COMP4650 Doc Analysis

Social Media Analysis (1 of 4)

Lexing Xie
Computer Science, ANU

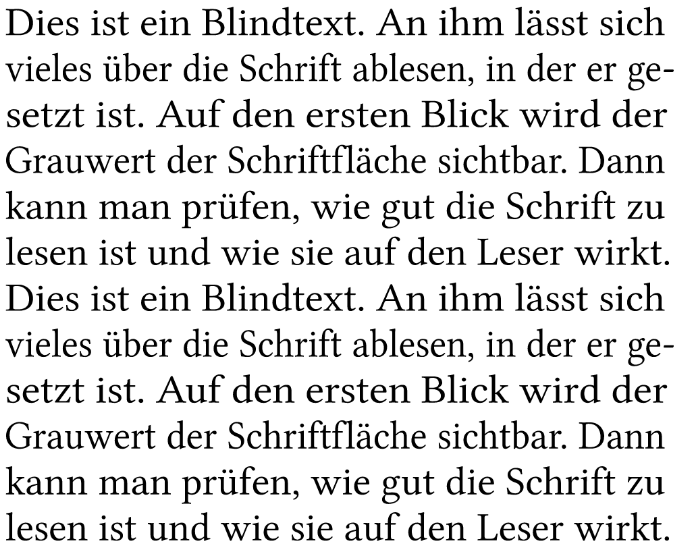
Lecture slides credit: Lada Adamic, Univ. Michigan
Jure Leskovec, Stanford University

Networks … of Documents

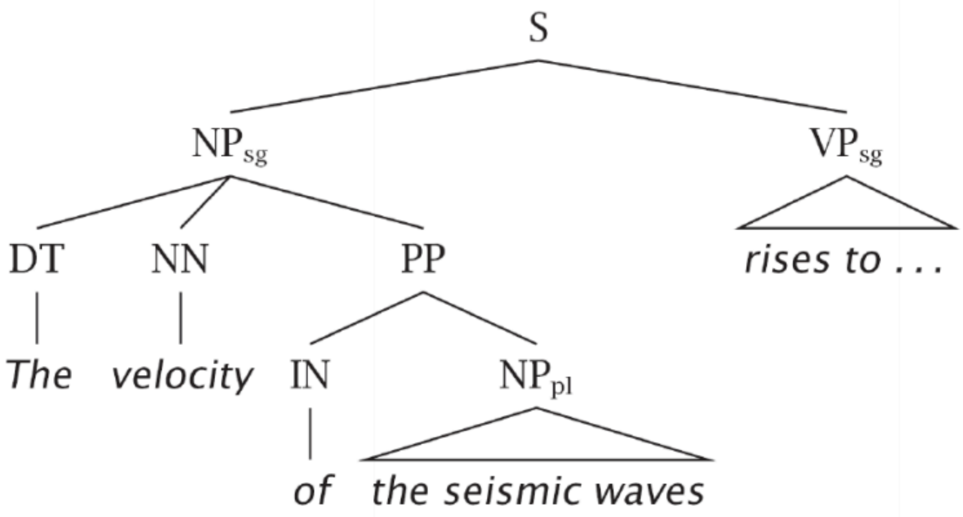
- The rest of this class covers
  - Extracting document elements: nouns/verbs, subjects/object, web page structure …
  - Labelling and matching documents: IR, document classification, clustering …

- What if we look at …
  - Documents and their relations to each other?
  - Document elements and their relations to each other?
  - People and their relations to each other

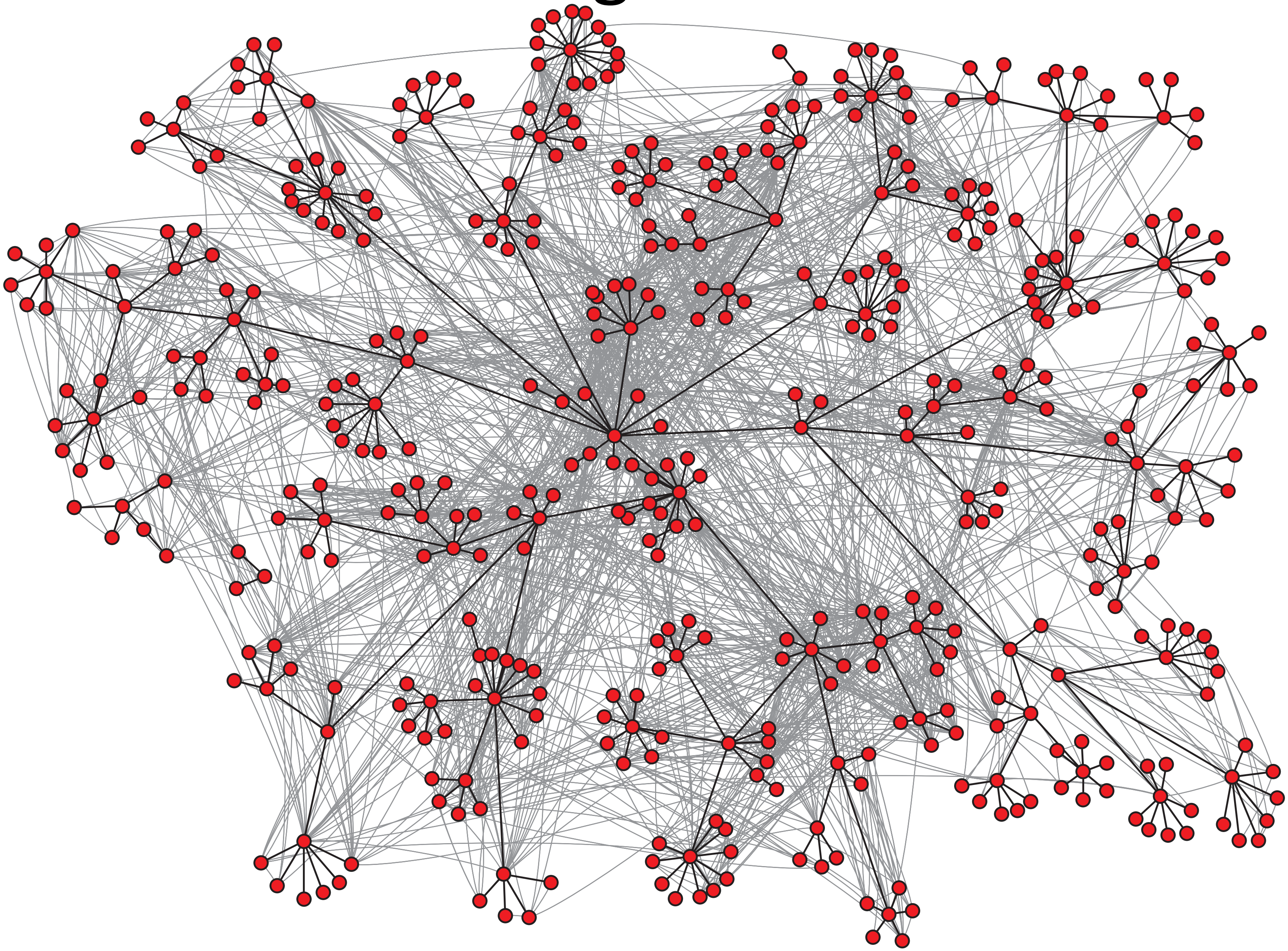
Earlier in This Course …



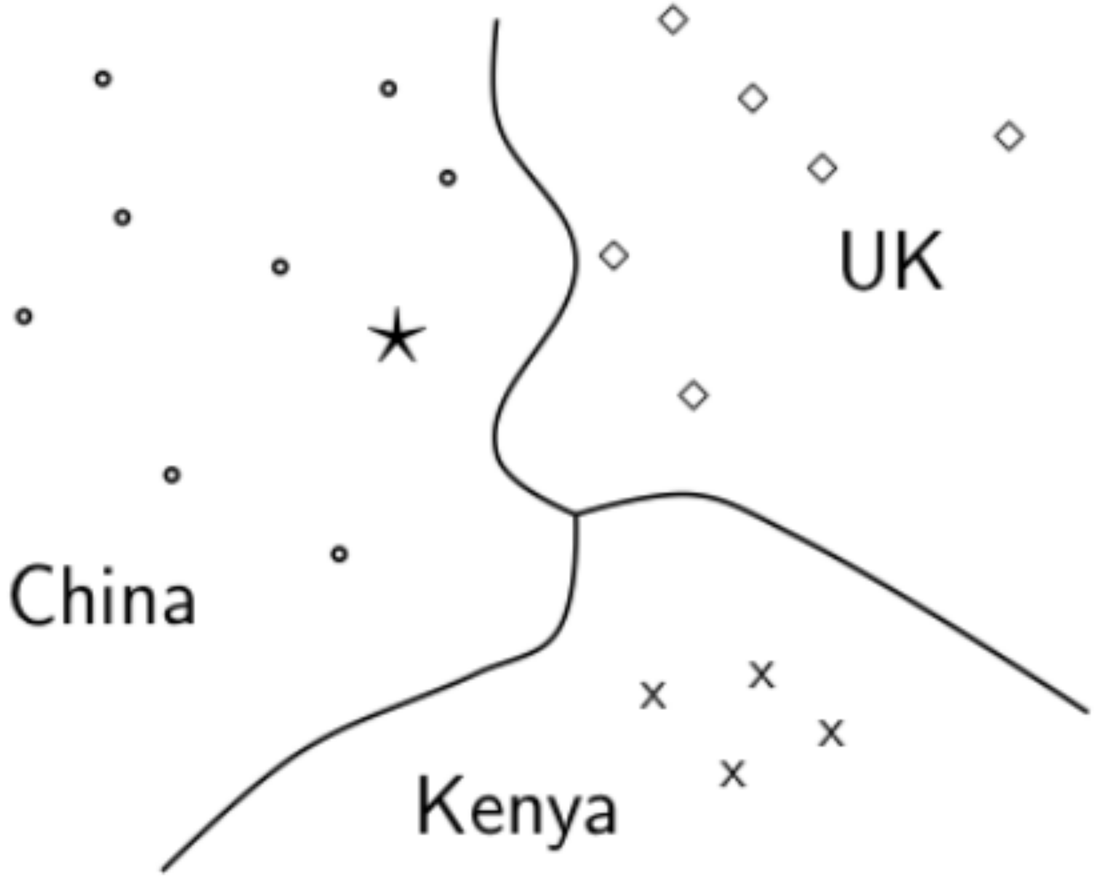
Dies ist ein Blindtext. An ihm lässt sich vieles über die Schrift ablesen, in der er gesetzt ist. Auf den ersten Blick wird der Grauwert der Schriftfläche sichtbar. Dann kann man prüfen, wie gut die Schrift zu lesen ist und wie sie auf den Leser wirkt. Dies ist ein Blindtext. An ihm lässt sich vieles über die Schrift ablesen, in der er gesetzt ist. Auf den ersten Blick wird der Grauwert der Schriftfläche sichtbar. Dann kann man prüfen, wie gut die Schrift zu lesen ist und wie sie auf den Leser wirkt.



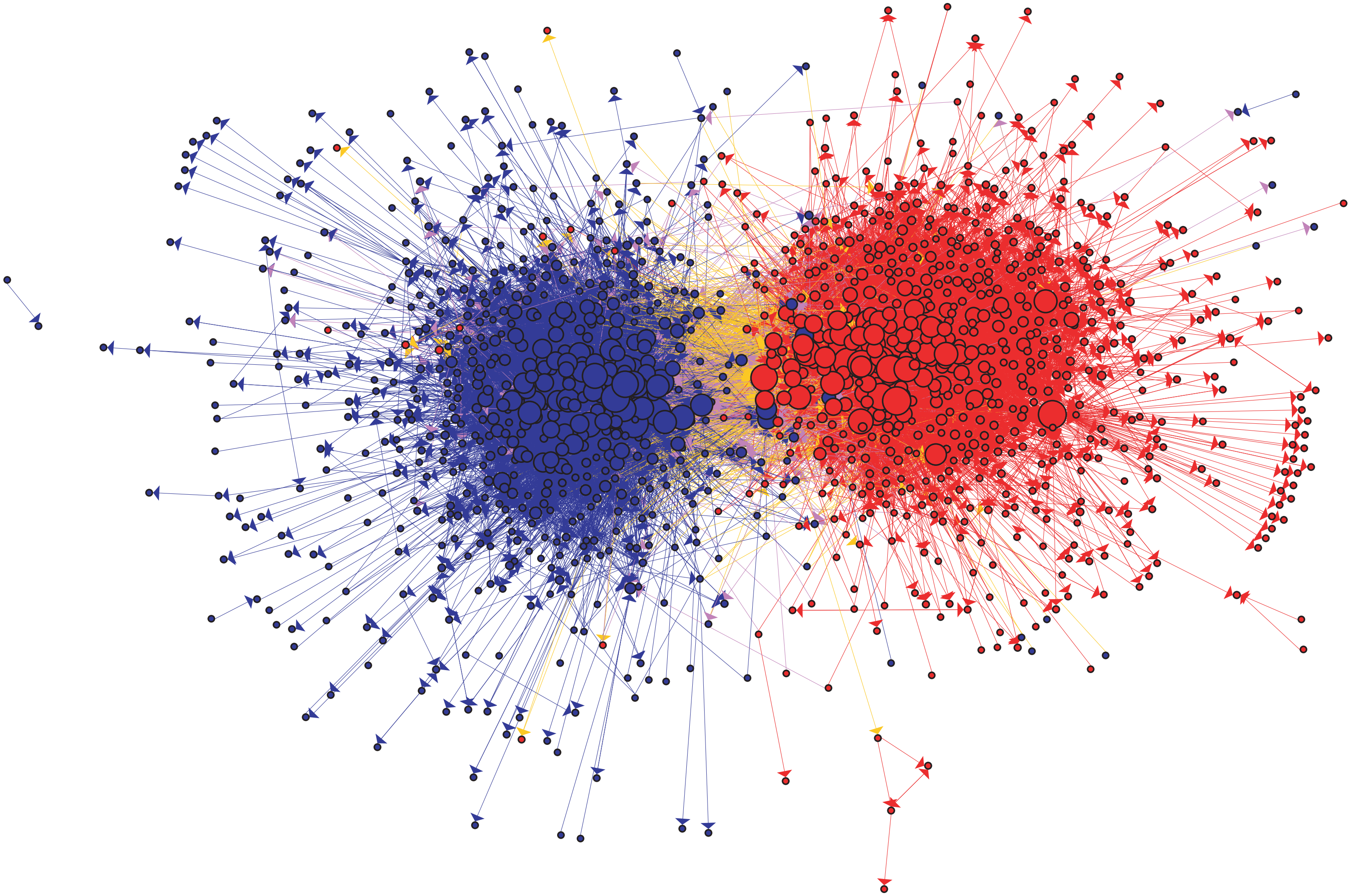
Networks: An organization - HP Labs



Earlier in this Course: classifying documents



Political blogs



Lada A. Adamic and Natalie Glance. 2005. The political blogosphere and the 2004 U.S. election: divided they blog. In Proceedings of the 3rd international workshop on Link discovery (LinkKDD '05)

