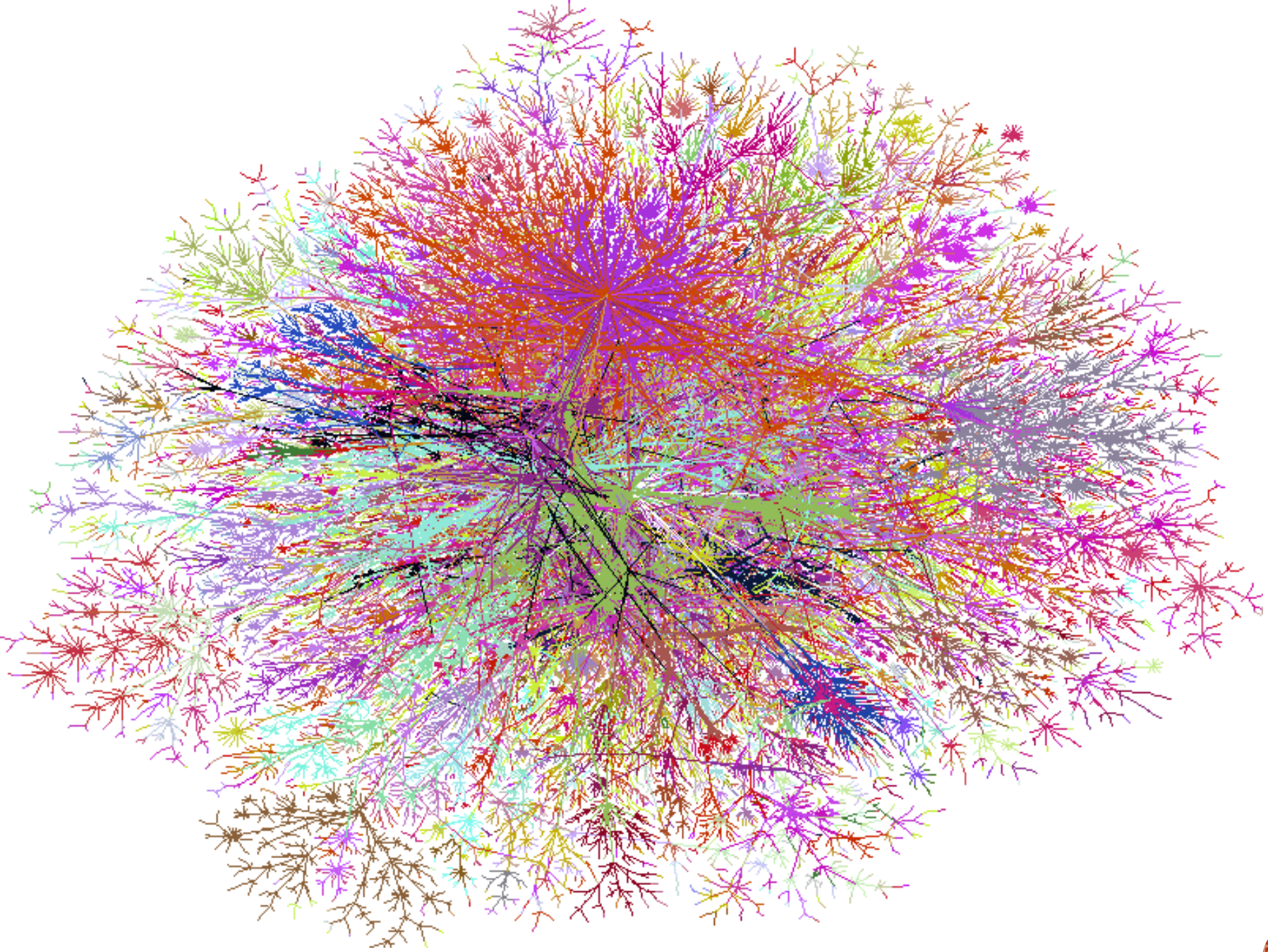
A brief introduction to

Social Network Analysis

and Visualization

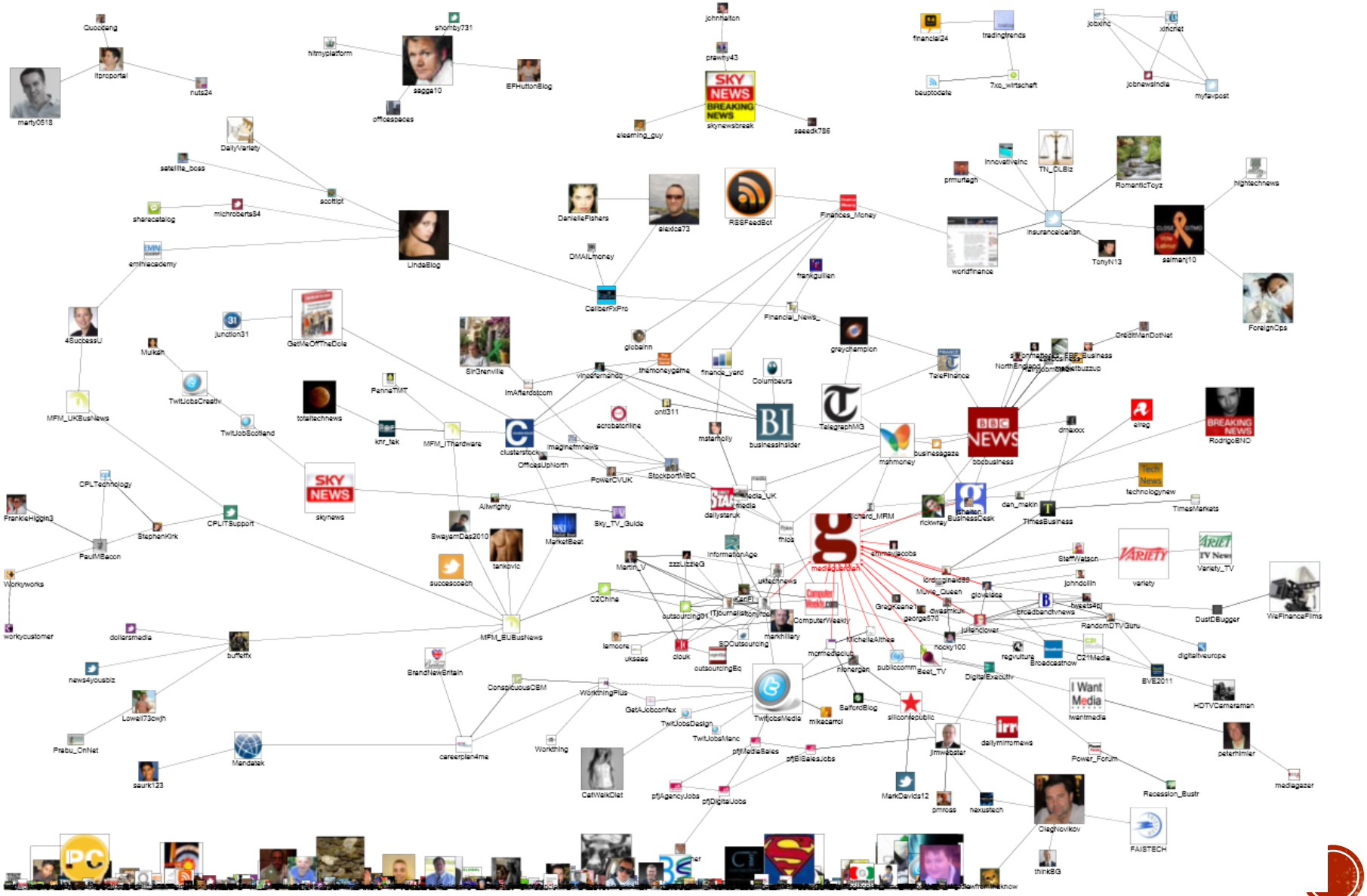
Emilio Ferrara @jabawack
Computer Scientist - USC Information Sciences Institute
Research Assistant Professor – USC Dept. Computer Science
emiliofe@usc.edu
www.emilio.ferrara.name

