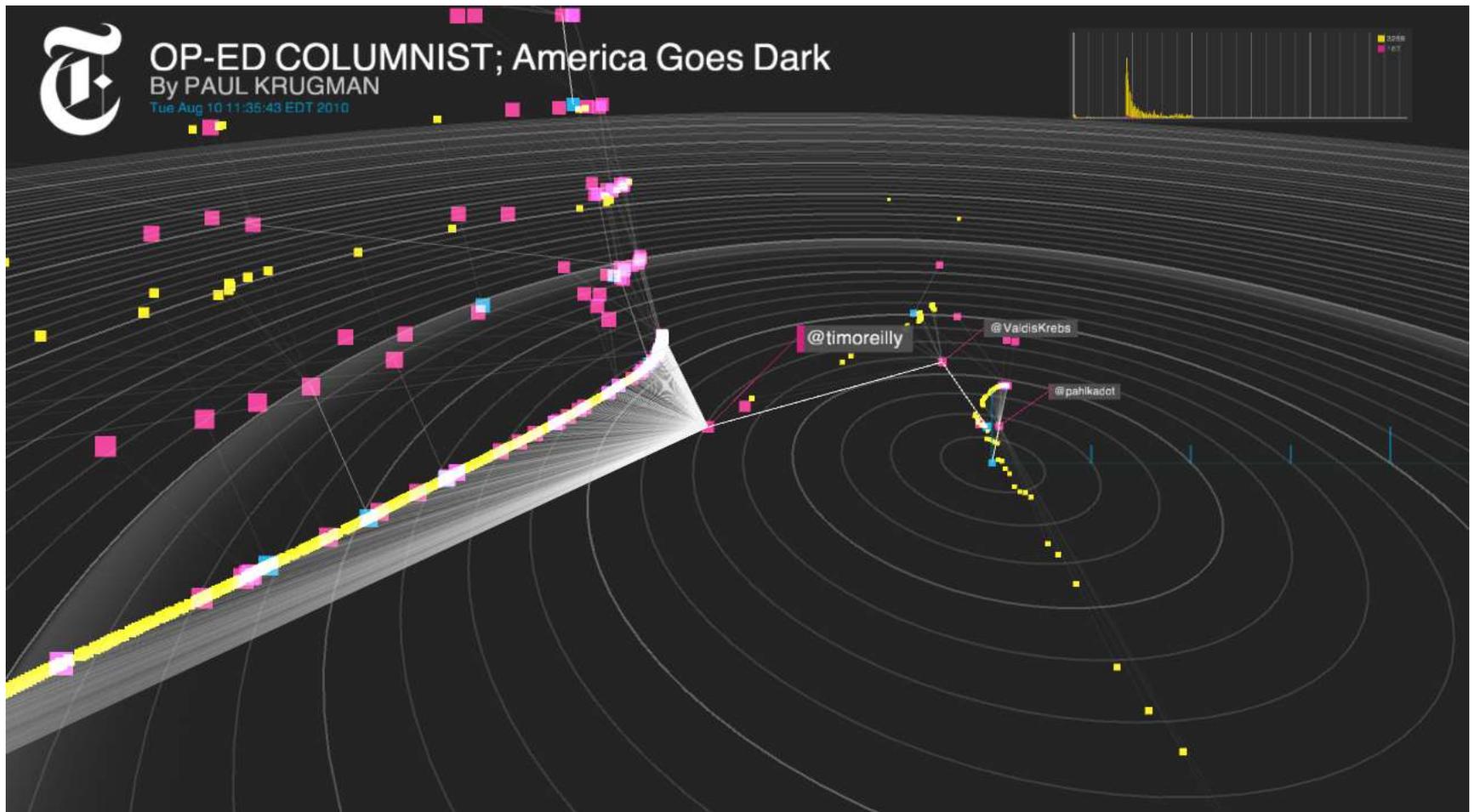
Propagation software tools

- 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

Visualization



Visualization



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!