by Lada Adamic, U Michigan

A View of Facebook via 10 M links

facebook
December 2010

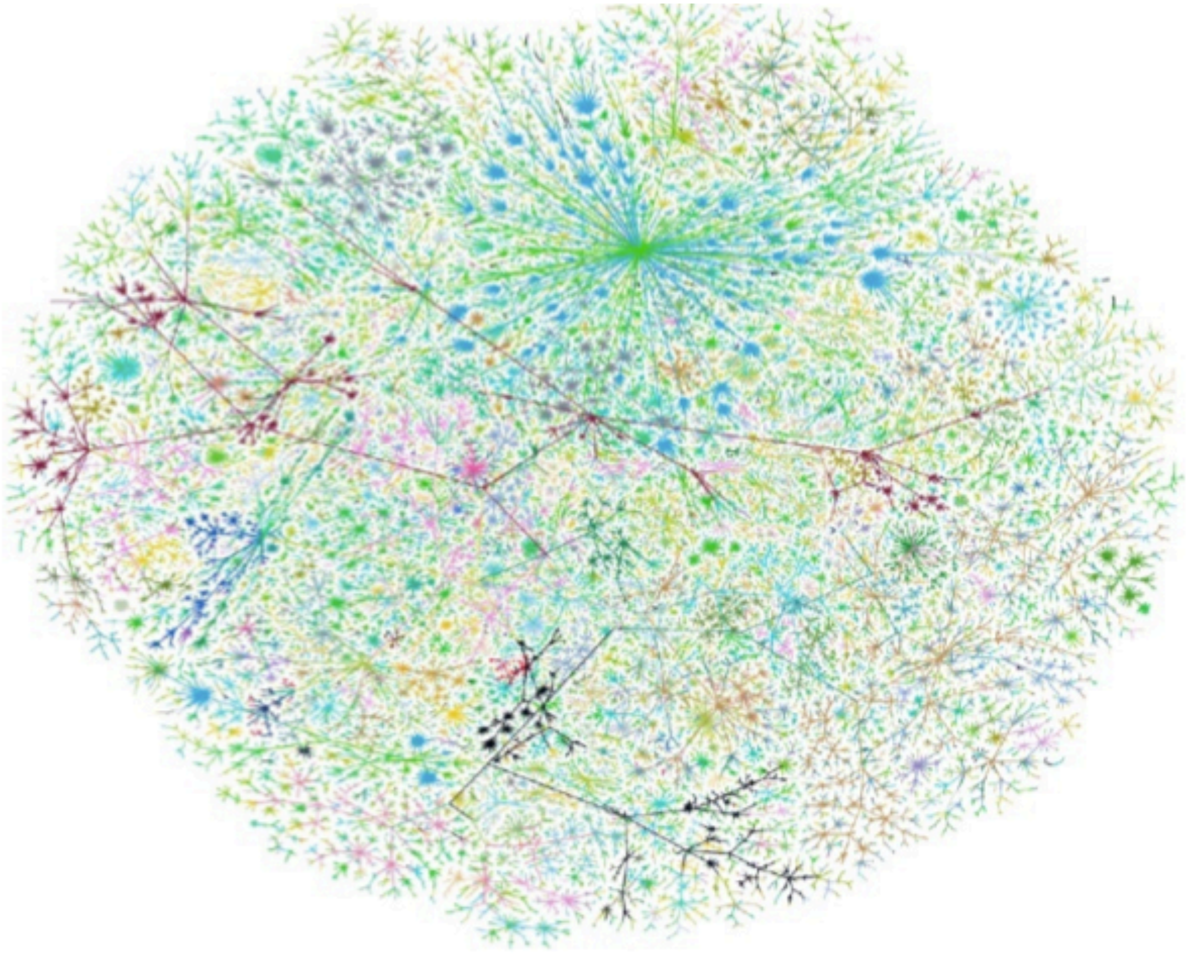
By Paul Butler, https://www.facebook.com/note.php?note_id=469716398919

Why Networks?

- Behind each of these complex systems there is an intricate wiring diagram,
  a network
  that defines the interactions between system components.

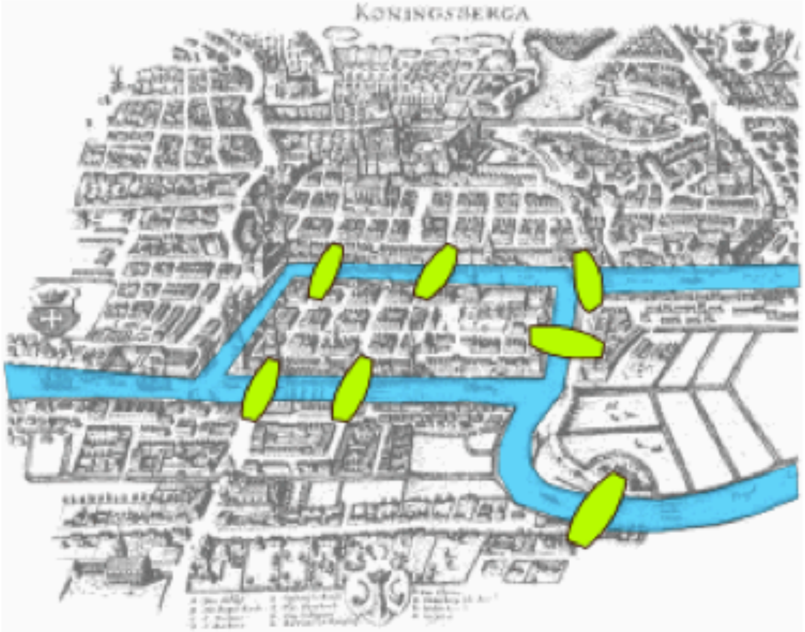
Understanding the network is key to understand the behaviors of such complex system.

Networks: Communications



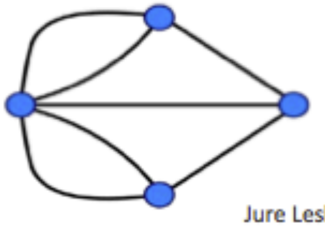
Graph of the Internet (Autonomous Systems)

Networks: Transportation



Seven Bridges of Königsberg (Euler 1735)

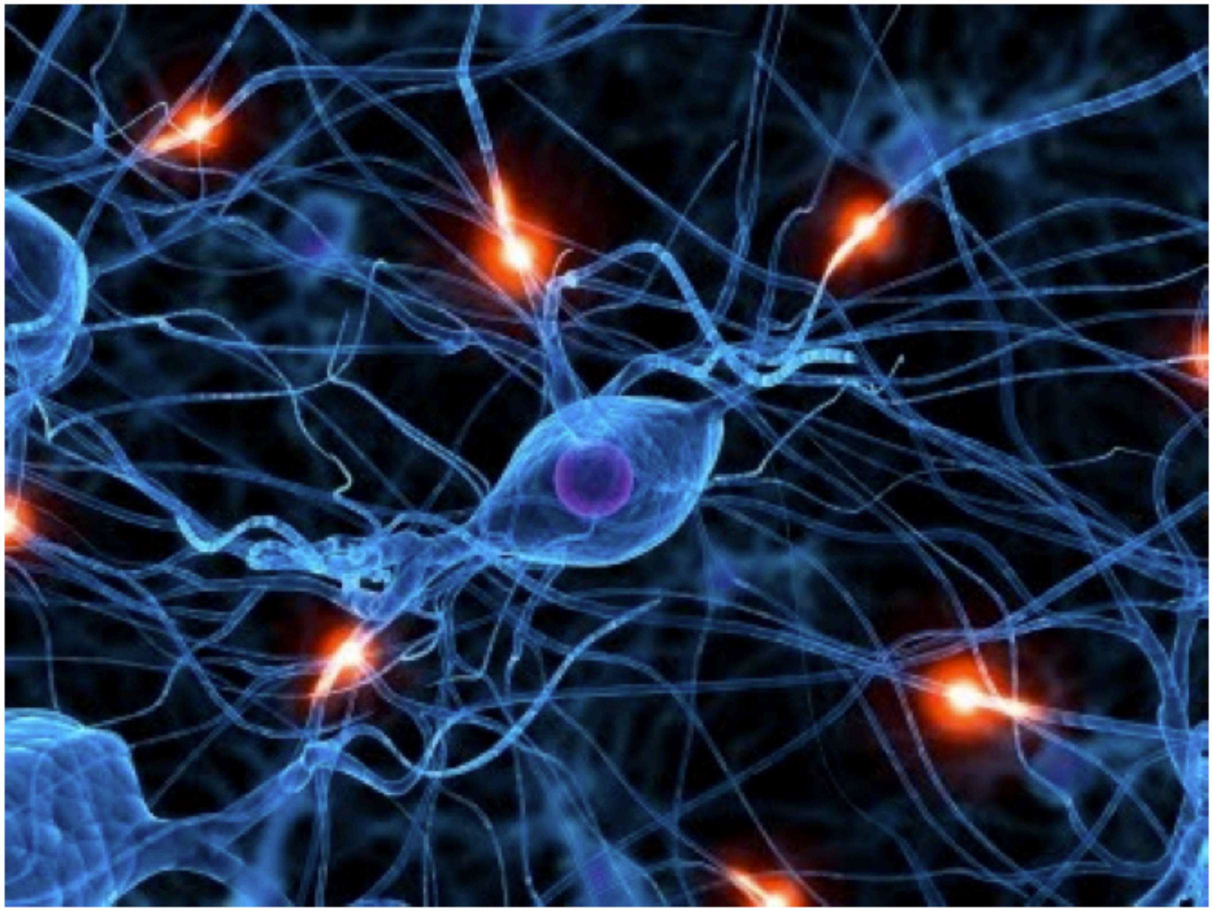
Return to the starting point by traveling each link of the graph once and only once.



Swiss Cottage
St. John's Wood
Kilburn High Road
South Hampstead
Mornington Crescent
King's Cross St. Pancras
Road & Barnsbury
Maida Vale
Paddington
Edgware Road
Baker Street
Great Portland Street
Euston
Angel
Paddington
Edgware Road
Marylebone
Warren Street
Regent's Park
Euston Square
Farringdon
Old Street
Bayswater
Russell Square
Barbican
Holland Park
Notting Hill Gate
Lancaster Gate
Bond Street
Oxford Circus
Goodge Street
Holborn
Chancery Lane
Moorgate
Queensway
Marble Arch
Tottenham Court Road
St. Paul's
Bank
Covent Garden
Cannon Street
High Street Kensington
Hyde Park Corner
Green Park
Piccadilly Circus
Leicester Square
Mansion House
Monument
Knightsbridge
Charing Cross
Blackfriars
Gloucester Road
Sloane Square
St. James's Park
Temple
Earl's Court
South Kensington
Victoria
Westminster
Embankment
London Bridge
Fulham Broadway
Waterloo