IPAM Culture Analytics
Spring 2016



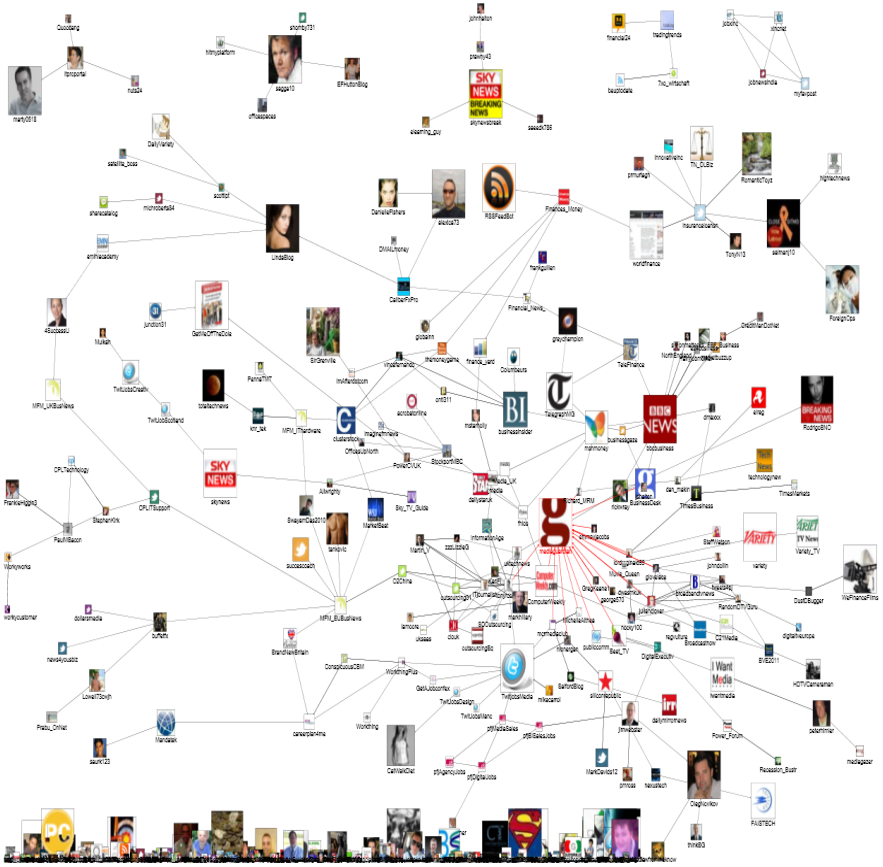
https://www.flickr.com/photos/jurvetson/916142

NETWORK DATA



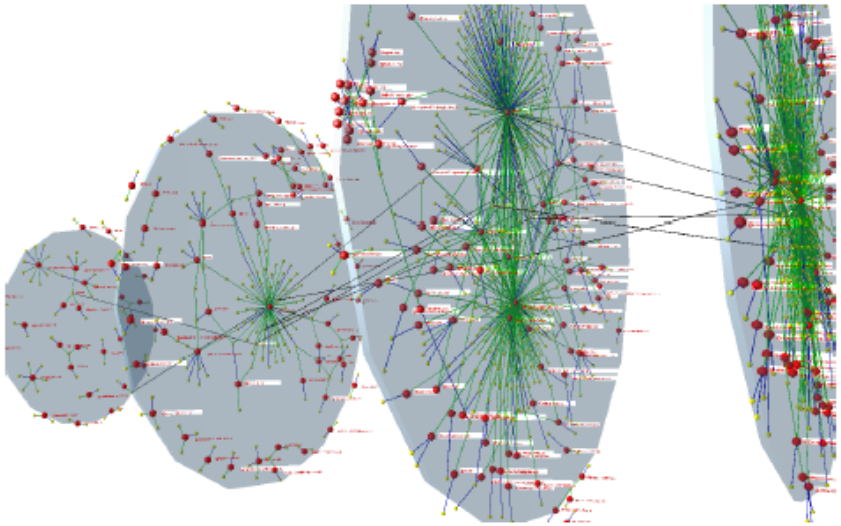
https://www.flickr.com/photos/marc_smith/4512483812

REPRESENTING NETWORK STRUCTURE



- **Graphs** of nodes connected by links
- **Nodes**: entities of interest (ie, a person, a protein, a Web page,…
- **Links**: relation between two nodes (ie, a communication, a common interest, etc.)

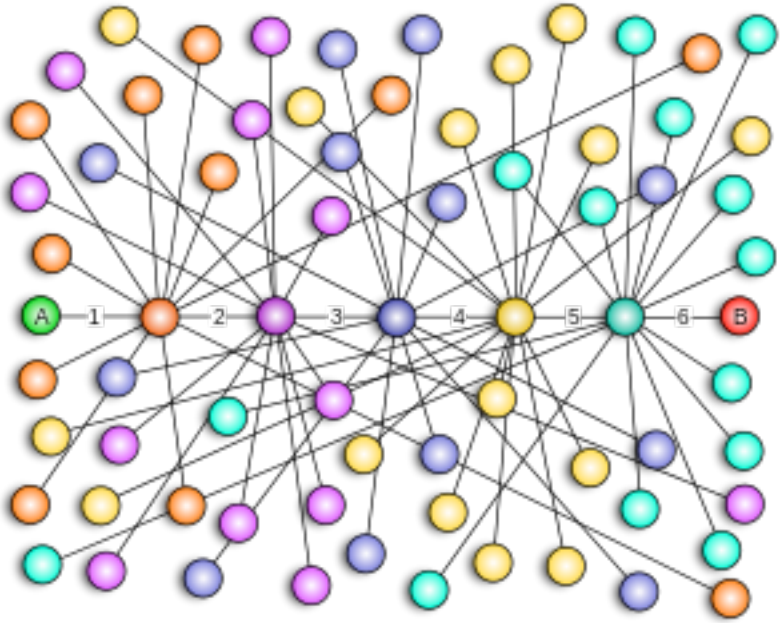
REPRESENTING NETWORK BEHAVIOR:
DYNAMIC NETWORKS



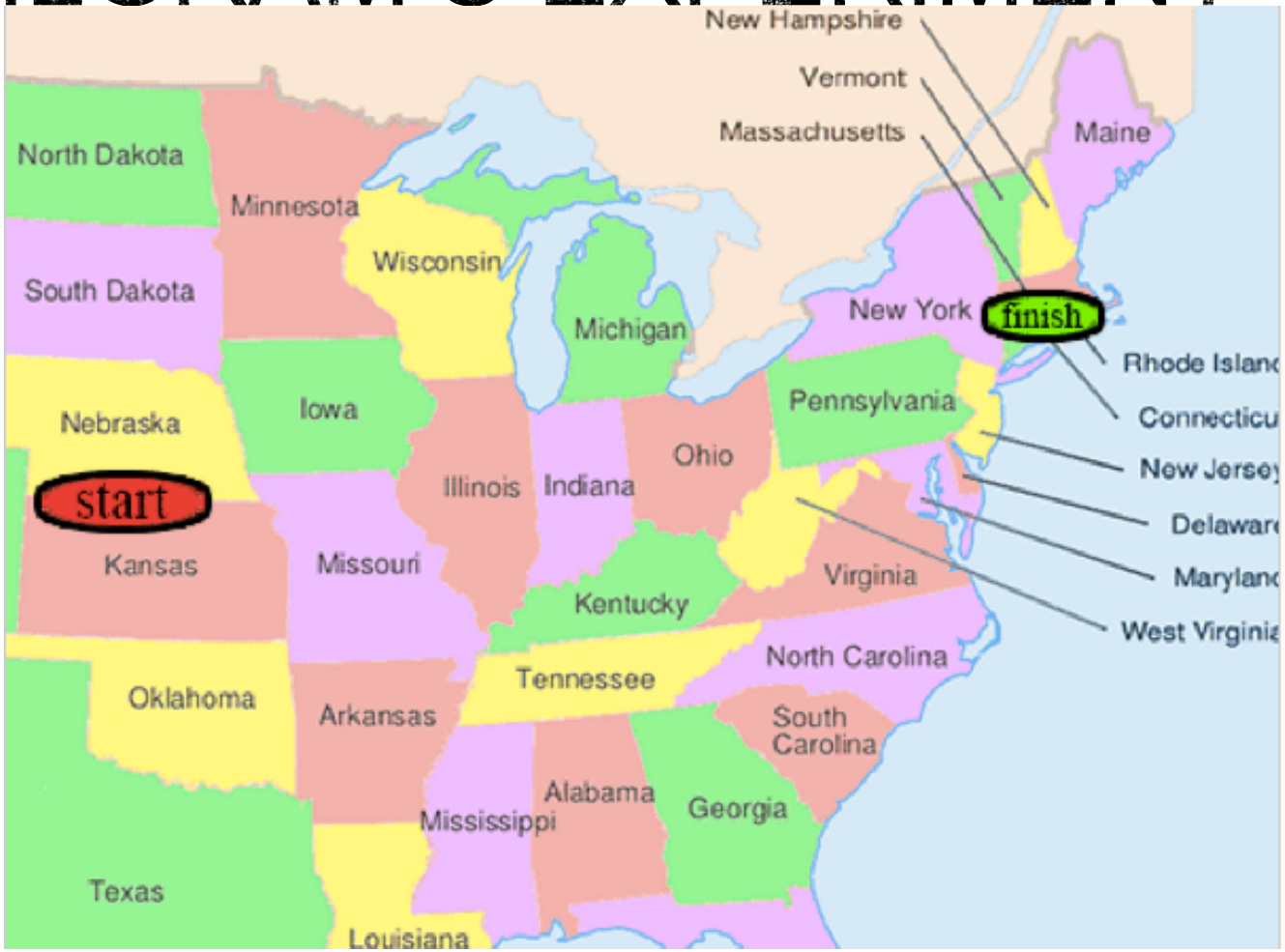
- Behaviors of the entities over time
- Network structure may also change over time
- Weak vs strong links
- Changes almost always have trickle effects on the rest of the network

INTERESTING QUESTIONS

- If you select two random people: what is the probability that they know one another?
- Can you connect two random people through their network of acquaintances?