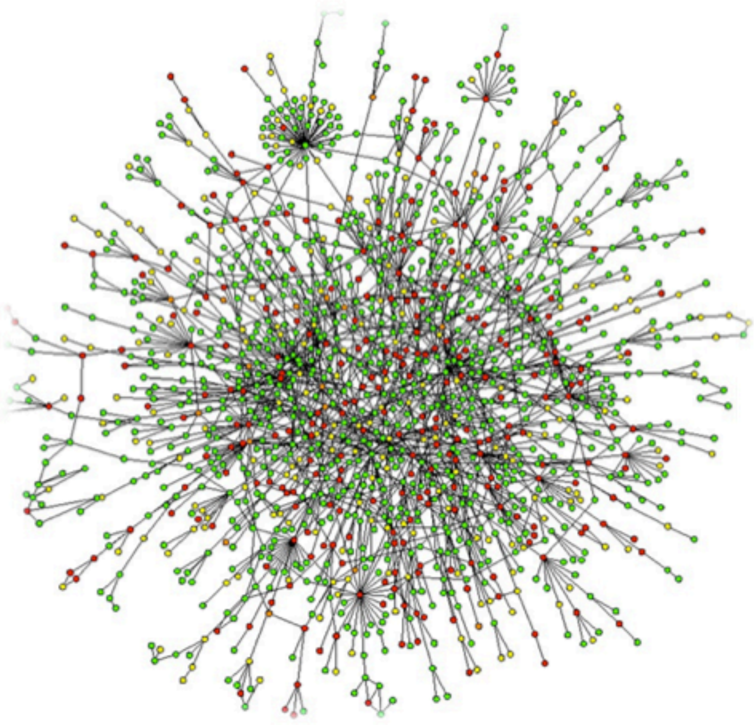
London Underground

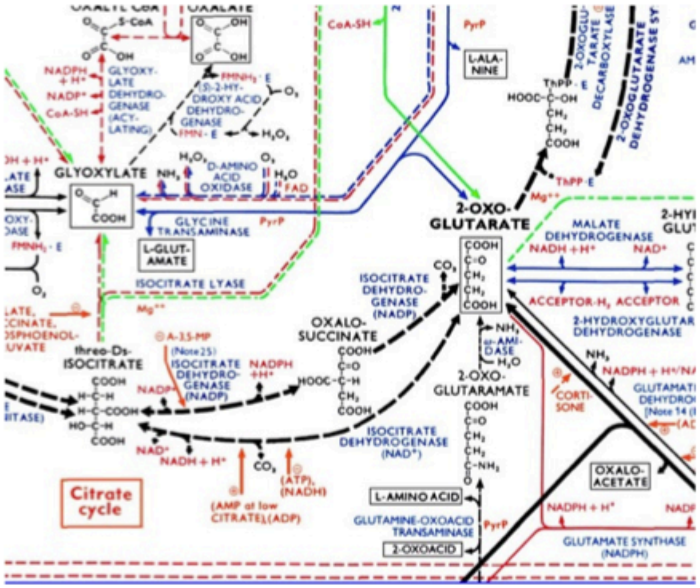
Networks: Brain



Human brain has between 10-100 billion neurons

Networsk: Cells

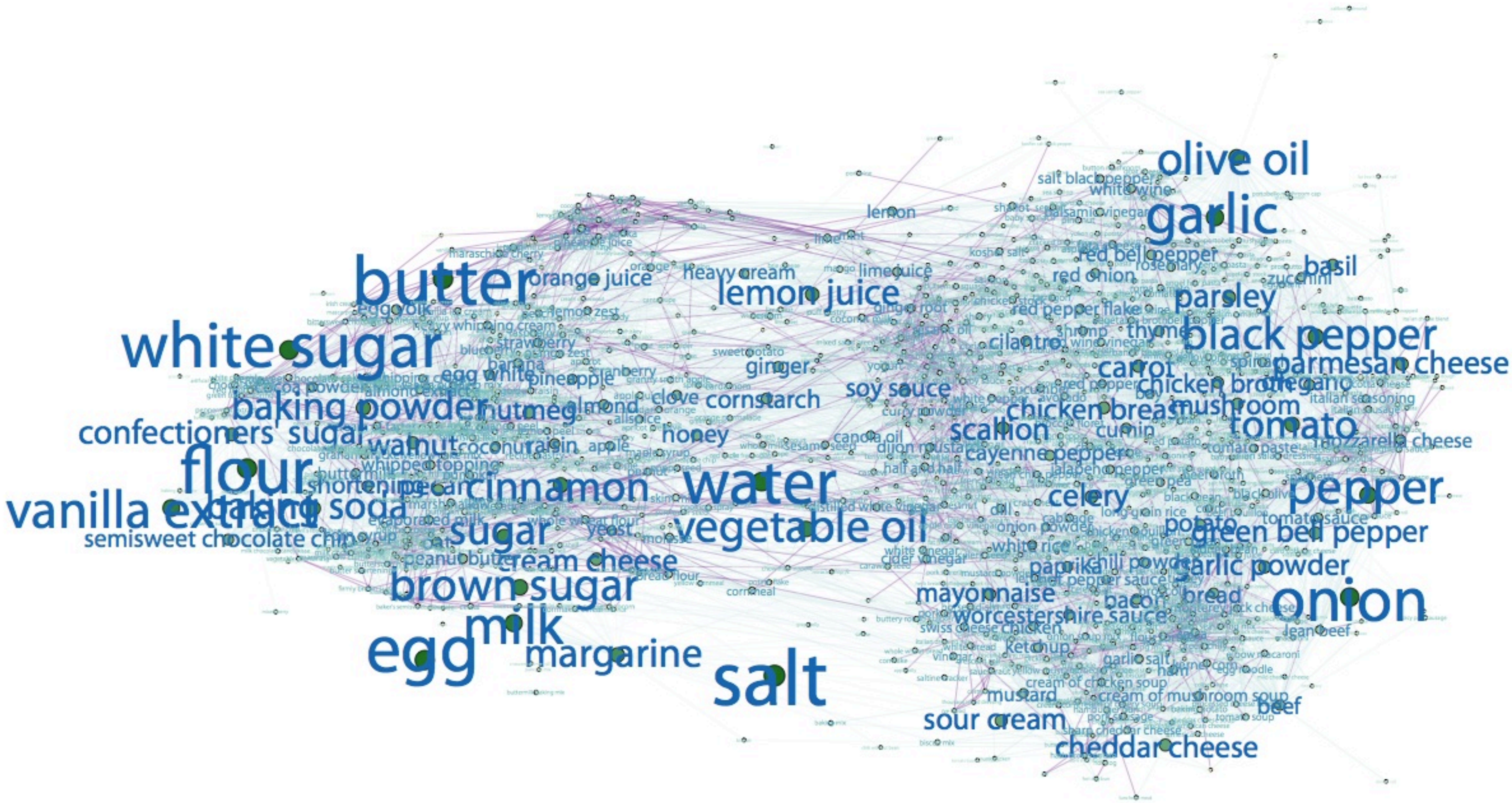



GLYOXYLATE
GLYOXYLATE DEHYDROGENASE (ACYLATING)
(S)-2-HYDROXY ACID DEHYDROGENASE
D-AMINO ACID OXIDASE
GLYCINE TRANSAMINASE
L-GLUTAMATE
L-ALANINE
ISOCITRATE LYASE
2-OXO-GLUTARATE
MALATE DEHYDROGENASE
ACCEPTOR-H₂ ACCEPTOR
2-HYDROXYGLUTARATE DEHYDROGENASE
2-OXOGLUTARATE DECARBOXYLASE
2-OXOGLUTARATE DEHYDROGENASE
ISOCITRATE DEHYDROGENASE (NADP)
OXALOSUCCINATE
threo-Ds-ISOCITRATE
ISOCITRATE DEHYDROGENASE (NAD⁺)
2-OXO-GLUTARAMATE
L-AMINO ACID
GLUTAMINE-OXOACID TRANSAMINASE
2-OXOACID
OXALOACETATE
GLUTAMATE SYNTHASE (NADPH)
CORTISONE
Citrate cycle
(AMP at low CITRATE),(ADP)
(ATP), (NADH)

Protein-Protein Interaction Networks:
Nodes: Proteins
Edges: 'physical' interactoins

Metabolic networks:
Nodes: Metabolites and enzymes
Edges: Chemical reactions

Ingredient networks



Recipe recommendation using ingredient networks. Chun-Yuen Teng, Yu-Ru Lin, Lada A. Adamic, WebSci 2011.

Why Networks?

- Behind each of these complex systems there is an intricate wiring diagram,
  a network
  that defines the interactions between system components.

Understanding the network is key to understand the behaviors of such complex system.

Application domains in network analysis

This class

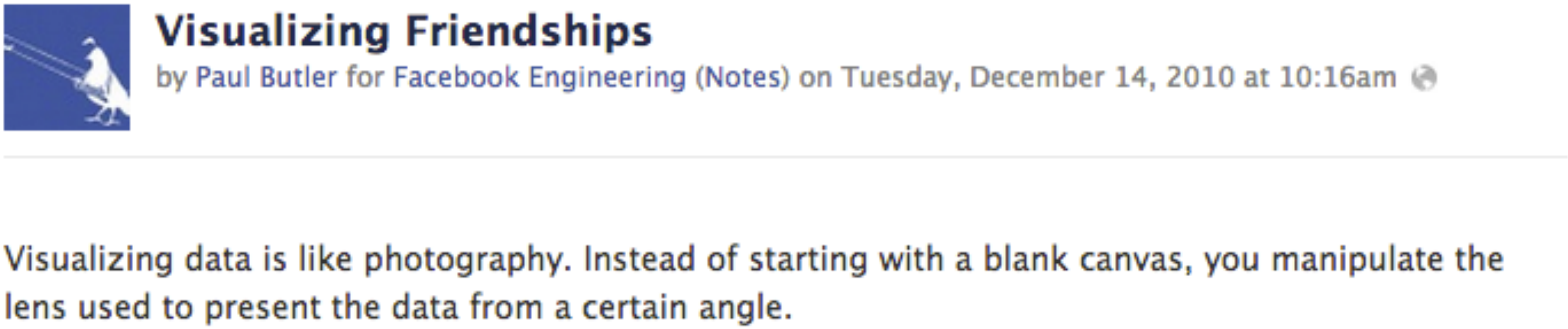
- Social (people-people) networks ✓
- Information networks ✓
- Organization and political networks ✓

- Computer networking
- Biology
- Transportation networks

CSS: Computational challenges

This class

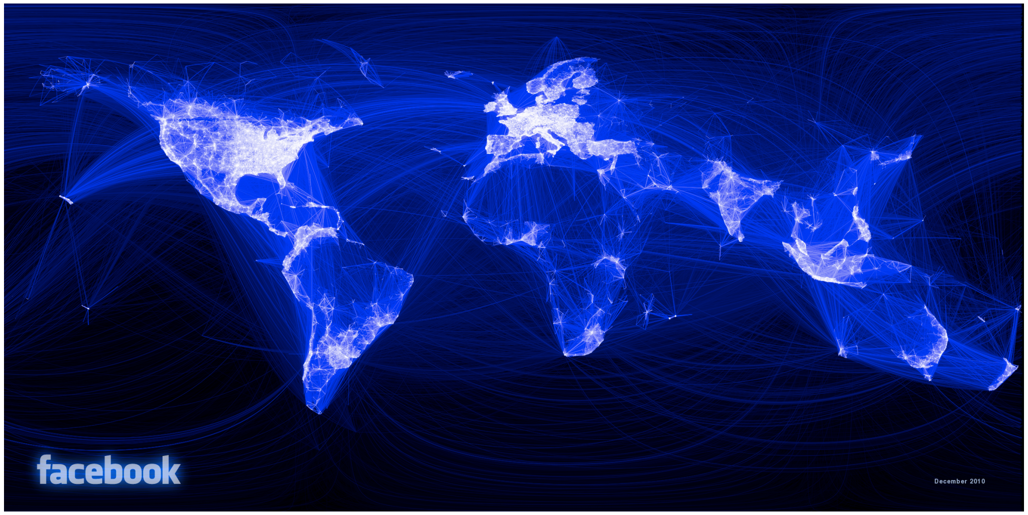
- Machine Learning and applied statistics:
  Predictive modeling, models for network structure, models for dynamic events, optimization ✔
- Natural Language Processing:
  linguistic styles and variations, document similarity, information retrieval ✔
- Visual analytics, visualization ✔
- Computer systems: scalability, reliability, programming languages and tools ✗
- …



Visualizing Friendships

by Paul Butler for Facebook Engineering (Notes) on Tuesday, December 14, 2010 at 10:16am

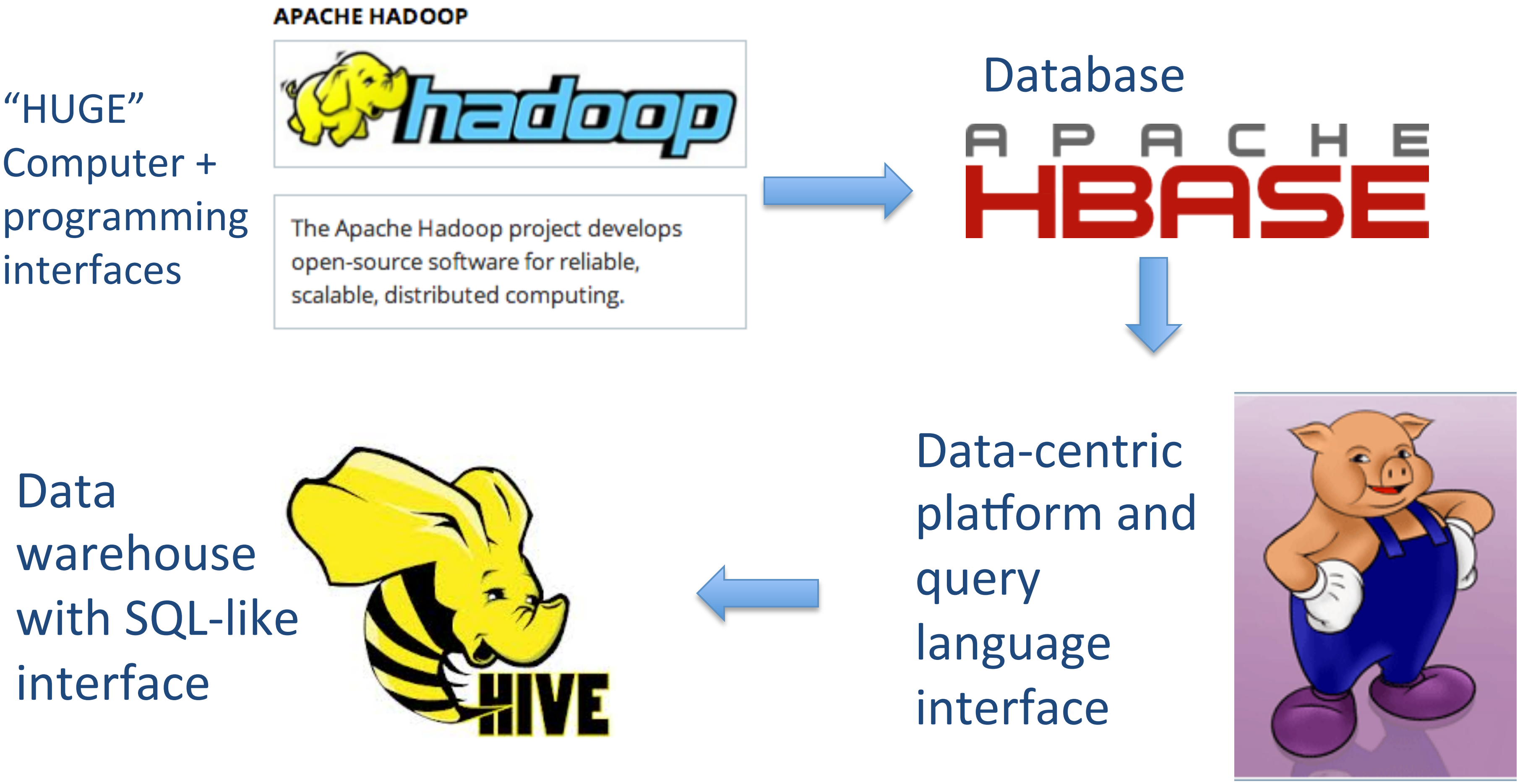
Visualizing data is like photography. Instead of starting with a blank canvas, you manipulate the lens used to present the data from a certain angle.



1. Sample 10 million links from Hive
2. Get city/lat/lon for each node
3. Compute connection strength between cities
4. Render links among city pairs in R
5. Tweak rendering and path drawing until satisfactory

https://www.facebook.com/note.php?note_id=469716398919

Storing and Manipulating 500M Links



"HUGE" Computer + programming interfaces

APACHE HADOOP

The Apache Hadoop project develops open-source software for reliable, scalable, distributed computing.

Database

Data-centric platform and query language interface

Data warehouse with SQL-like interface

http://developer.yahoo.com/blogs/hadoop/posts/2010/08/pig_and_hive_at_yahoo/

Why Social Computing?

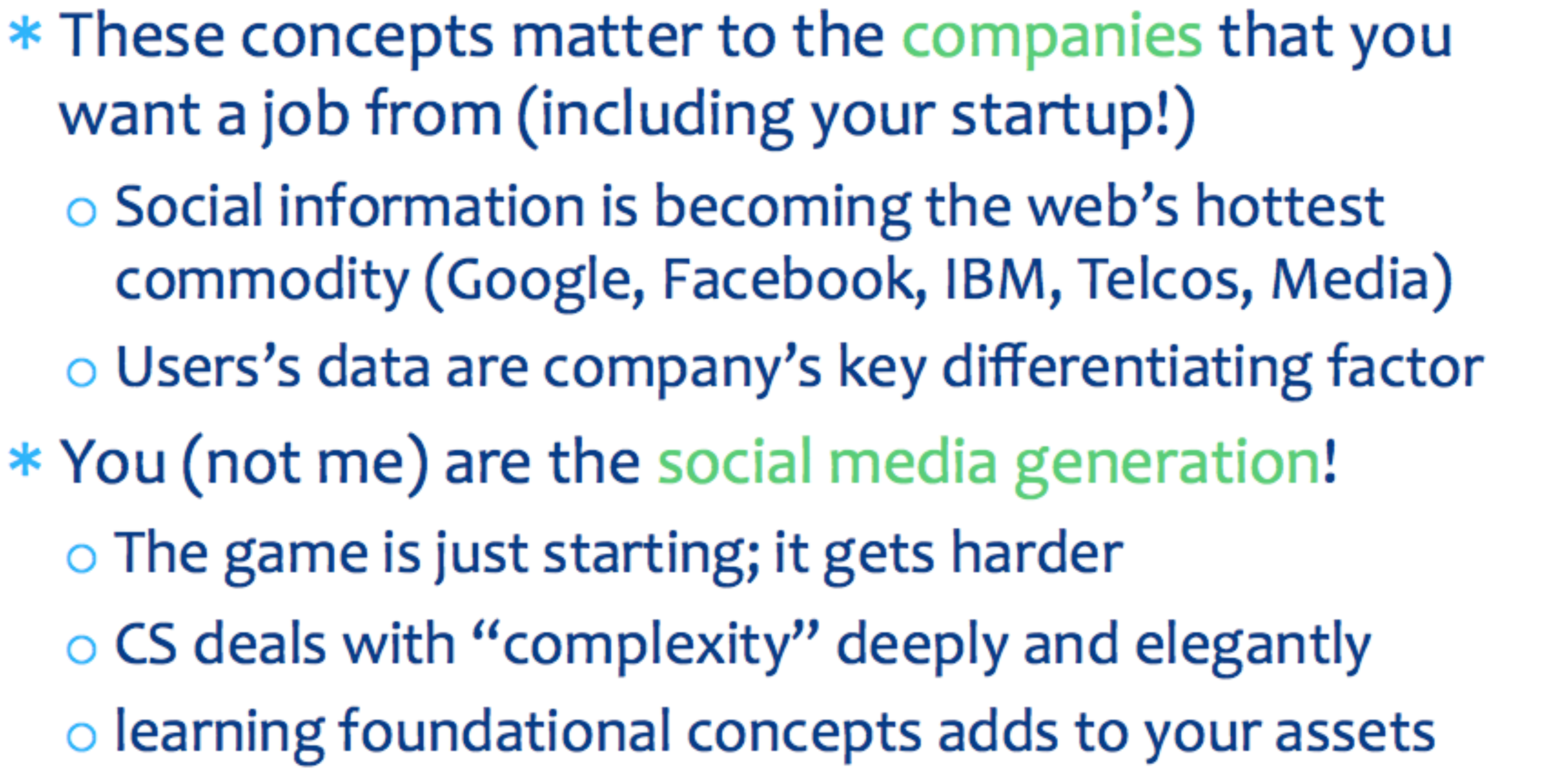
- The next generation could be the one with access to an unprecedented amount of behavioral data
- This can solve real problems