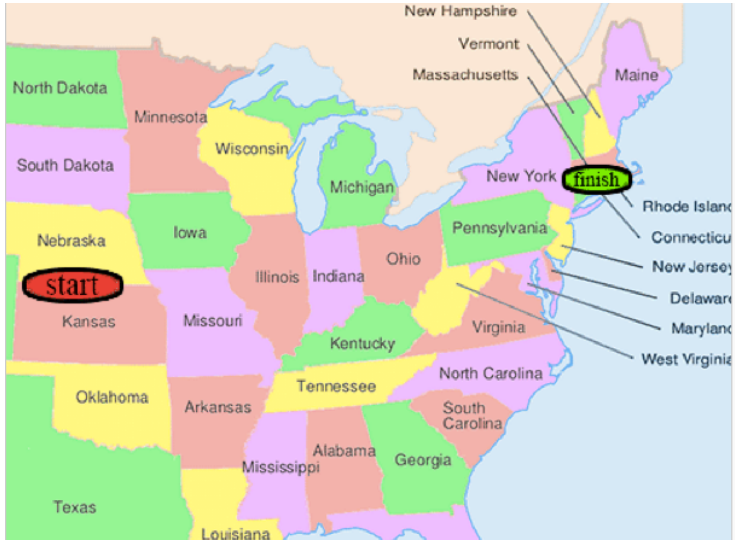


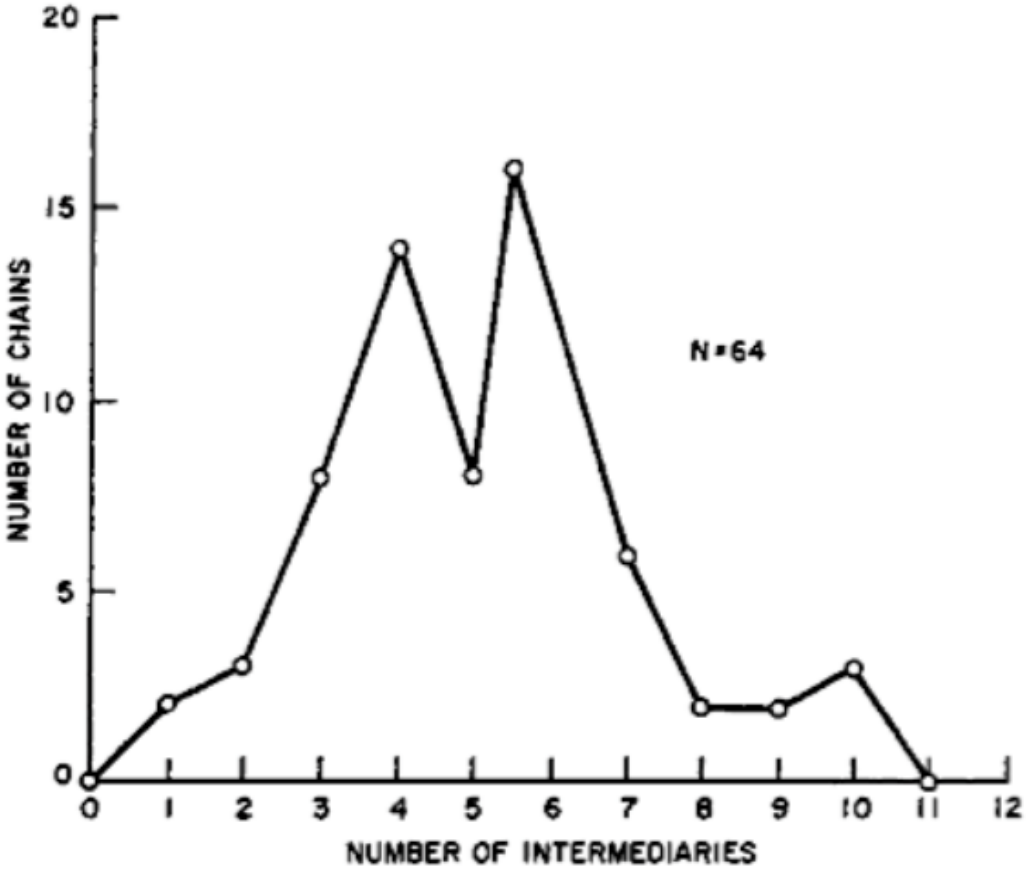
SMALL-WORLD NETWORKS: MILGRAM'S EXPERIMENT



https://en.wikipedia.org/wiki/Small-world_experiment#/media/File:Experement_Small_World_%28possible_option%29.gif

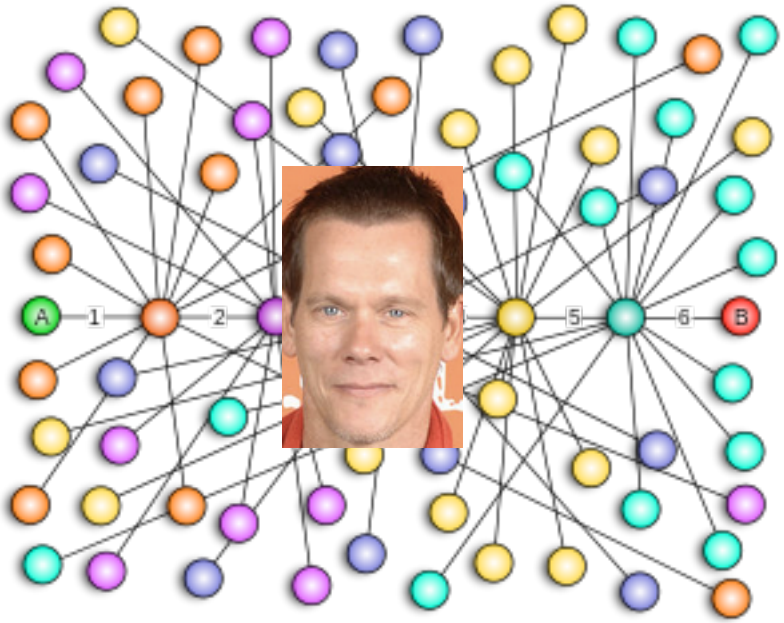
SMALL-WORLD NETWORKS: MILGRAM'S EXPERIMENT





https://en.wikipedia.org/wiki/Small-world_experiment#/media/File:Experement_Small_World_%28possible_option%29.gif

“SIX DEGREES OF SEPARATION”



"I know a guy who knows a guy who knows a guy who knows a guy who knows Kevin Bacon.”

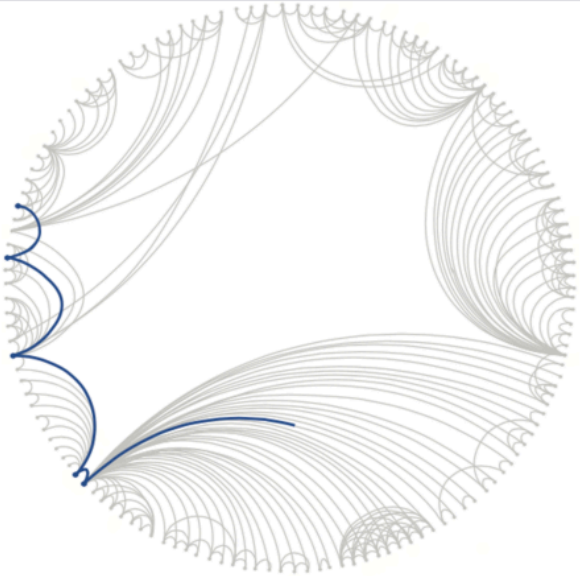
Bacon number = 6

https://en.wikipedia.org/wiki/Six_Degrees_of_Kevin_Bacon#/media/File:Kevin_Bacon.jpg
https://en.wikipedia.org/wiki/Small-world_experiment#/media/File:Six_degrees_of_separation.svg

FACEBOOK 3.5 DEGREES OF SEPARATION

Three and a half degrees of separation

Blog



Sergey Edunov, Carlos Diuk, Ismail Onur Filiz, Smriti Bhagat, Moira Burke

4 febbraio

Core Data Science

"I read somewhere that everybody on this planet is separated by only six other people. Six degrees of separation. Between us and everybody else on this planet. The president of the United States. A gondolier in Venice. Fill in the names. . . . How every person is a new door, opening up into other worlds. Six degrees of separation between me and everyone else on this planet. But to find the right six people . . ." - John Guare, **Six Degrees of Separation** (1990)

How connected is the world? Playwrights [1], poets [2], and scientists [3] have proposed that everyone on the planet is connected to everyone else by six other people. In honor of Friends Day, we've crunched the Facebook friend graph and determined that the number is 3.57. **Each person in the world (at least among the 1.59 billion people active on Facebook) is connected to every other person by an average of three and a half other people.** The average distance we observe is 4.57, corresponding to 3.57 intermediaries or "degrees of separation." Within the US, people are connected to each other by an average of 3.46 degrees.

Our collective "degrees of separation" have shrunk over the past five years. In 2011, researchers at Cornell, the Università degli Studi di Milano, and Facebook computed the average across the 721 million people using the site then, and found that it was 3.74 [4,5]. Now, with twice as many people using the site, we've grown more interconnected, thus shortening the distance between any two people in the world.

Calculating this number across billions of people and hundreds of billions of friendship connections is challenging; we use statistical techniques described below to precisely estimate distance based on de-identified, aggregate data.

My degrees of separation

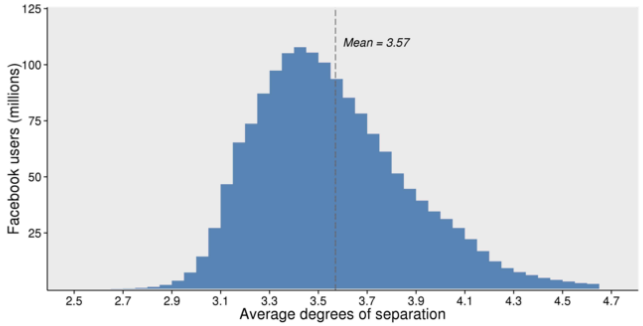
Emilio Ferrara's average degrees of separation from everyone is **3.17**.

Some Facebook employees

Mark Zuckerberg
3.17 degrees of separation

Sheryl Sandberg
2.92 degrees of separation

The majority of the people on Facebook have averages between 2.9 and 4.2 degrees of separation. Figure 1 (below) shows the distribution of averages for each person.

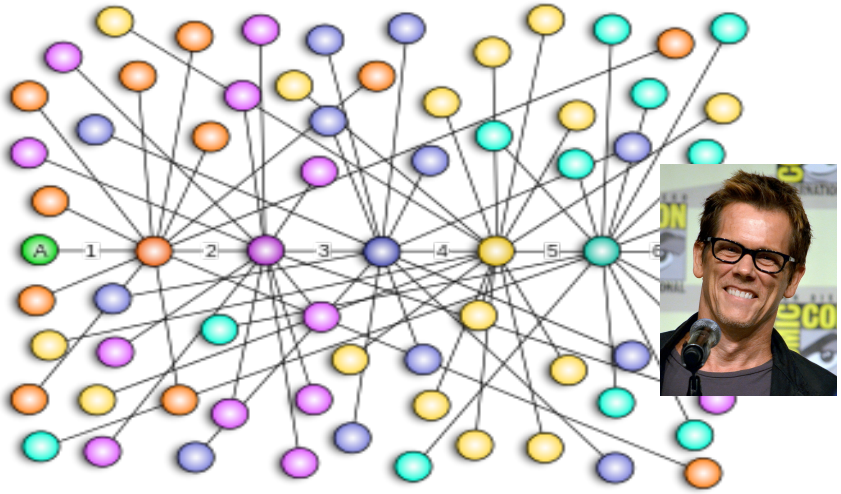


J. Ugander, B. Karrer, L. Backstrom, C. Marlow. The Anatomy of the Facebook Social Graph, http://arxiv.org/abs/1111.4503
L. Backstrom, P. Boldi, M. Rosa, J. Ugander, S. Vigna. Four Degrees of Separation, http://arxiv.org/abs/1111.4570
https://research.facebook.com/blog/three-and-a-half-degrees-of-separation/

ORACLEOFBACON.ORG

Kevin Bacon Number	# of People
0	1
1	3105
2	366567
3	1313581
4	332813
5	27993
6	3274
7	453
8	52
9	6
10	1

Total number of linkable actors: 2047846
Weighted total of linkable actors: 6171494
Average Kevin Bacon number: 3.014



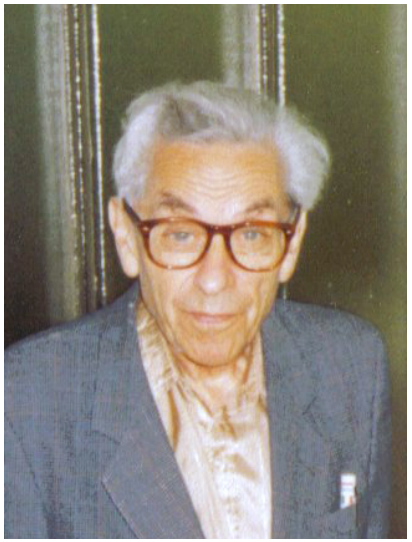
George Lucas (I) has a Bacon number of 2.

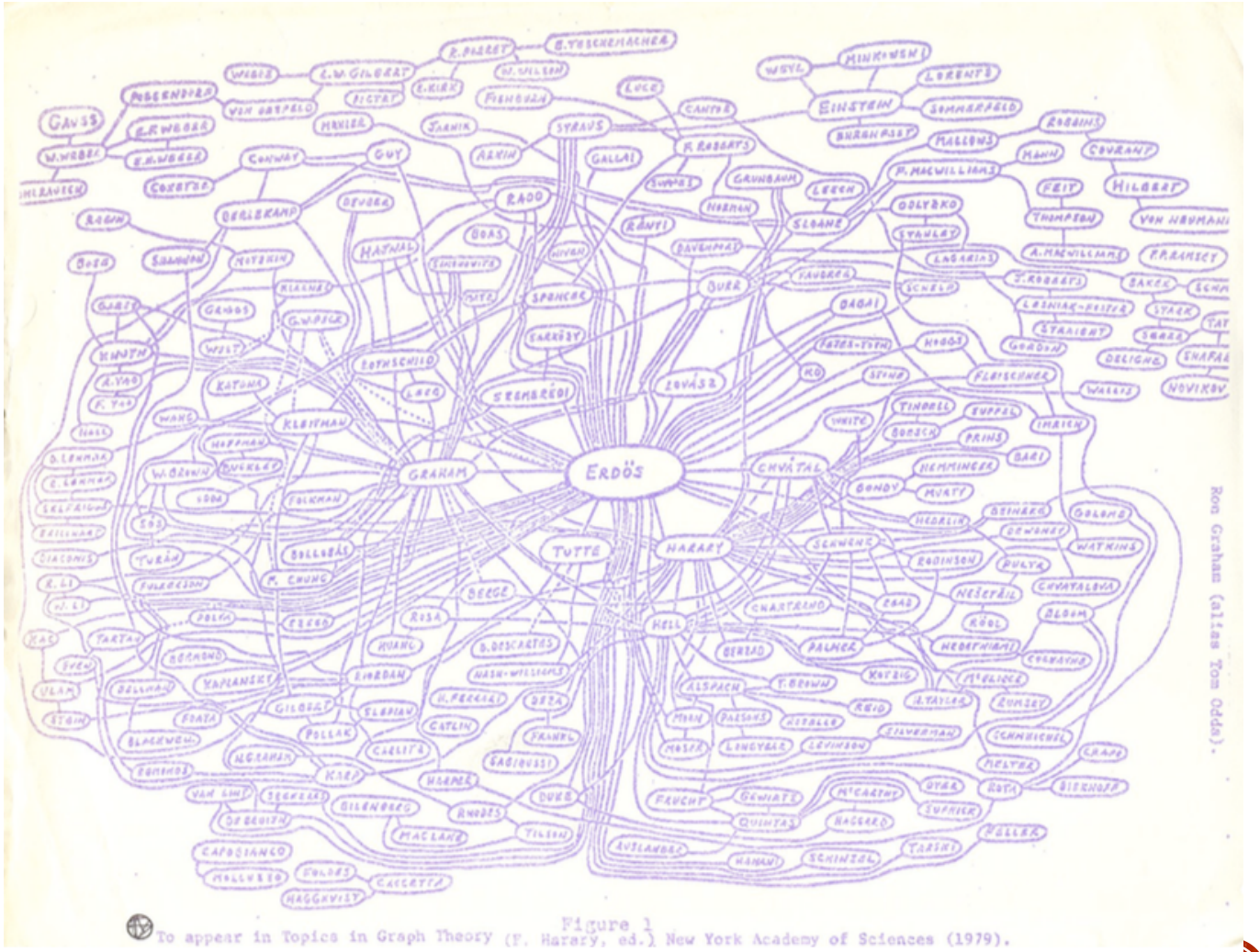
George Lucas (I)
was in
Beverly Hills Cop III (1994)
with
Meadow Williams
was in
Apollo 13 (1995)
with
Kevin Bacon

George Lucas (I)
was in
Hook (1991)
with
Julia Roberts (I)
was in
Flatliners (1990)
with
Kevin Bacon

http://upload.wikimedia.org/wikipedia/commons/7/78/Kevin_Bacon_Comic-Con_2012.jpg

THE ERDOS NUMBER





http://wwwp.oakland.edu/enp/thedata/
https://commons.wikimedia.org/wiki/File:Erdos_budapest_fall_1992_(cropped).jpg

SOURCES OF NETWORKED DATA

- Messages across people (e.g., emails, blogs)
- Social network sites (e.g., Facebook)
- Social media (e.g., twitter)
- Constructing networks from other data

http://www.sas.com/content/sascom/en_us/industry/banking/fraud-network-analysis/_jcr_content/par/styledcontainer_9a94/par/image_f35c.img.png/1431023853128.png

NETWORKS CREATED FROM DIFFERENT KINDS OF DATA

TouchGraph's visualization of senator co-sponsorship patterns in the 110th congress shows that he has acting like a democrat for years.