  … not just finding a movie or a restaurant

  - ensuring energy efficiency
  - monitoring our environment
  - reduce inequality
  - informing social decision

Slide by Augustine Chaintreau, Columbia University

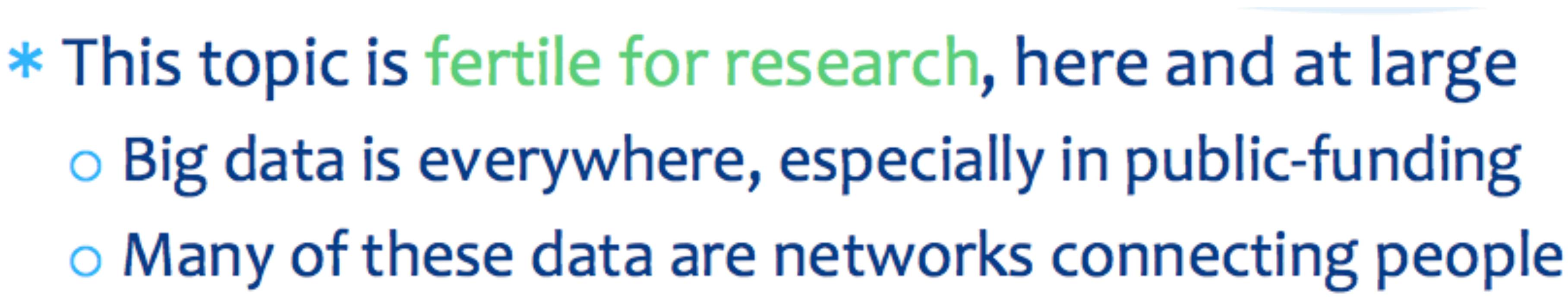
Why should you learn about it? #1



- These concepts matter to the companies that you want a job from (including your startup!)
  - Social information is becoming the web's hottest commodity (Google, Facebook, IBM, Telcos, Media)
  - Users's data are company's key differentiating factor
- You (not me) are the social media generation!
  - The game is just starting; it gets harder
  - CS deals with "complexity" deeply and elegantly
  - learning foundational concepts adds to your assets

Slide by Augustine Chaintreau, Columbia University

Why should you learn about it? #2



* This topic is fertile for research, here and at large
  - Big data is everywhere, especially in public-funding
  - Many of these data are networks connecting people

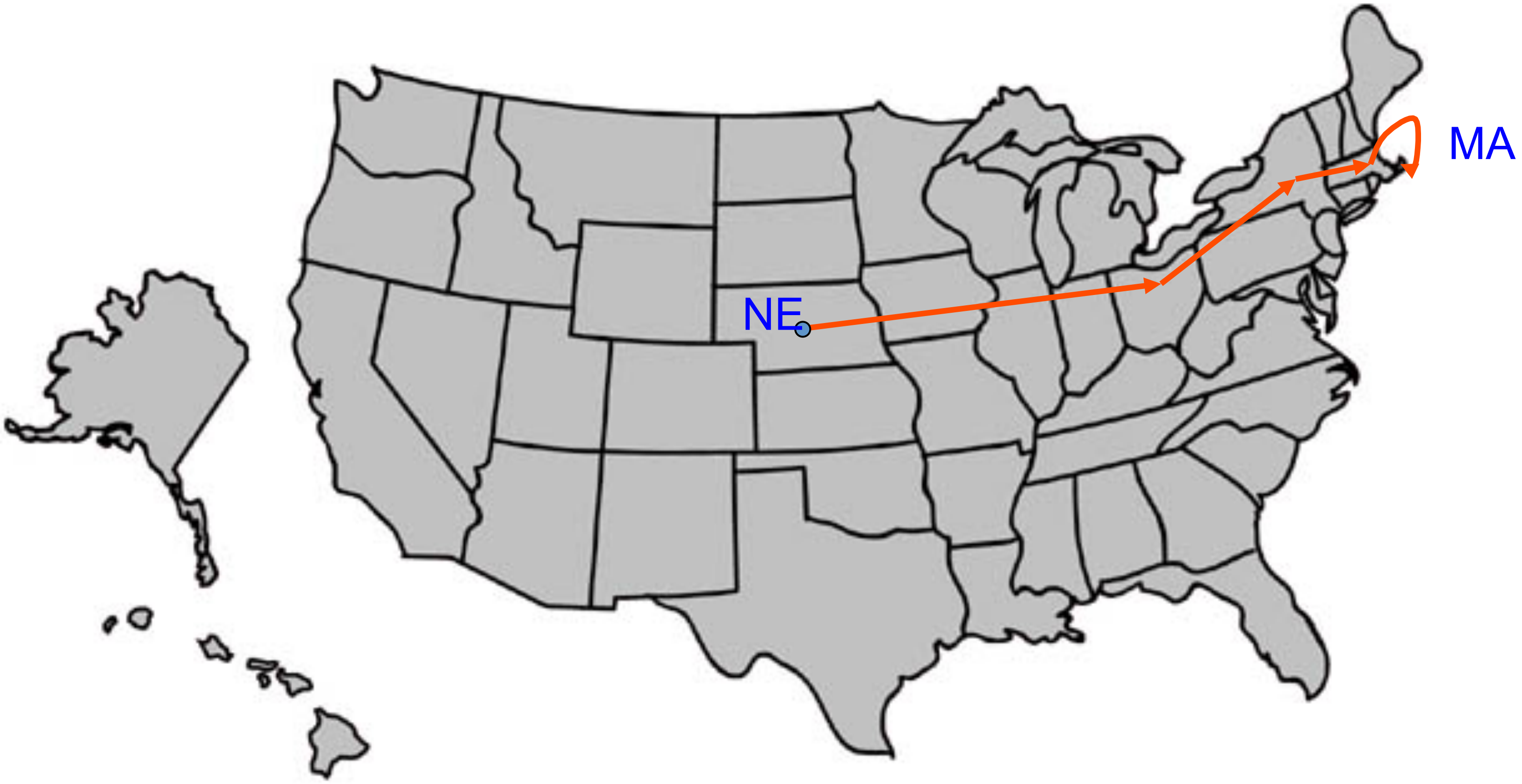
- Many opportunities in/around ANU
  - Big data research theme at CS
  - Machine learning group at NICTA
  - Master degree with Network Science concentration at CASS
  - Overall, great topic to look for an academic job!

Slide by Augustine Chaintreau, Columbia University

What are the classic, solve, and open questions in CSS?

PROBLEMS IN SOCIAL NETWORK

Small-world phenomena: Milgram's Experiment



http://en.wikipedia.org/wiki/Small_world_experiment

1967 -- 1969

Slide by Lada Adamic, U Michigan

Milgram’s experiment

Instructions:
Given a target individual (stockbroker in Boston), pass the message to a person you correspond with who is “closest” to the target.

Outcome:

20% of initiated chains reached target
average chain length = 6.5

- “Six degrees of separation”

Slide by Lada Adamic, U Michigan

Milgram’s experiment repeated

email experiment
Dodds, Muhamad, Watts,
Science 301, (2003)
(optional reading)

- 18 targets
- 13 different countries

- 60,000+ participants
- 24,163 message chains
- 384 reached their targets
- average path length 4.0

Source: NASA, U.S. Government; http://visibleearth.nasa.gov/view_rec.php?id=2429

Slide by Lada Adamic, U Michigan

- Is 6 is a ***surprising*** number?

  In the 1960s? Today? Why?

  What is the mechanism behind "small-world" networks?

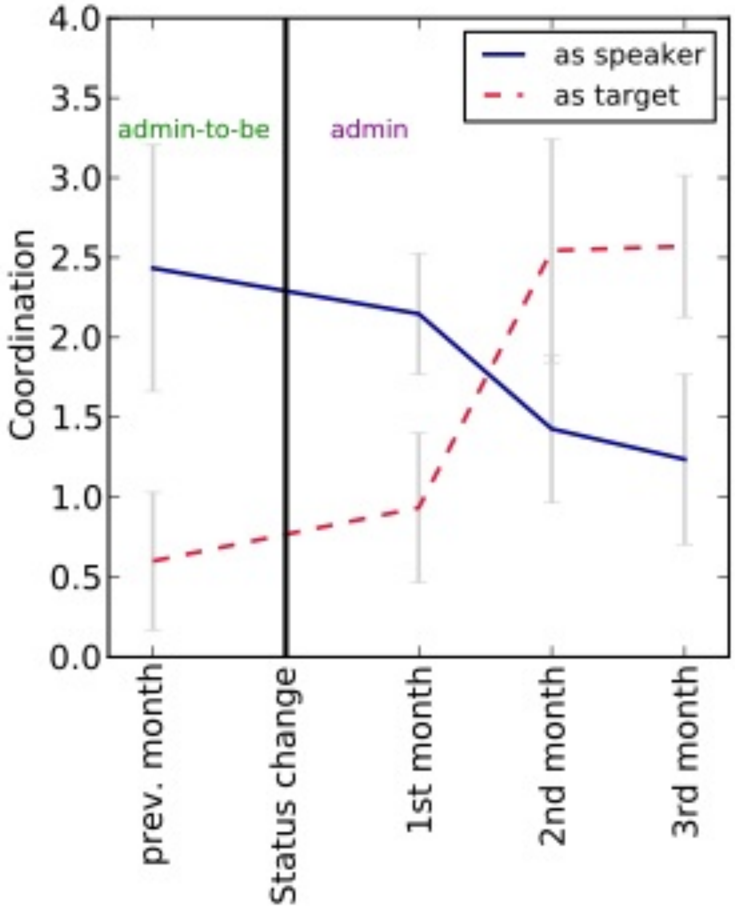
Link prediction

“Given a snapshot of a social network, can we infer which new interactions among its members are likely to occur in the near future? “

What are the measurements?
-- “proximity” and “similarity” between two unconnected nodes
What are application domains?
-- social networks, friend recommendation; product / webpage recommendation; predicting academic collaborations; predicting merger and acquisitions …
How to measure performance?
-- use future held-out data, conversion rate, …

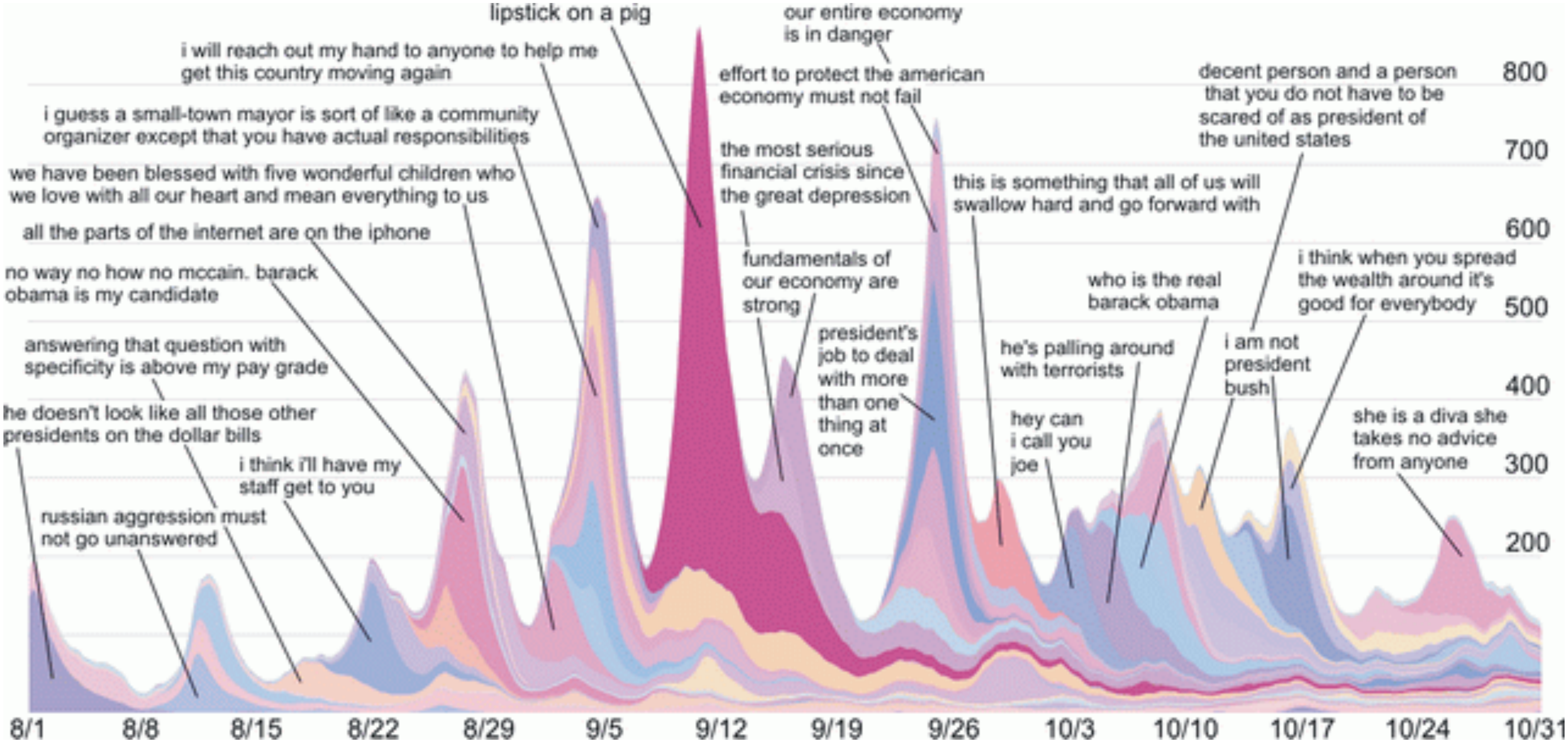
Liben-Nowell, D. and Kleinberg, J. The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology, 58(7) 1019{1031 (2007)

Language Use in Social Media



Echoes of power: Language effects and power differences in social interaction
Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang and Jon Kleinberg.
Proceedings of WWW, 2012.

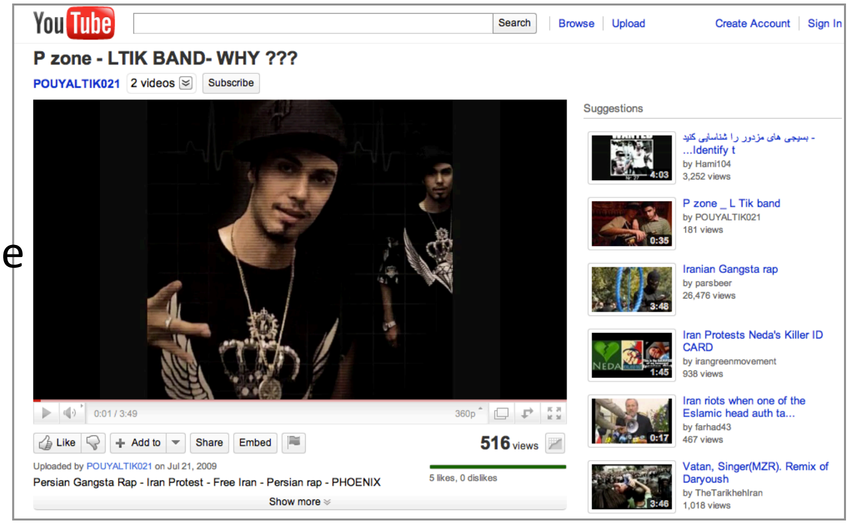
Tracking Memes



Remixing on YouTube – “Iran” topic

Youtube
Video page

Video A



P zone - LTIK BAND- WHY ???