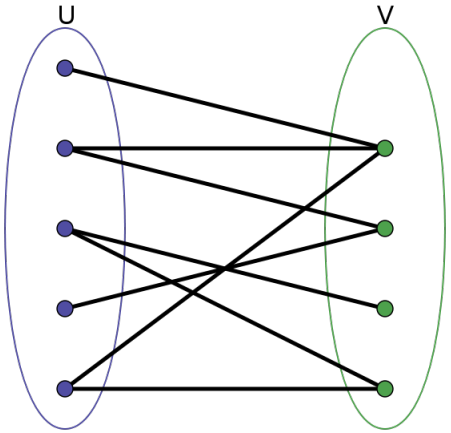
http://www.touchgraph.com/news

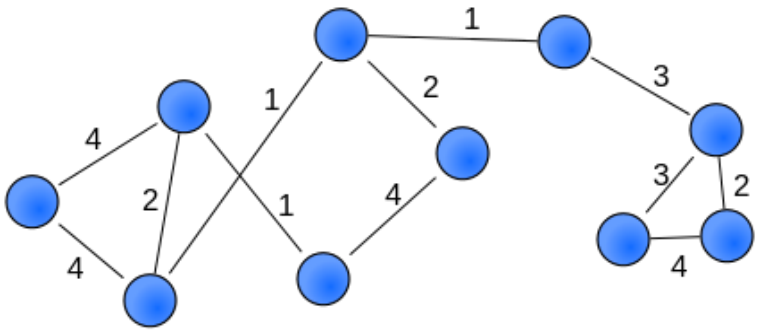
US Senators that share an Alma Mater

EXAMPLES OF NETWORKS

NETWORK	NODES	LINKS
Cellular metabolism	Molecules involved in burning food for energy	Participation in the same biochemical reaction
Hollywood	Actors	Appearance in the same movie
Internet	Routers	Optical and other physical connections
Protein regulatory network	Proteins that help to regulate a cell's activities	Interactions among proteins
Research collaborations	Scientists	Co-authorship of papers
Sexual relationships	People	Sexual contact
World Wide Web	Web pages	URLs

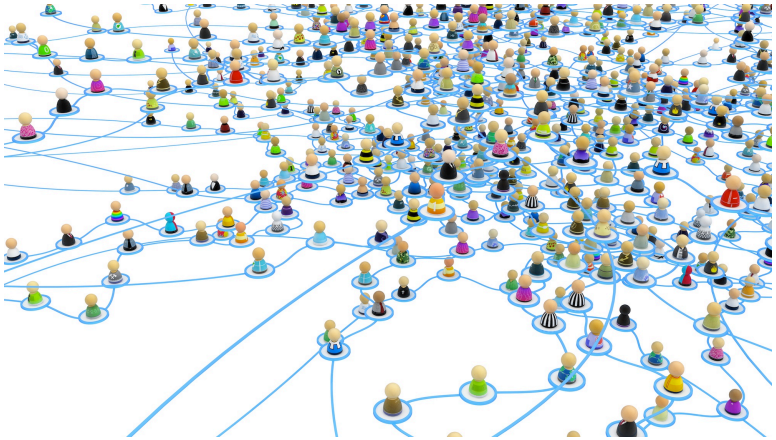
TYPES OF NETWORKS





- Homogeneous networks
- < Bipartite networks
- Heterogeneous networks
- < Weighted networks

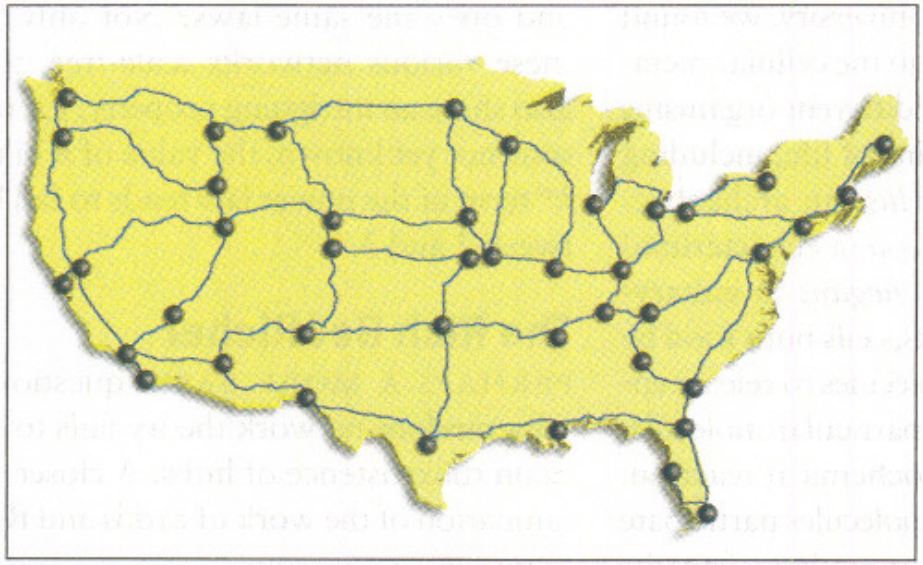
INTERESTING TYPES OF NETWORKS



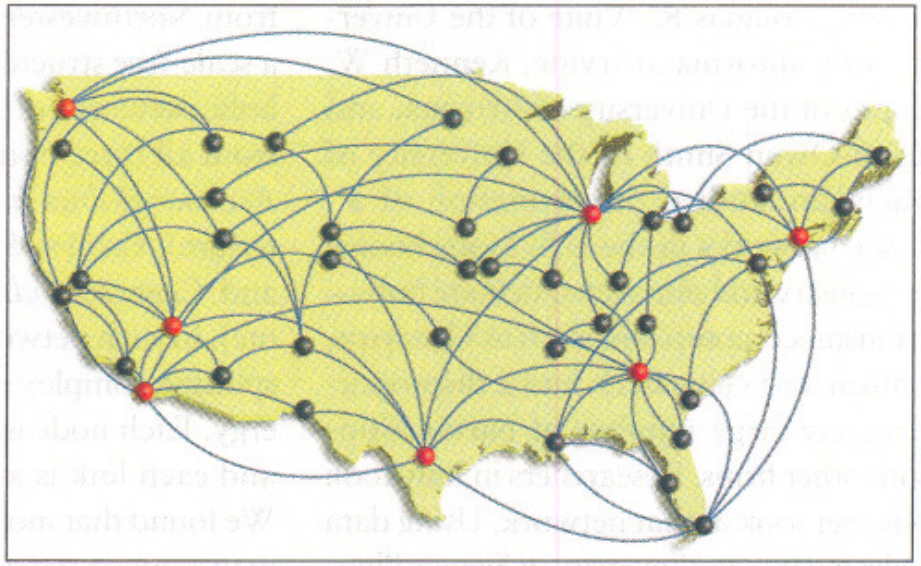
- Social networks
  - People as nodes
- Scale-free networks
  - Each node has few (and different) connections
  - A few nodes may have many connections

RANDOM VS SCALE-FREE NETWORKS

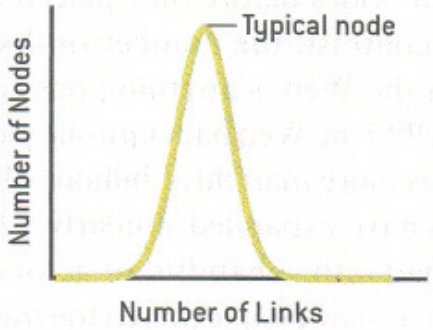
Random Network



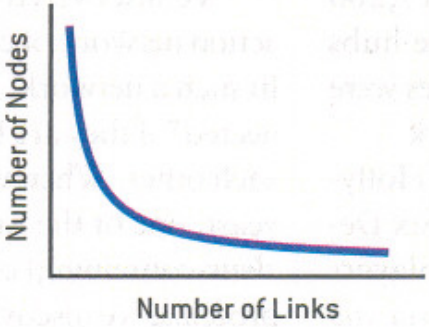
Scale-Free Network

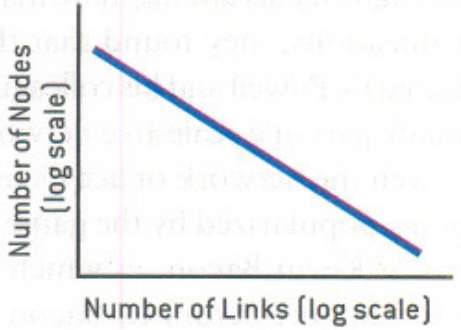


Bell Curve Distribution of Node Linkages

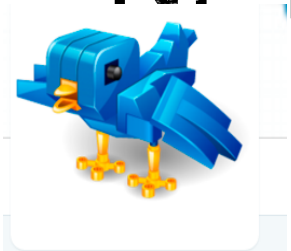


Power Law Distribution of Node Linkages





SCALE-FREE SOCIAL NETWORKS



Truthy BotOrNot
@TruthyBotOrNot FOLLOWS YOU

Online classification tool for Twitter users. Bot or Not?