POUYALTIK021 2 videos

516 views

Uploaded by POUYALTIK021 on Jul 21, 2009

Persian Gangsta Rap - Iran Protest - Free Iran - Persian rap - PHOENIX

5 likes, 0 dislikes

بسیجی های مزدور را شناسایی کنید - ...Identify t
by Hami104
3,252 views

P zone _ L Tik band
by POUYALTIK021
181 views

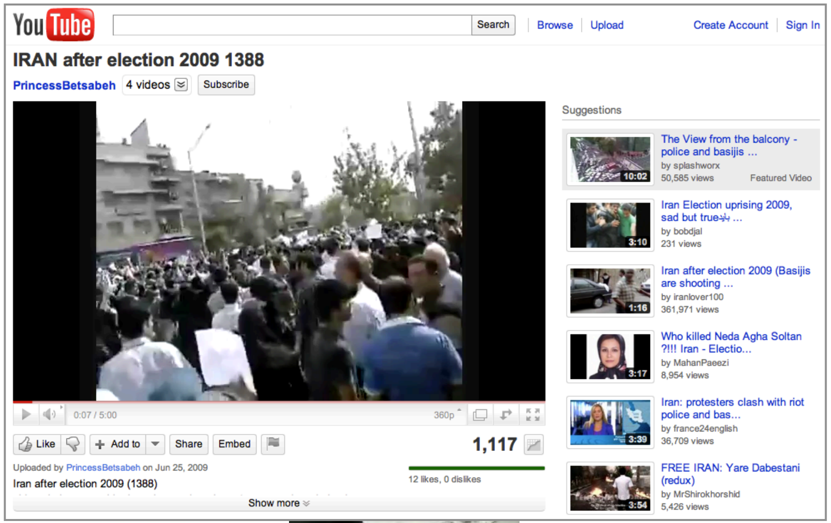
Iranian Gangsta rap
by parsbeer
26,476 views

Iran Protests Neda's Killer ID CARD
by irangreenmovement
938 views

Iran riots when one of the Eslamic head auth ta...
by farhad43
467 views

Vatan, Singer(MZR). Remix of Daryoush
by TheTarikhehIran
1,018 views

Video B



IRAN after election 2009 1388

PrincessBetsabeh 4 videos

1,117

Uploaded by PrincessBetsabeh on Jun 25, 2009

Iran after election 2009 (1388)

12 likes, 0 dislikes

The View from the balcony - police and basijis ...
by splashworx
50,585 views Featured Video

Iran Election uprising 2009, sad but trueباید ...
by bobdjal
231 views

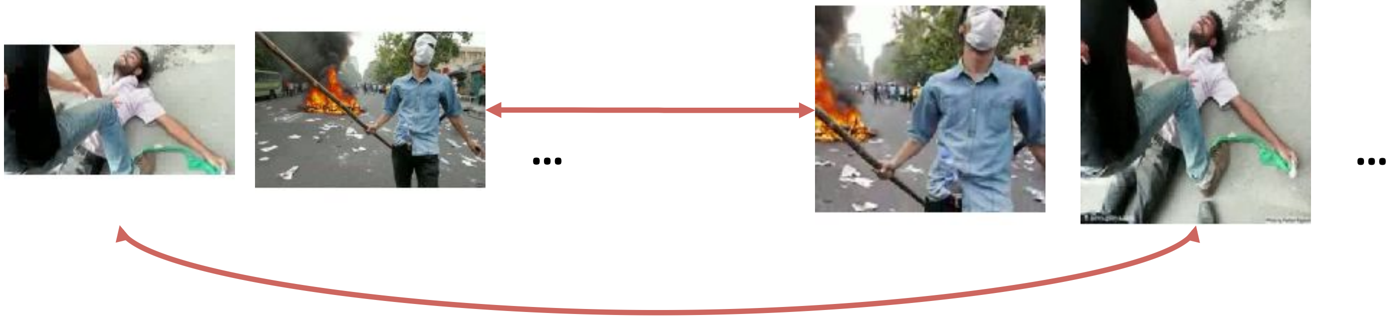
Iran after election 2009 (Basijis are shooting ...
by iranlover100
361,971 views

Who killed Neda Agha Soltan ?!!! Iran - Electio...
by MahanPaeezi
8,954 views

Iran: protesters clash with riot police and bas...
by france24english
36,709 views

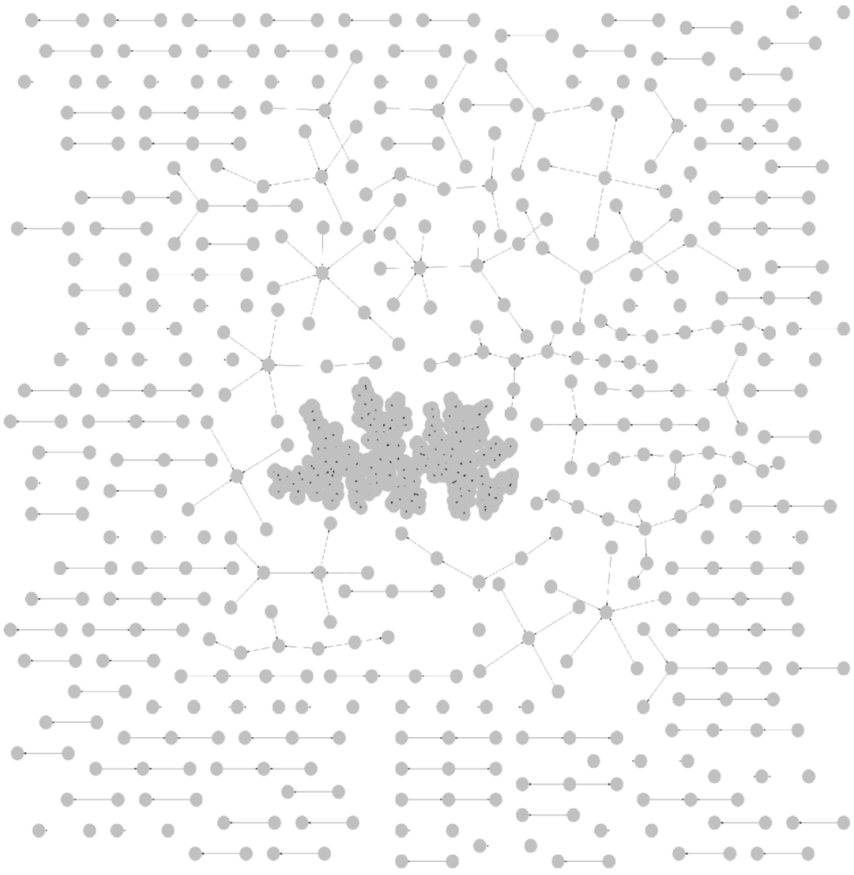
FREE IRAN: Yare Dabestani (redux)
by MrShirokhorshid
5,426 views

Meme
shot
examples

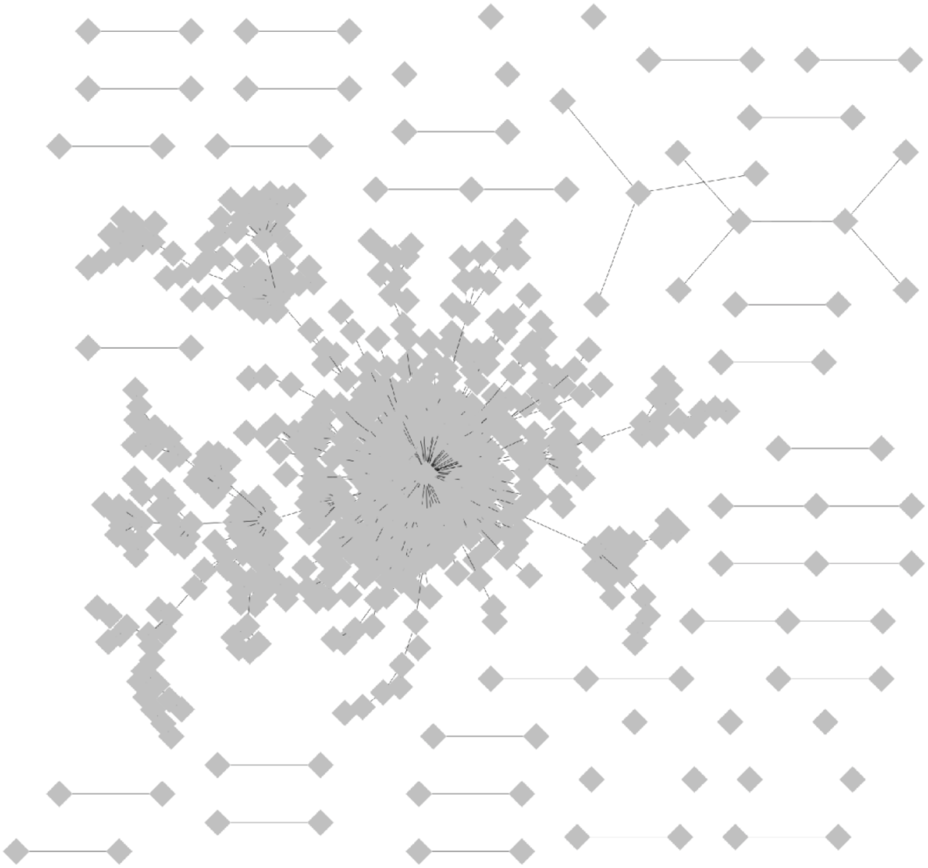


YouTube Remix Network

Video graph



Author graph

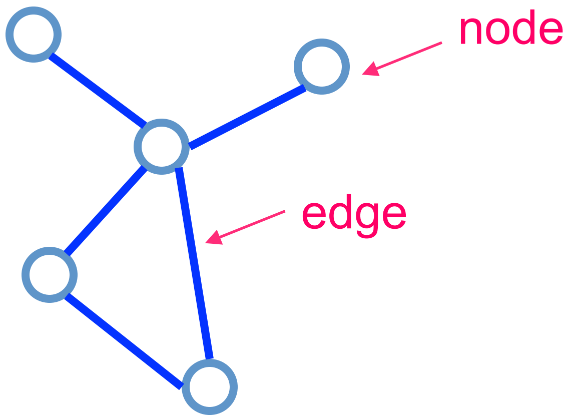


Visual Memes in Social Media: Tracking Real-world News in YouTube Videos (2011), Lexing Xie, Apostol Natsev, Matthew Hill, John Kender, John R Smith, ACM Multimedia 2011, Scottsdale, AZ, USA, Nov 2011

What are networks?

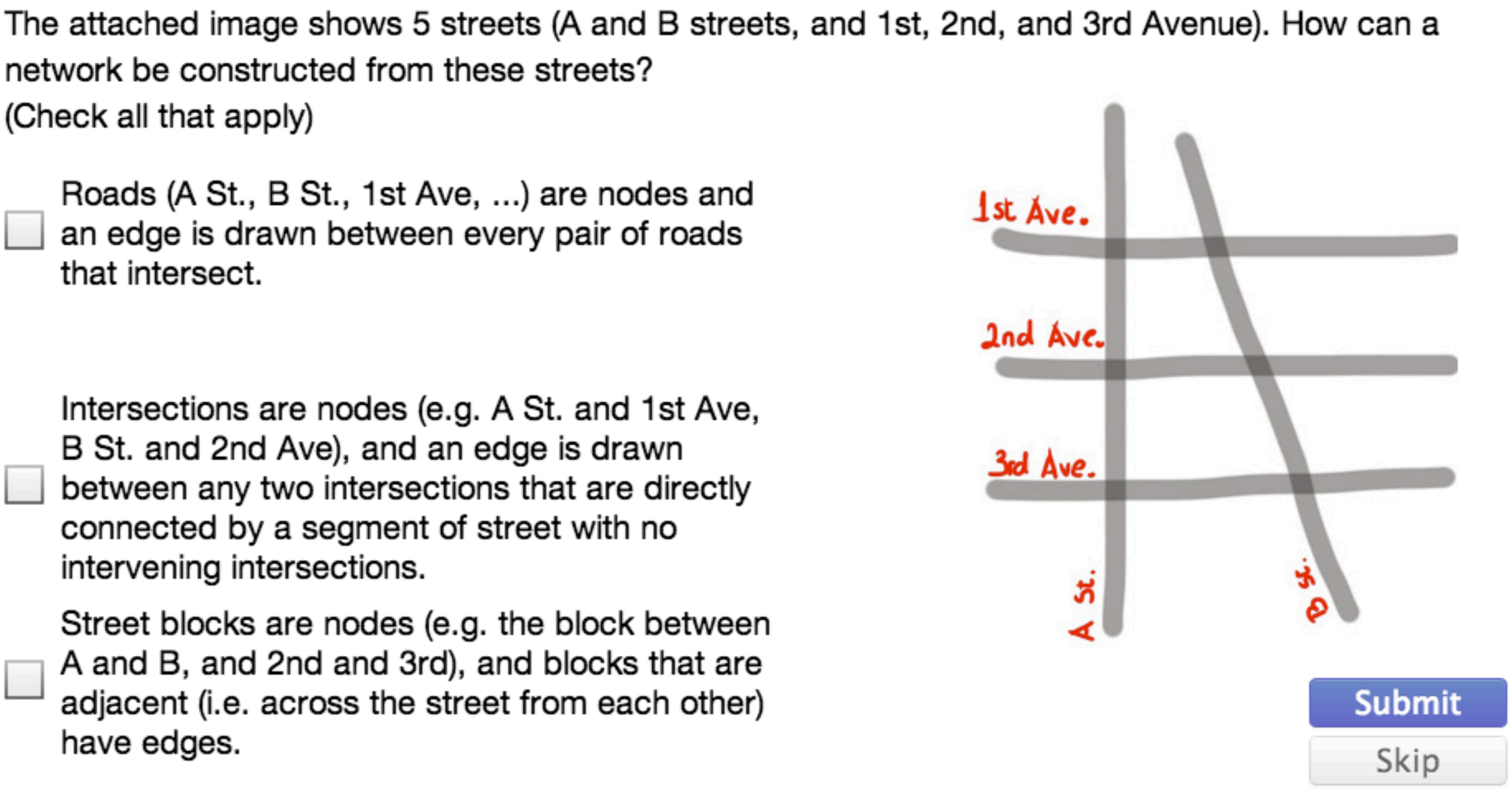
- Networks are sets of nodes connected by edges.

“Network” ≡ “Graph”



points	**lines**	
vertices	edges, arcs	math
nodes	links	computer science
sites	bonds	physics
actors	ties, relations	sociology

Slide by Lada Adamic, U Michigan



The attached image shows 5 streets (A and B streets, and 1st, 2nd, and 3rd Avenue). How can a network be constructed from these streets?

(Check all that apply)

□ Roads (A St., B St., 1st Ave, ...) are nodes and an edge is drawn between every pair of roads that intersect.

□ Intersections are nodes (e.g. A St. and 1st Ave, B St. and 2nd Ave), and an edge is drawn between any two intersections that are directly connected by a segment of street with no intervening intersections.

□ Street blocks are nodes (e.g. the block between A and B, and 2nd and 3rd), and blocks that are adjacent (i.e. across the street from each other) have edges.

https://www.coursera.org/course/sna

Network elements: edges

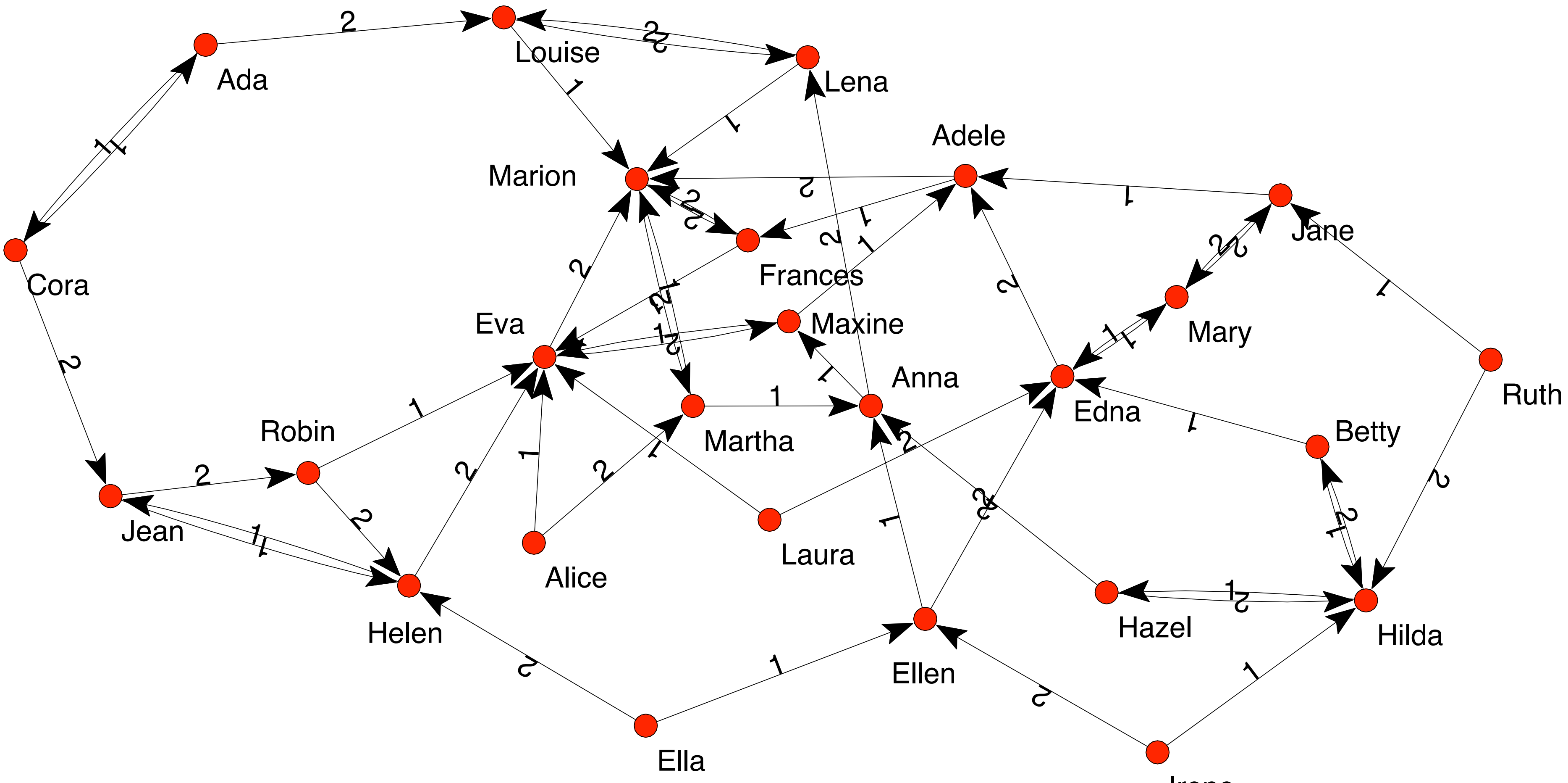
- Directed (also called arcs, links)
  - A -> B
    - A likes B, A gave a gift to B, A is B' s child
- Undirected
  - A <-> B or A – B
    - A and B like each other
    - A and B are siblings
    - A and B are co-authors

Edge attributes

- Examples
  - weight (e.g. frequency of communication)
  - ranking (best friend, second best friend…)
  - type (friend, relative, co-worker)
  - properties depending on the structure of the rest of the graph: e.g. betweenness

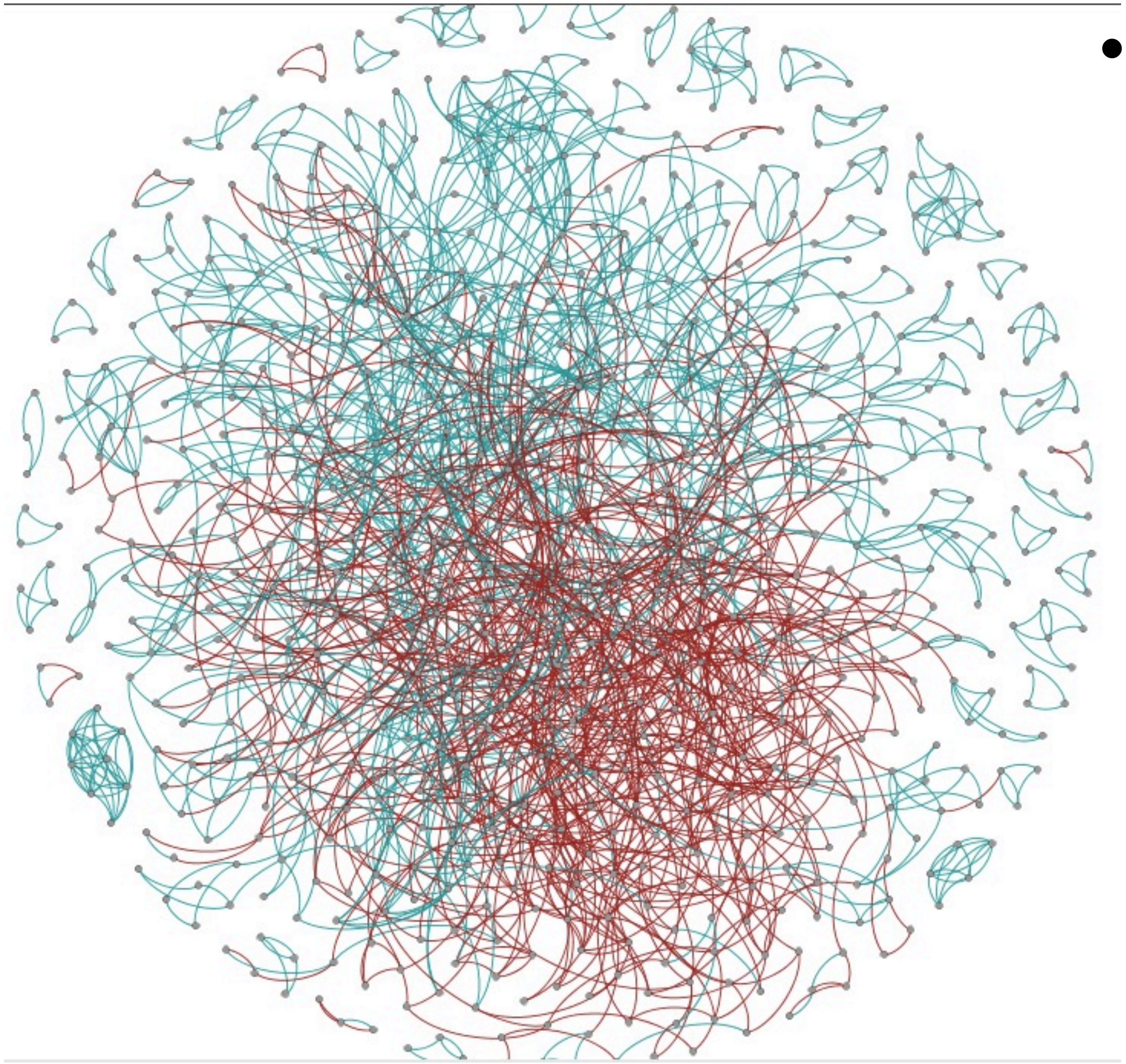
Directed networks

- girls’ school dormitory dining-table partners, 1st and 2nd choices
  (Moreno, *The sociometry reader*, 1960)



by Lada Adamic, U Michigan

Positive and negative weights



- e.g. one person trusting/distrusting another
  - Research challenge: How does one ‘propagate’ negative feelings in a social network? Is my enemy’s enemy my friend?

sample of positive & negative ratings from Epinions network

by Lada Adamic, U Michigan

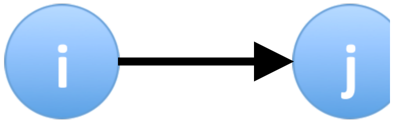
Data representation

- adjacency matrix
- edgelist
- adjacency list

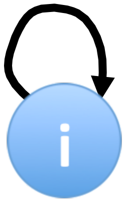
by Lada Adamic, U Michigan

Adjacency matrices

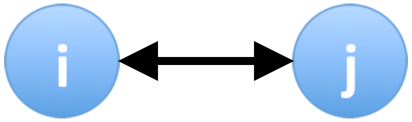
- Representing edges (who is adjacent to whom) as a matrix
  - A_{ij} = 1 if node *i* has an edge to node *j*
    = 0 if node *i* does not have an edge to j



  - A_{ii} = 0 unless the network has self-loops

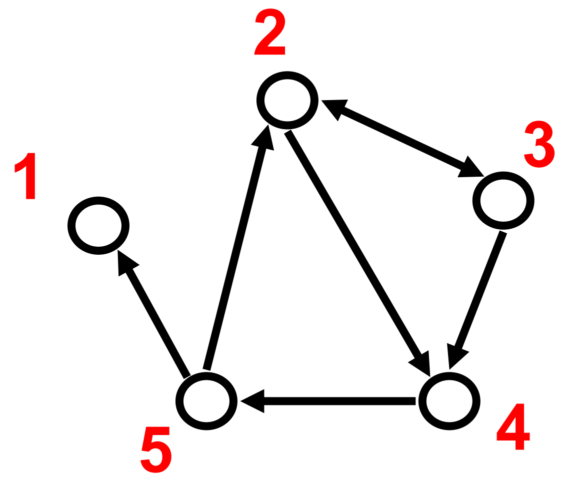


  - $A_{ij} = A_{ji}$ if the network is undirected, or if *i* and *j* share a reciprocated edge



by Lada Adamic, U Michigan

Example adjacency matrix

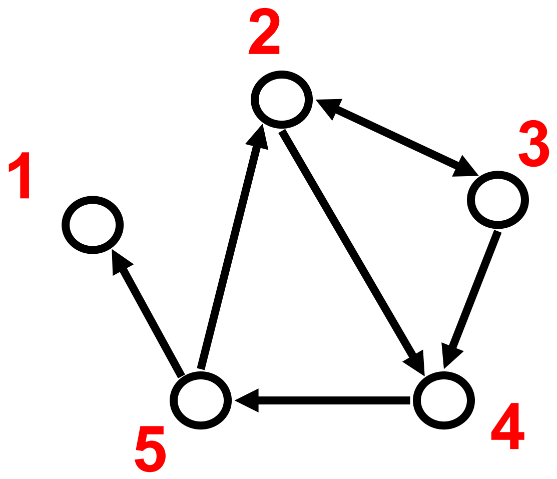


$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$

by Lada Adamic, U Michigan

Edge list

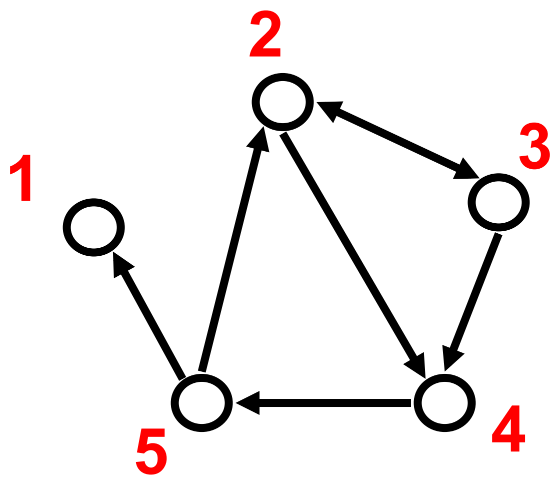
- Edge list
  - 2, 3
  - 2, 4
  - 3, 2
  - 3, 4
  - 4, 5
  - 5, 2
  - 5, 1



by Lada Adamic, U Michigan

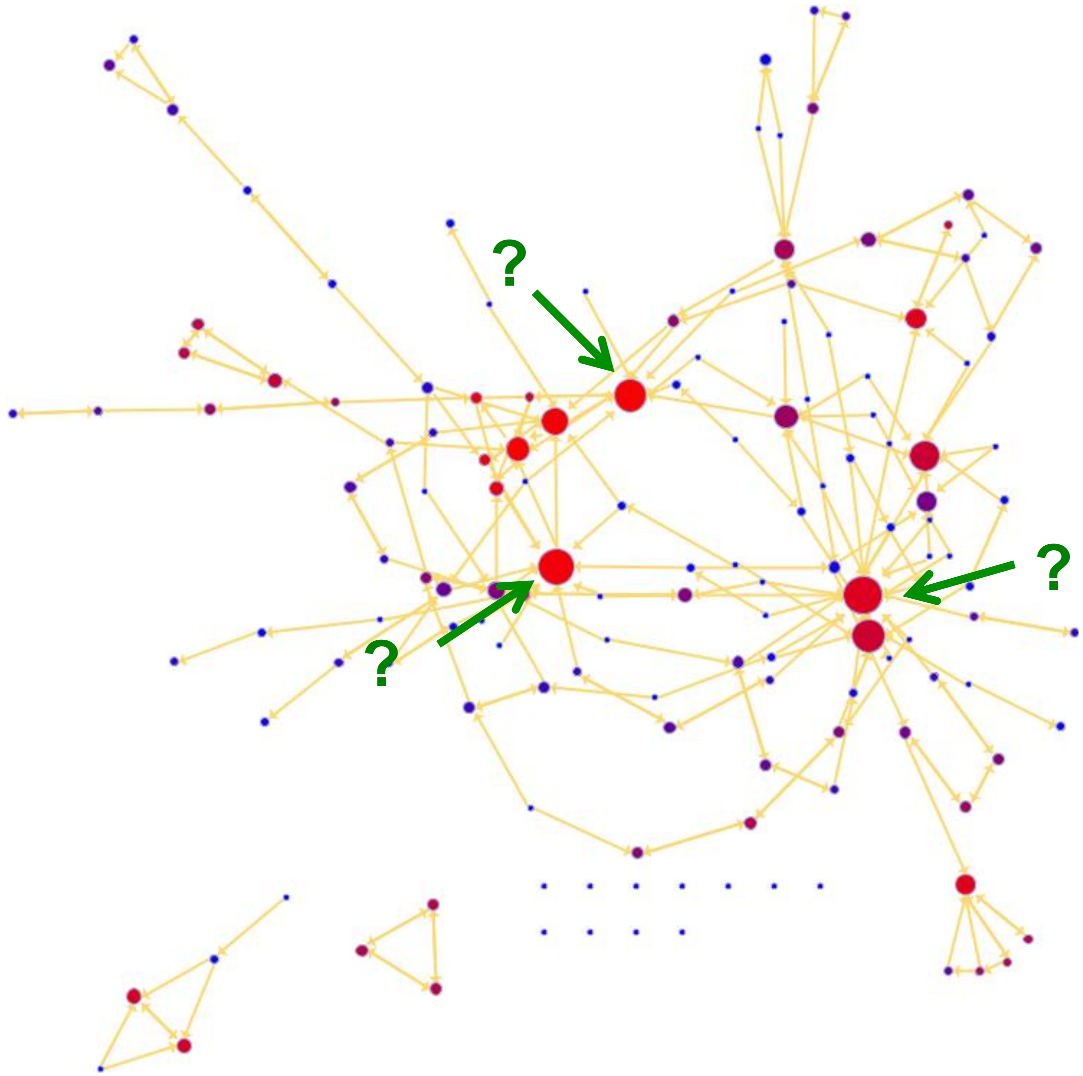
Adjacency lists

- Adjacency list
  - is easier to work with if network is
    - large
    - sparse
  - quickly retrieve all neighbors for a node
    - 1:
    - 2: 3 4
    - 3: 2 4
    - 4: 5
    - 5: 1 2



by Lada Adamic, U Michigan

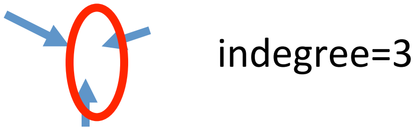
Degree: which node has the most edges?