Bloomington, IN
truthy.indiana.edu/botornot/

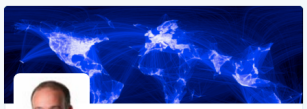
TWEETS	FOLLOWING	FOLLOWERS
155	34	288

IU SOIC
@iusoic FOLLOWS YOU

Official news, events, and announcements from the IU School of Informatics and Computing, Bloomington.

Bloomington, Indiana
soic.indiana.edu

TWEETS	FOLLOWING	FOLLOWERS
6,256	544	2,688



Emilio Ferrara
@jabawack

TWEETS	FOLLOWING	FOLLOWERS
6,995	869	30.6K

x10

Stanford University
@Stanford

This year, Stanford celebrates 125 years of impact in people's lives and around the world. Visit the Stanford 125 website for information about upcoming events.

Stanford, CA

TWEETS	FOLLOWING	FOLLOWERS
13.5K	502	383K

LeBron James
@LeBronJames

#StriveForGreatness !!
#TeamLeBron parody. ★ Advertising email ★ Hero4774@Gmail.com

Worldwide
vine.co/KingLeBron

TWEETS	FOLLOWING	FOLLOWERS
1,614	930K	3.64M

jimmy fallon
@jimmyfallon

astrophysicist

New York, New York
tonightshow.com

TWEETS	FOLLOWING	FOLLOWERS
8,862	6,793	36.6M

YES, EVEN ACADEMICS!

x10

Qing Ke
Indiana University
Email verificata su indiana.edu
Citato da 23
Complex Systems Network Science Economics of Science Statistical Learning

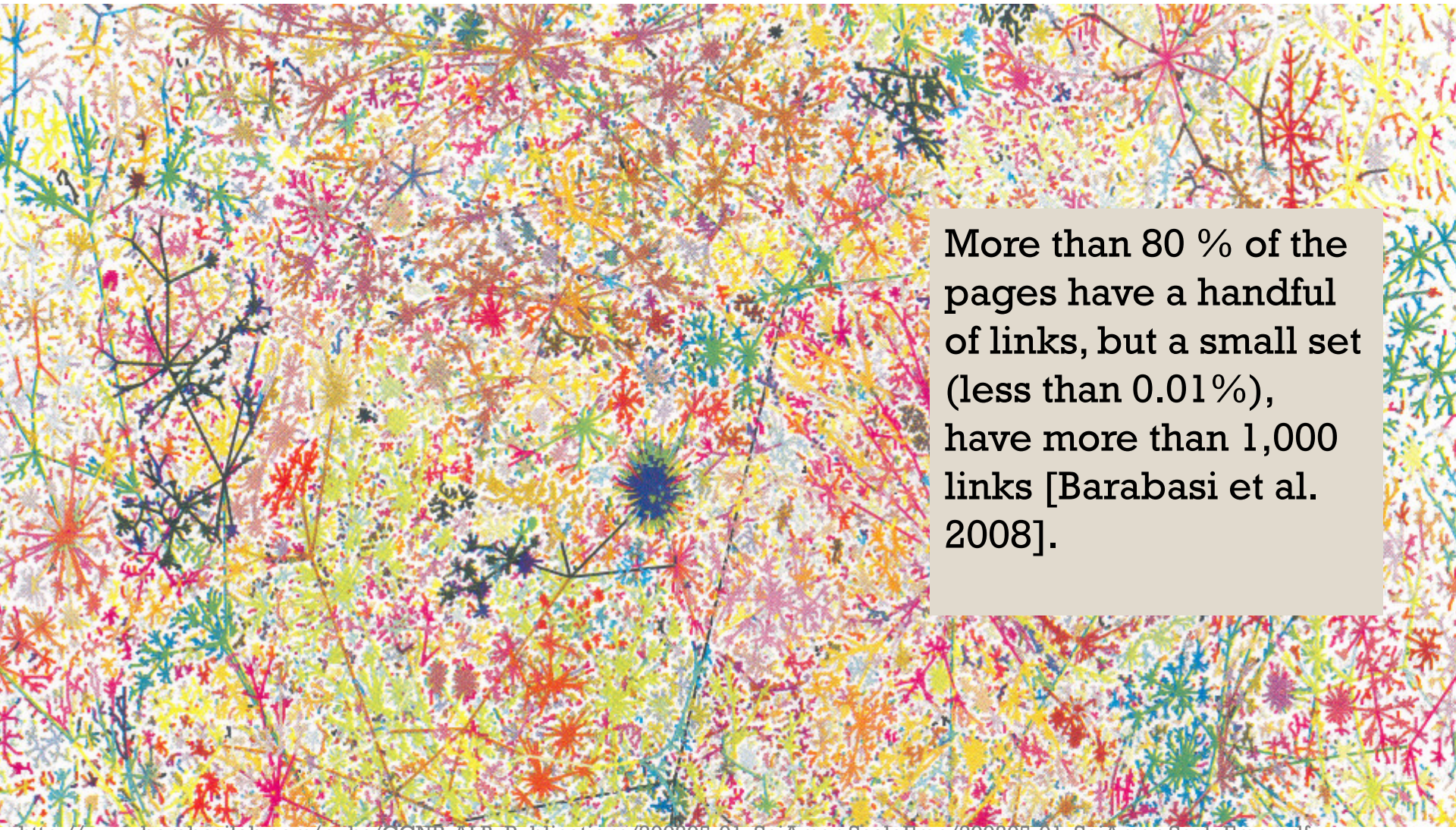
Onur Varol
Indiana University
Email verificata su indiana.edu
Citato da 112

Emilio Ferrara
University of Southern California - Information Sciences Institute
Email verificata su isi.edu
Citato da 1108
Data Science Network Science Machine Learning Data Mining Computational Social Science

Aaron Clauset
Assistant Professor of Computer Science, University of Colorado Boulder
Email verificata su colorado.edu
Citato da 11024
Network Science Complex Networks Machine Learning Computational Biology
Social Science

Albert-László Barabási
Northeastern University, Harvard Medical School
Email verificata su neu.edu
Citato da 141841
network science statistical physics biological physics physics medicine

SCALE-FREE NETWORKS: THE INTERNET AND THE WEB



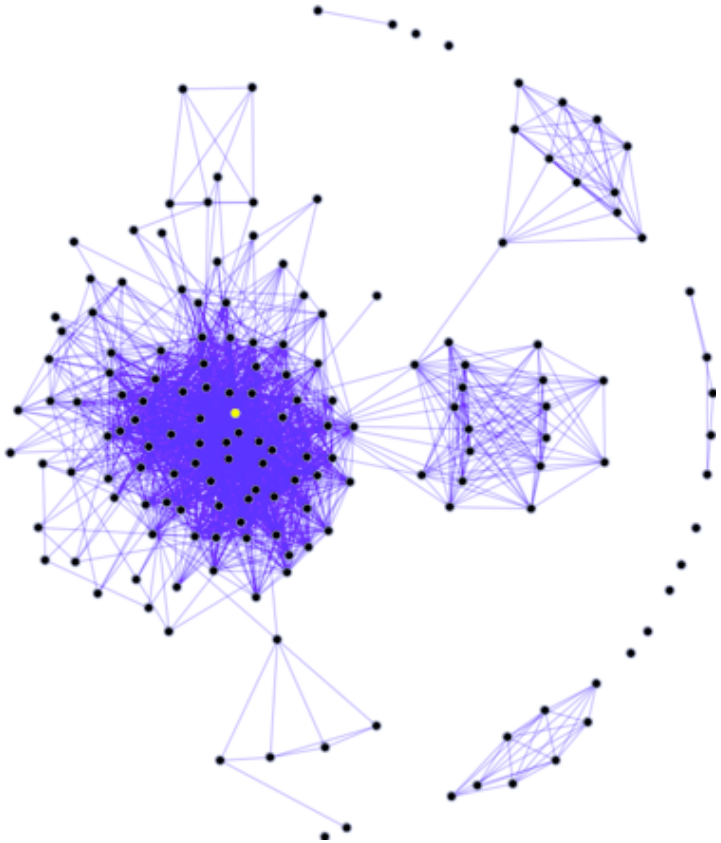
More than 80 % of the pages have a handful of links, but a small set (less than 0.01%), have more than 1,000 links [Barabasi et al. 2008].

http://www.barabasilab.com/pubs/CCNR-ALB_Publications/200305-01_SciAmer-ScaleFree/200305-01_SciAmer-ScaleFree.pdf

THESE ARE ALL SCALE-FREE NETWORKS!