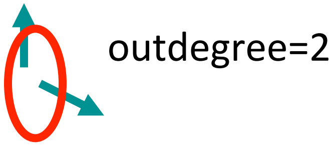
by Lada Adamic, U Michigan

Node degrees

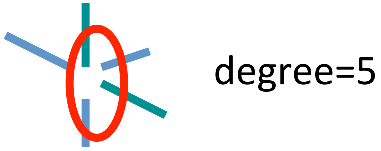
- Node network properties
  - from immediate connections
    - indegree
      how many directed edges (arcs) are incident on a node



    - outdegree
      how many directed edges (arcs) originate at a node

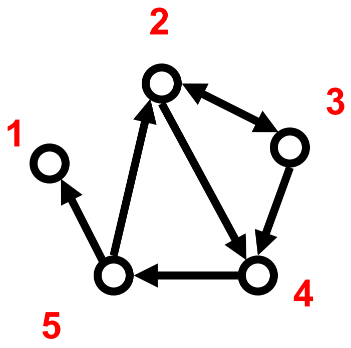


    - degree (in or out)
      number of edges incident on a node



  - from the entire graph
    - centrality (betweenness, closeness)

by Lada Adamic, U Michigan



Node degree from matrix values

- Outdegree = $\sum_{j=1}^{n} A_{ij}$

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

example: outdegree for node 3 is 2, which we obtain by summing the number of non-zero entries in the 3rd *row*

$$\sum_{j=1}^{n} A_{3j}$$

- Indegree = $\sum_{i=1}^{n} A_{ij}$

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

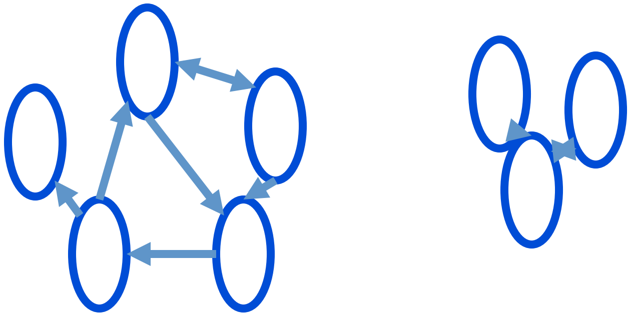
example: the indegree for node 3 is 1, which we obtain by summing the number of non-zero entries in the 3rd *column*

$$\sum_{i=1}^{n} A_{i3}$$

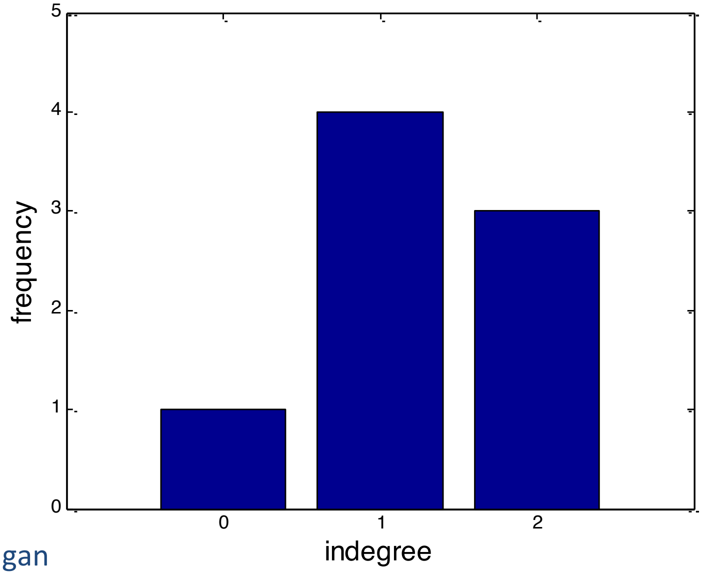
by Lada Adamic, U Michigan

Network metrics: degree sequence and degree distribution

- Degree sequence: An ordered list of the (in,out) degree of each node
  - In-degree sequence:
    - [2, 2, 2, 1, 1, 1, 1, 0]
  - Out-degree sequence:
    - [2, 2, 2, 2, 1, 1, 1, 0]
  - (undirected) degree sequence:
    - [3, 3, 3, 2, 2, 1, 1, 1]

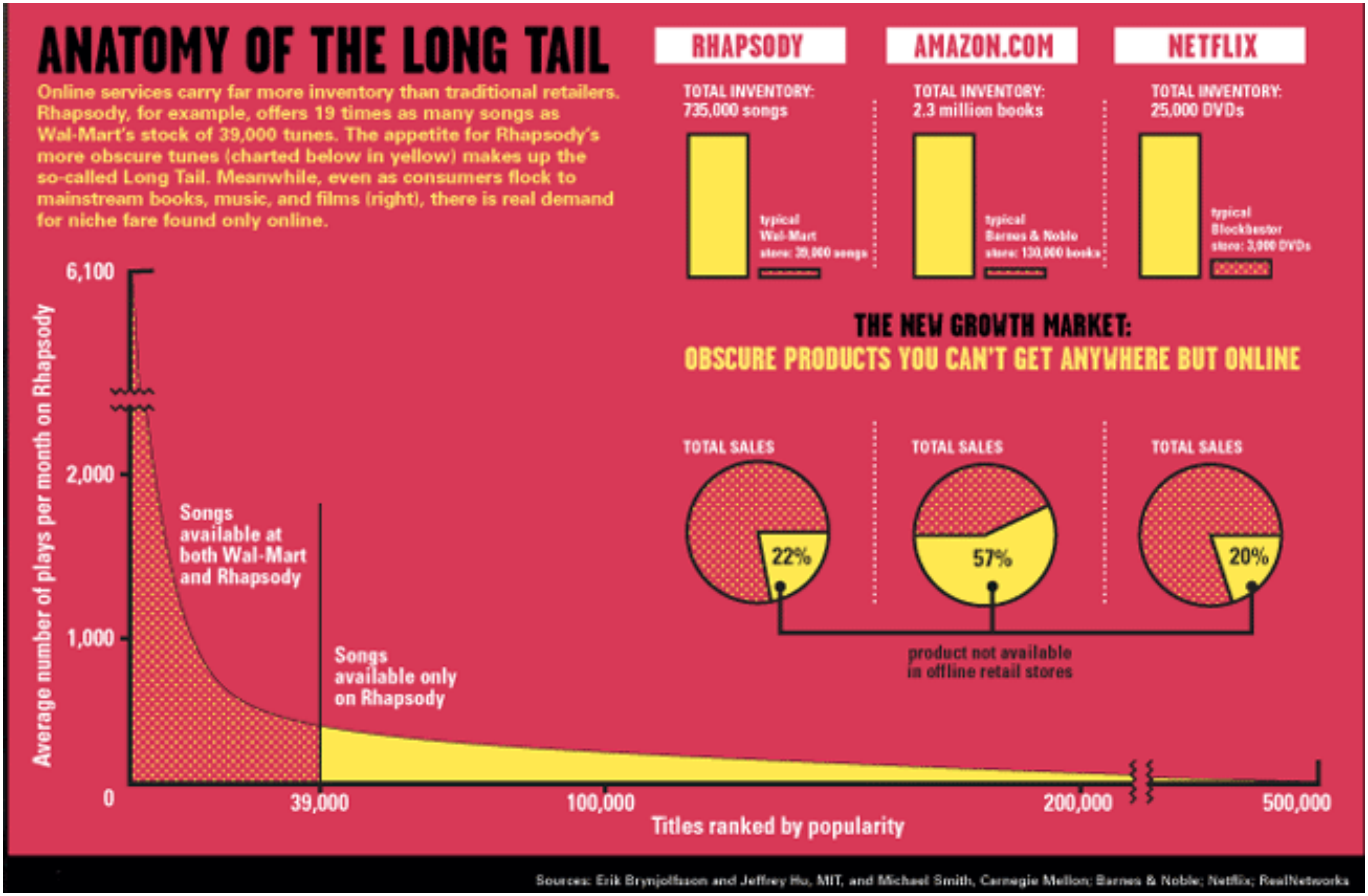


- Degree distribution: A frequency count of the occurrence of each degree
  - In-degree distribution:
    - [(2,3) (1,4) (0,1)]
  - Out-degree distribution:
    - [(2,4) (1,3) (0,1)]
  - (undirected) distribution:
    - [(3,3) (2,2) (1,3)]



by Lada Adamic, U Michigan

Long-Tail and power law

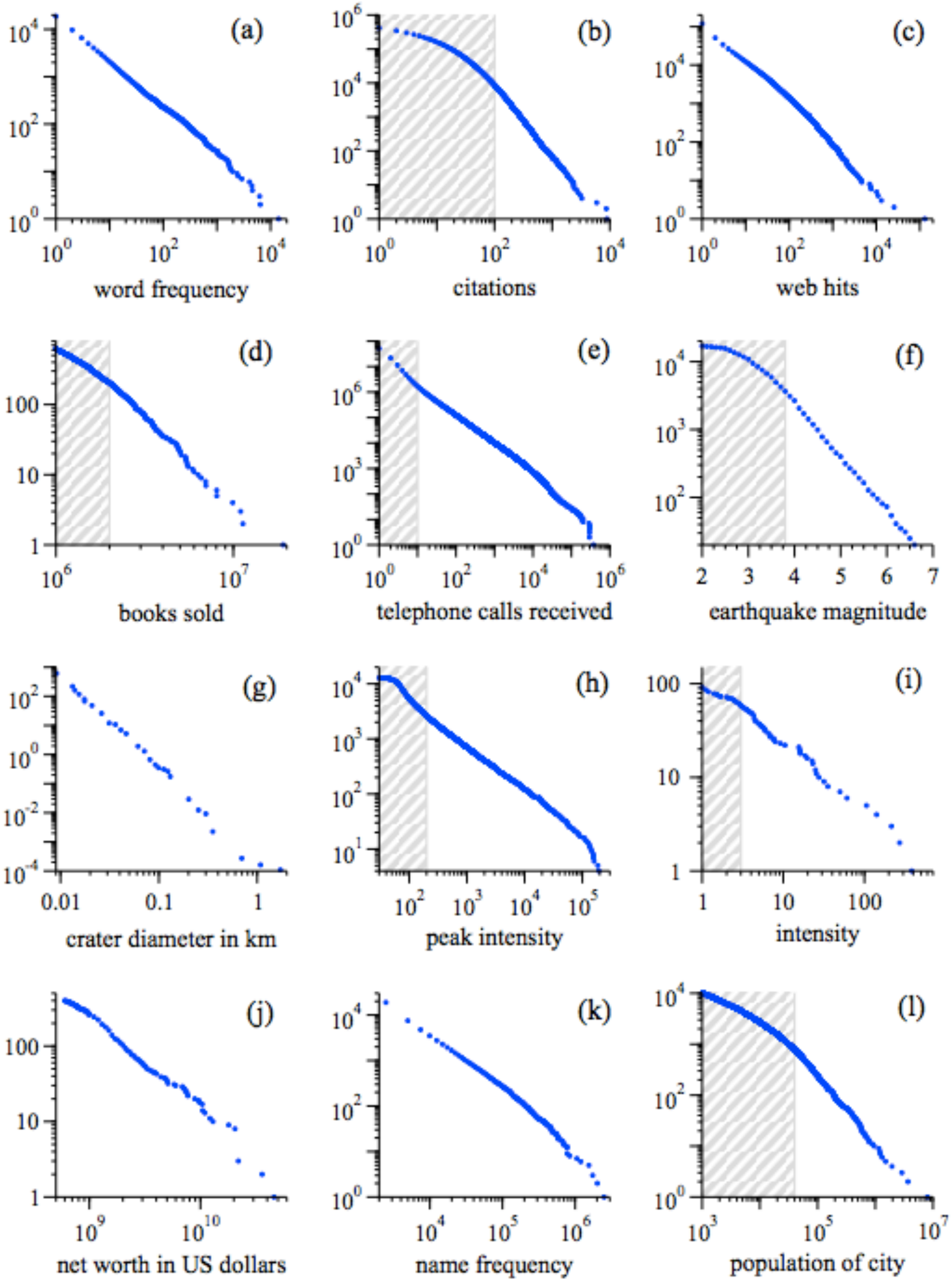


ANATOMY OF THE LONG TAIL

Online services carry far more inventory than traditional retailers. Rhapsody, for example, offers 19 times as many songs as Wal-Mart's stock of 39,000 tunes. The appetite for Rhapsody's more obscure tunes (charted below in yellow) makes up the so-called Long Tail. Meanwhile, even as consumers flock to mainstream books, music, and films (right), there is real demand for niche fare found only online.