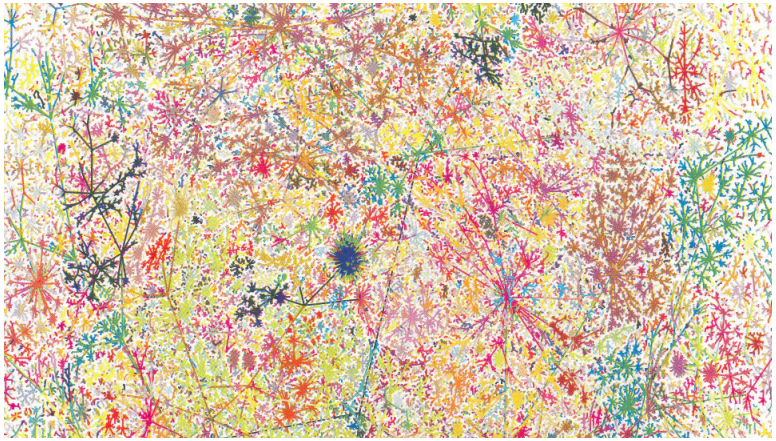
NETWORK	NODES	LINKS
Cellular metabolism	Molecules involved in burning food for energy	Participation in the same biochemical reaction
Hollywood	Actors	Appearance in the same movie
Internet	Routers	Optical and other physical connections
Protein regulatory network	Proteins that help to regulate a cell's activities	Interactions among proteins
Research collaborations	Scientists	Co-authorship of papers
Sexual relationships	People	Sexual contact
World Wide Web	Web pages	URLs

SOCIAL NETWORK DATA



- Major issue: networks in social sites are often not publicly accessible
  - E.g., Facebook
- Some exceptions:
  - 'Hand-crafted' datasets
  - Twitter (et similia)

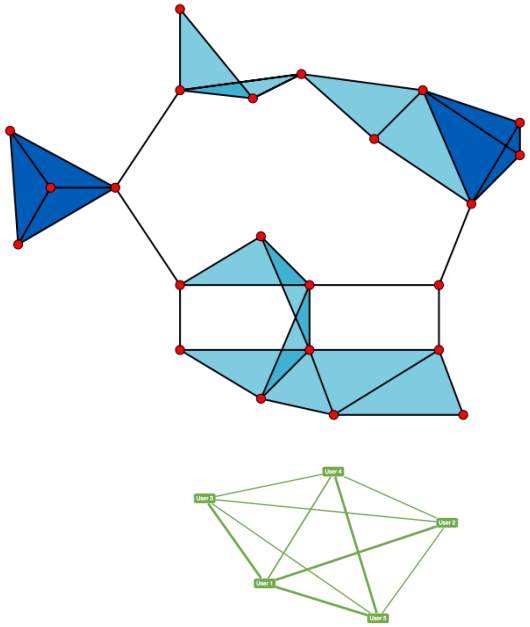
NETWORK ANALYSIS

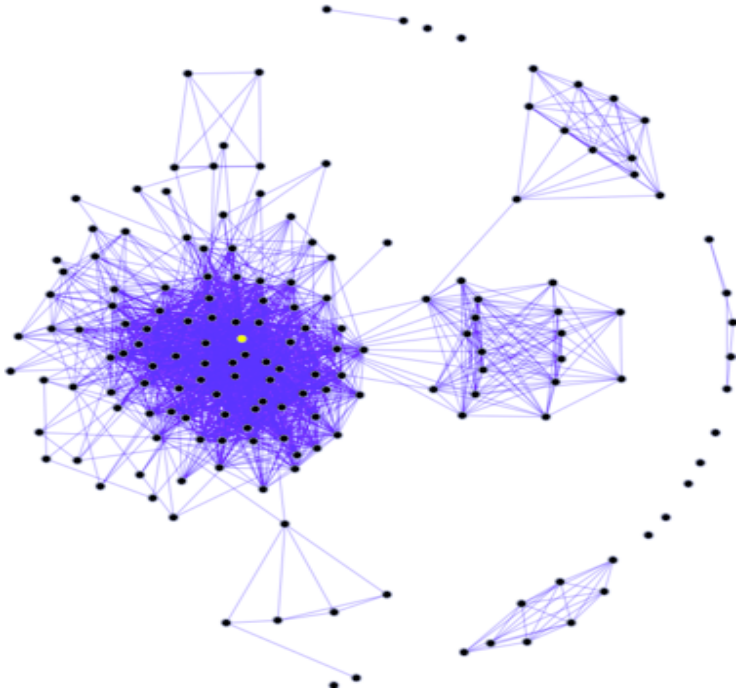


- Cliques
- Distance
- Bridges
- Centrality
- Characterize network structure
  - Scale-free or not

http://www.barabasilab.com/pubs/CCNR-ALB_Publications/200305-01_SciAmer-ScaleFree/200305-01_SciAmer-ScaleFree.pdf

CLIQUES AND CONNECTED COMPONENTS

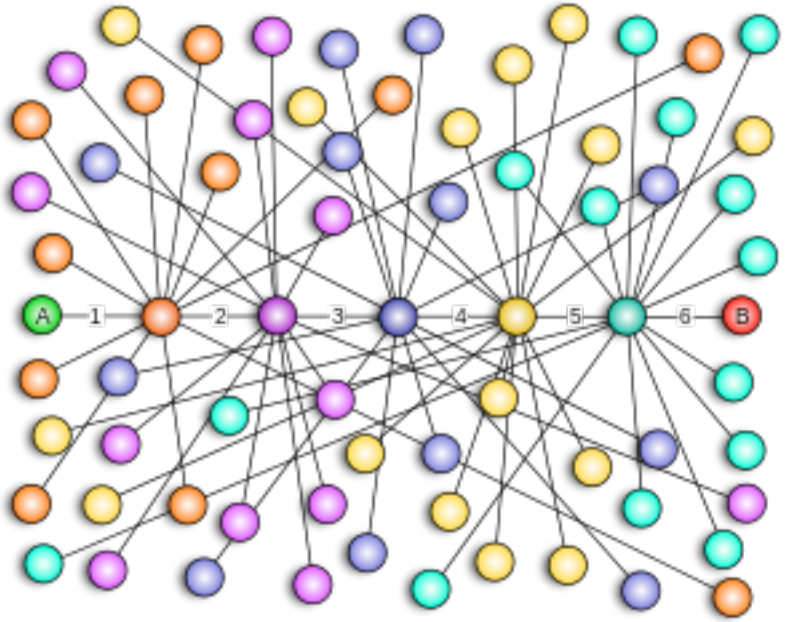




Clique: a subset of nodes of an undirected network, such that every two distinct nodes in the clique are connected.

Connected component: a subset of nodes of an undirected network in which any two nodes are connected to each other by (at least one) path(s)

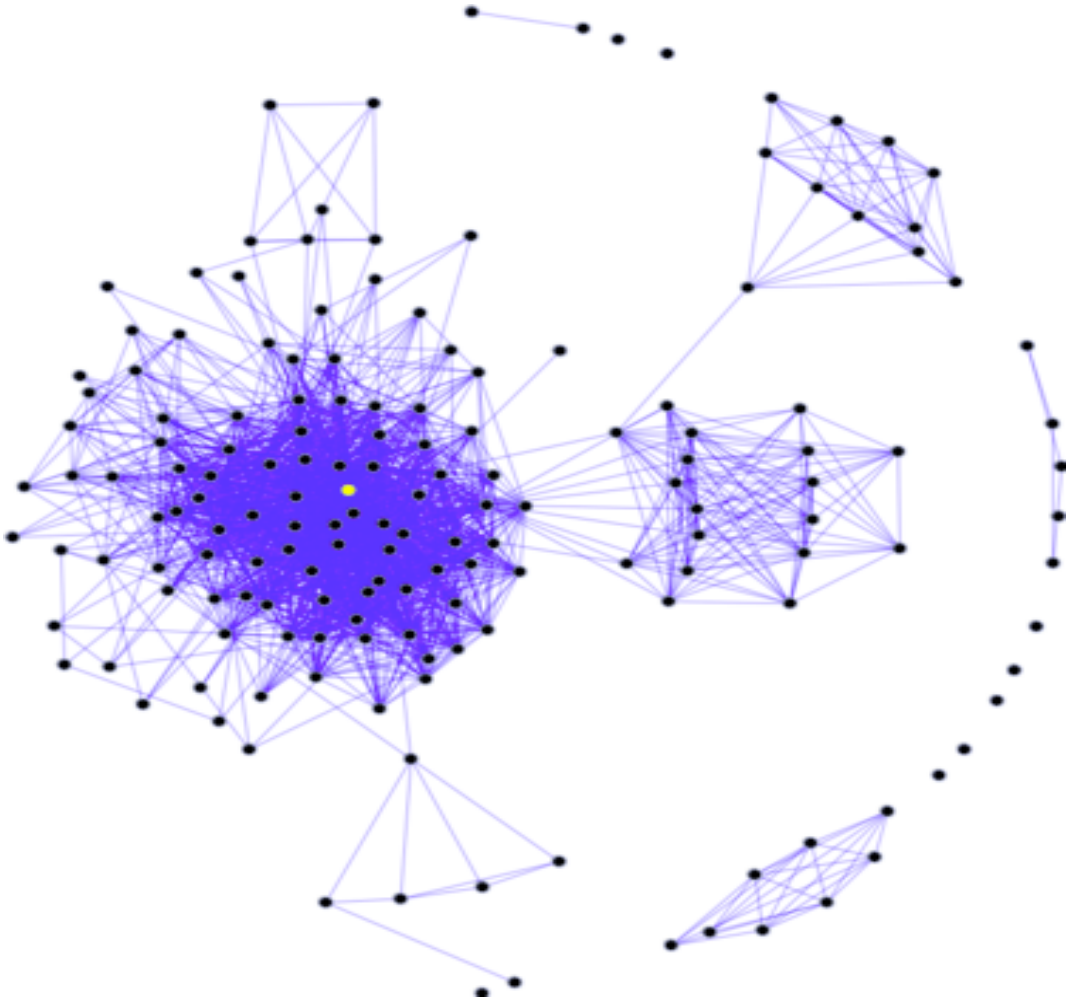
DISTANCE



?