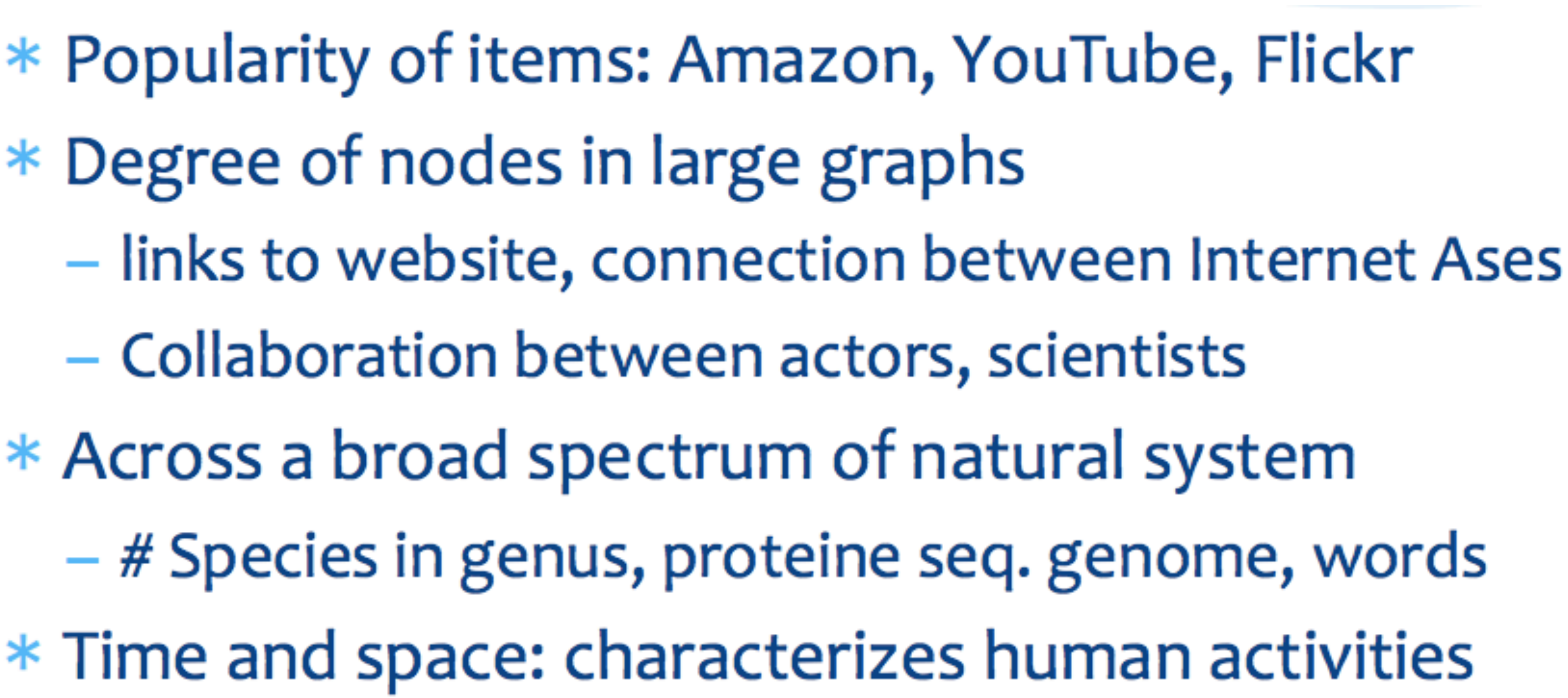
RHAPSODY
TOTAL INVENTORY: 735,000 songs
typical Wal-Mart store: 39,000 songs
AMAZON.COM
TOTAL INVENTORY: 2.3 million books
typical Barnes & Noble store: 130,000 books
NETFLIX
TOTAL INVENTORY: 25,000 DVDs
typical Blockbuster store: 3,000 DVDs
THE NEW GROWTH MARKET:
OBSCURE PRODUCTS YOU CAN'T GET ANYWHERE BUT ONLINE
TOTAL SALES
22%
TOTAL SALES
57%
TOTAL SALES
20%
product not available in offline retail stores
6,100
2,000
1,000
0
Average number of plays per month on Rhapsody
Songs available at both Wal-Mart and Rhapsody
Songs available only on Rhapsody
39,000
100,000
200,000
500,000
Titles ranked by popularity

Sources: Erik Brynjolfsson and Jeffrey Hu, MIT, and Michael Smith, Carnegie Mellon; Barnes & Noble; Netflix; RealNetworks



Power laws, Pareto distributions and Zipf’s law. M. Newman, Cont.Physics (2005)

Power Law Seems Ubiquitous



- Popularity of items: Amazon, YouTube, Flickr
- Degree of nodes in large graphs
  - links to website, connection between Internet Ases
  - Collaboration between actors, scientists
- Across a broad spectrum of natural system
  - # Species in genus, proteine seq. genome, words
- Time and space: characterizes human activities

How does power laws arise?

- A process for generating web links:

1) Nodes (i.e. webpages) join the graph in sequence, creating edges linking to previous nodes.
2) Each node u creates one outgoing edge as follows
   a) Choose a node v uniformly
   b) With prob. p, edge u->v is created,
   c) With prob. (1-p), edge u->w for w a (uniformly chose) children of v

   c) is equivalent to: "with probability 1−p, page u chooses a page w with probability proportional to w's current number of in-links, and creates a link to w"

Rich-Get-Richer Process

Initial condition

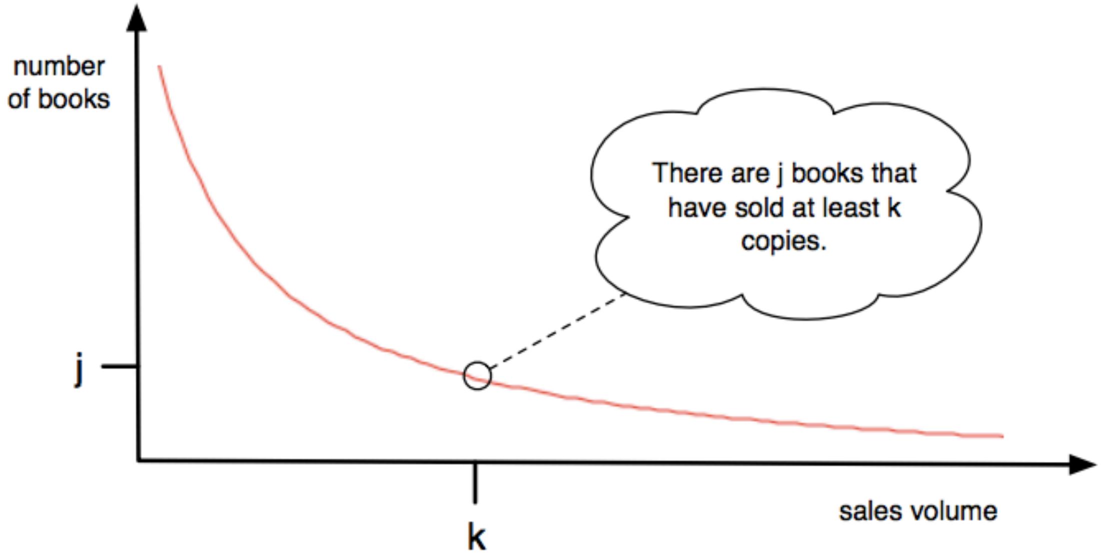
$$x_j(j) = 0$$

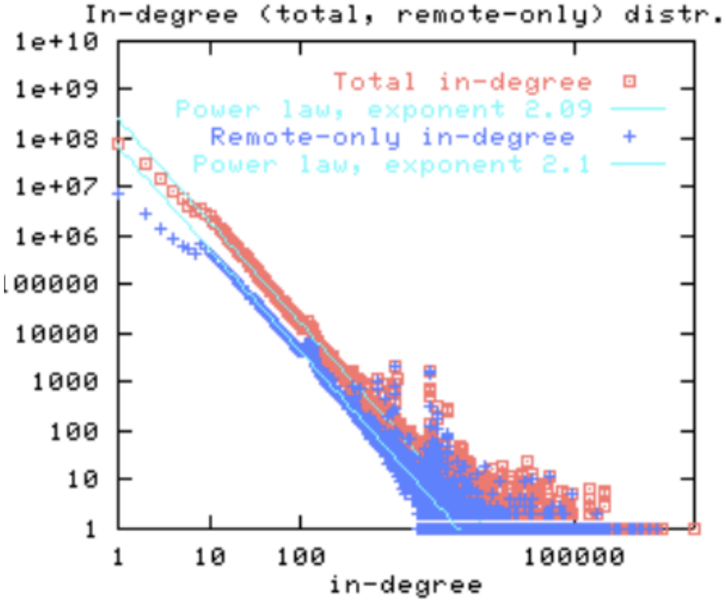
Rate of growth

$$\frac{dx_j}{dt} = \frac{p}{t} + \frac{(1-p)x_j}{t}.$$

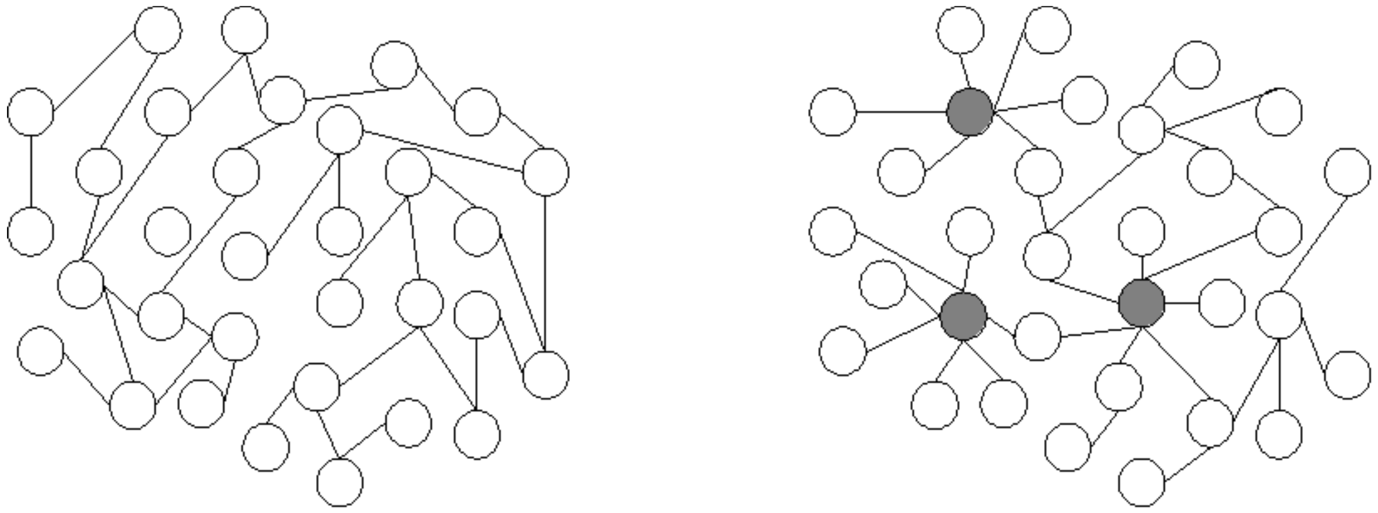
Fraction of nodes with k incoming links $\sim k^{-\alpha}$, $\alpha = 1 + 1/(1-p)$

[Easly and Kleinberg, Chapter 18] Fig 18.2 and 18.3





Random vs Power-law networks



(a) Random network

(b) Scale-free network

Terminologies that describe (different aspects of) the same process:

Power-law, Long-tail

Scale-free

Rich-get-richer, preferential attachment

Networks from Document Collections

Dataset	Nodes	Edges	Edge weight?	Directed?	Potential use of the network
“Divided they Blog”	blog				
Email	people				
Research Papers	authors				
Research Papers	papers				
Research Papers	Journals/ conferences				
Recipe					
…					

Summary for Today

- What are networks of documents
- A collection of network analysis problems
- Network representation
- Power-law degree distribution and rich-get-richer processes

- HW: representing Twitter as network
- Coming up in Wed: network description

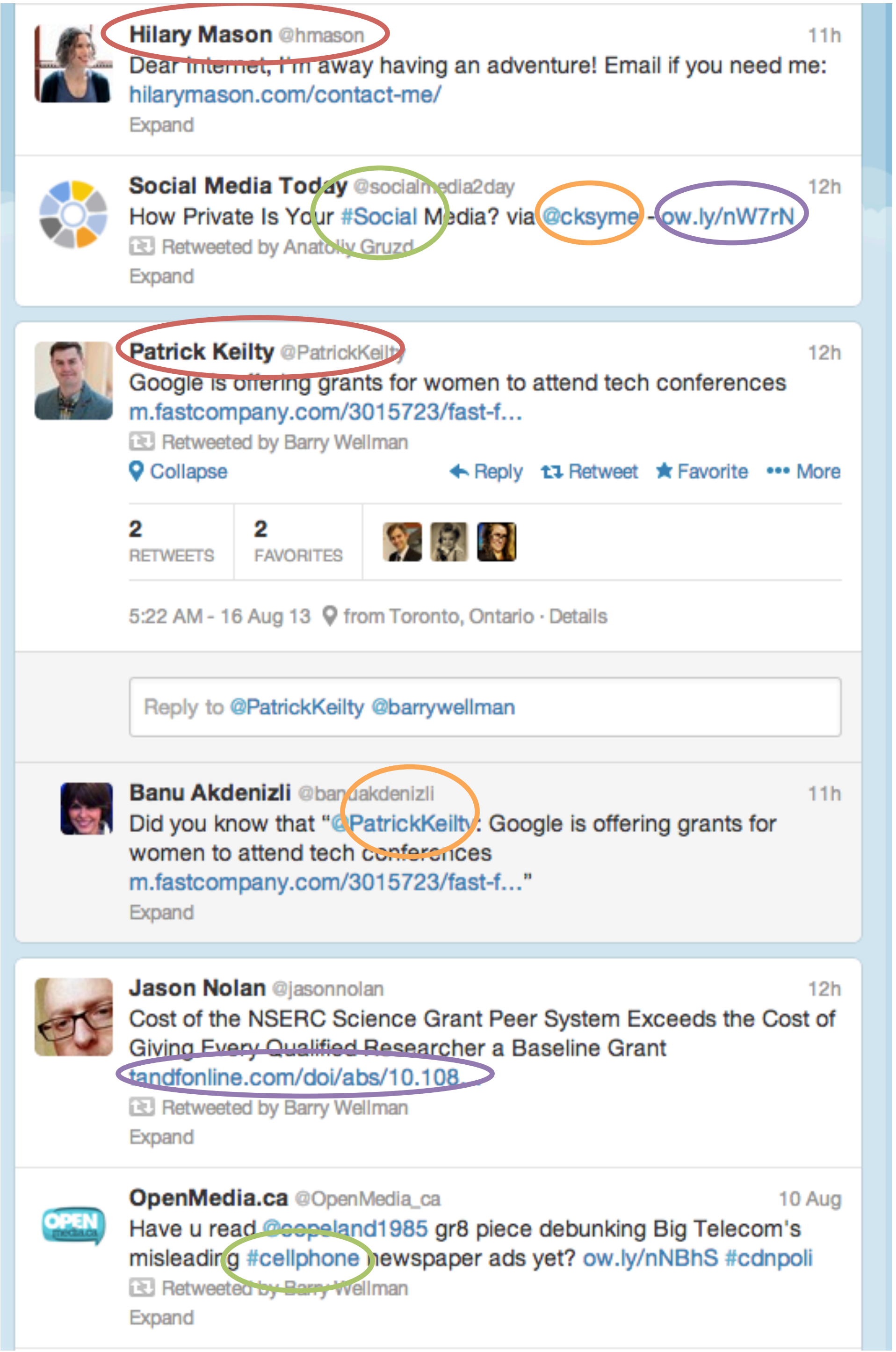
Document Elements In Twitter

User

Mention

Hashtag

Hyperlink



Hilary Mason @hmason 11h
Dear Internet, I'm away having an adventure! Email if you need me:
hilarymason.com/contact-me/
Expand

Social Media Today @socialmedia2day 12h
How Private Is Your #Social Media? via @cksyme - ow.ly/nW7rN
Retweeted by Anatoliy Gruzd
Expand

Patrick Keilty @PatrickKeilty 12h
Google is offering grants for women to attend tech conferences
m.fastcompany.com/3015723/fast-f...
Retweeted by Barry Wellman
Collapse   Reply Retweet Favorite ••• More

2 RETWEETS 2 FAVORITES

5:22 AM - 16 Aug 13 from Toronto, Ontario · Details

Reply to @PatrickKeilty @barrywellman

Banu Akdenizli @banuakdenizli 11h
Did you know that "@PatrickKeilty: Google is offering grants for women to attend tech conferences
m.fastcompany.com/3015723/fast-f..."
Expand

Jason Nolan @jasonnolan 12h
Cost of the NSERC Science Grant Peer System Exceeds the Cost of Giving Every Qualified Researcher a Baseline Grant
tandfonline.com/doi/abs/10.108...
Retweeted by Barry Wellman
Expand

OpenMedia.ca @OpenMedia_ca 10 Aug
Have u read @copeland1985 gr8 piece debunking Big Telecom's misleading #cellphone newspaper ads yet? ow.ly/nNBhS #cdnpoli
Retweeted by Barry Wellman
Expand

Networks from Twitter

Nodes	Edges	Edge weight?	Directed?	Potential use of the network
User				
User				
Hashtag				
Website				
Website, User				
User				