The **distance** between two nodes in a network is the number of edges in a shortest path (also called a **graph geodesic**) connecting them.

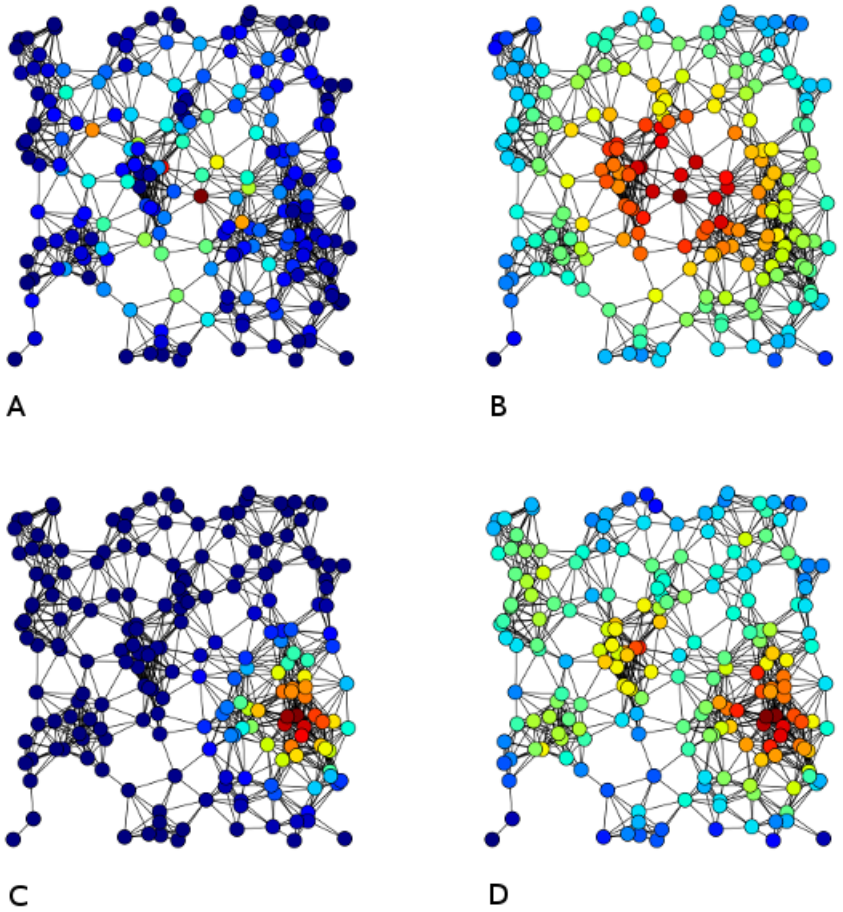
BRIDGES



A **bridge** is an edge of a network whose deletion increases its number of connected components.

NETWORK CENTRALITY

In network analysis, indicators of **centrality** identify **the most important nodes** within a network. Applications include identifying the *most influential person(s)* in a social network, *key infrastructure nodes* in the Internet or urban networks, and *super-spreaders* of disease.



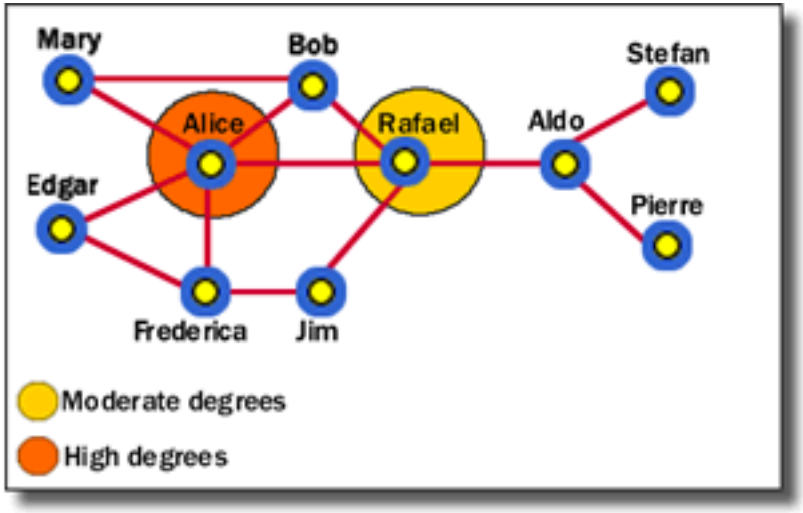
Examples of A) Betweenness centrality, B) Closeness centrality, C) Eigenvector centrality, D) Degree centrality

https://en.wikipedia.org/wiki/Centrality

DEGREE CENTRALITY

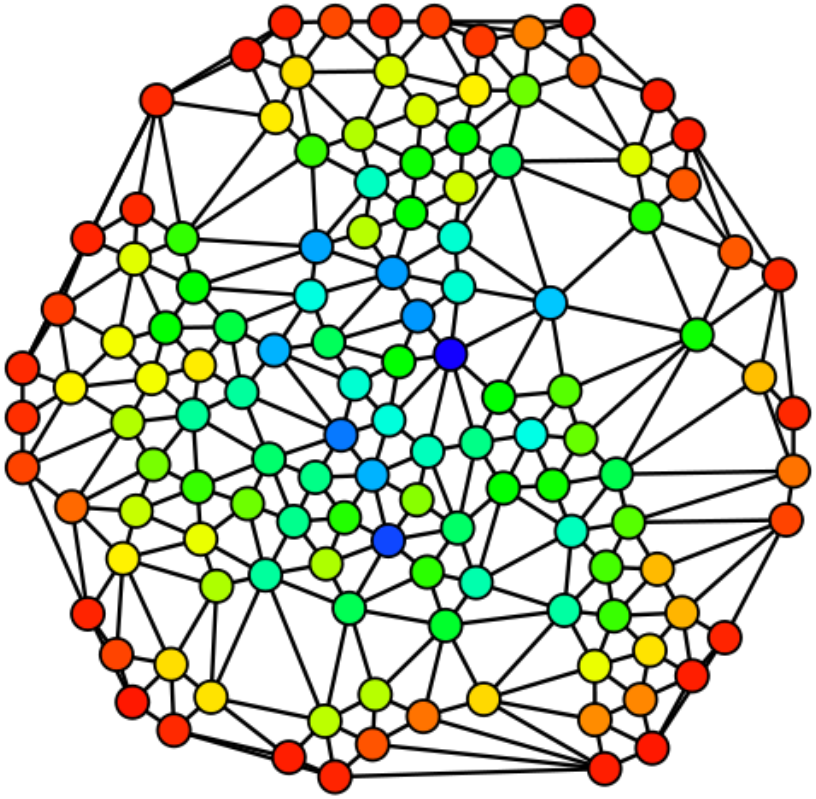
The simplest network centrality measure is **degree centrality**, defined as the *number of links incident upon a node* (i.e., the number of ties that a node has).

The degree can be interpreted in terms of the immediate risk of a node for catching whatever is flowing through the network (such as a virus, or some information).

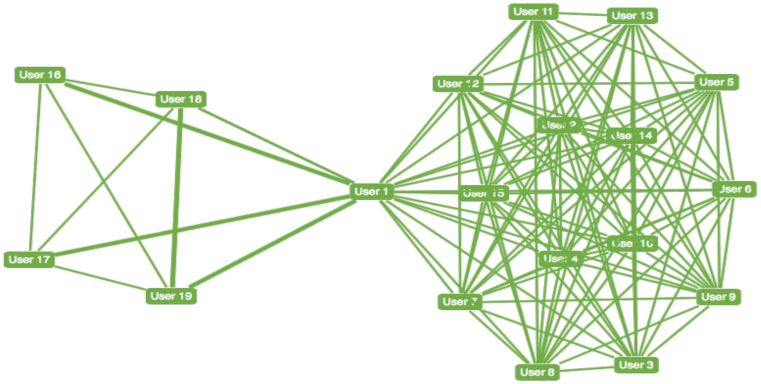


http://www.fmsasg.com/socialnetworkanalysis/SocialNetworkAnalysis_Degrees.gif

BETWEENNESS CENTRALITY



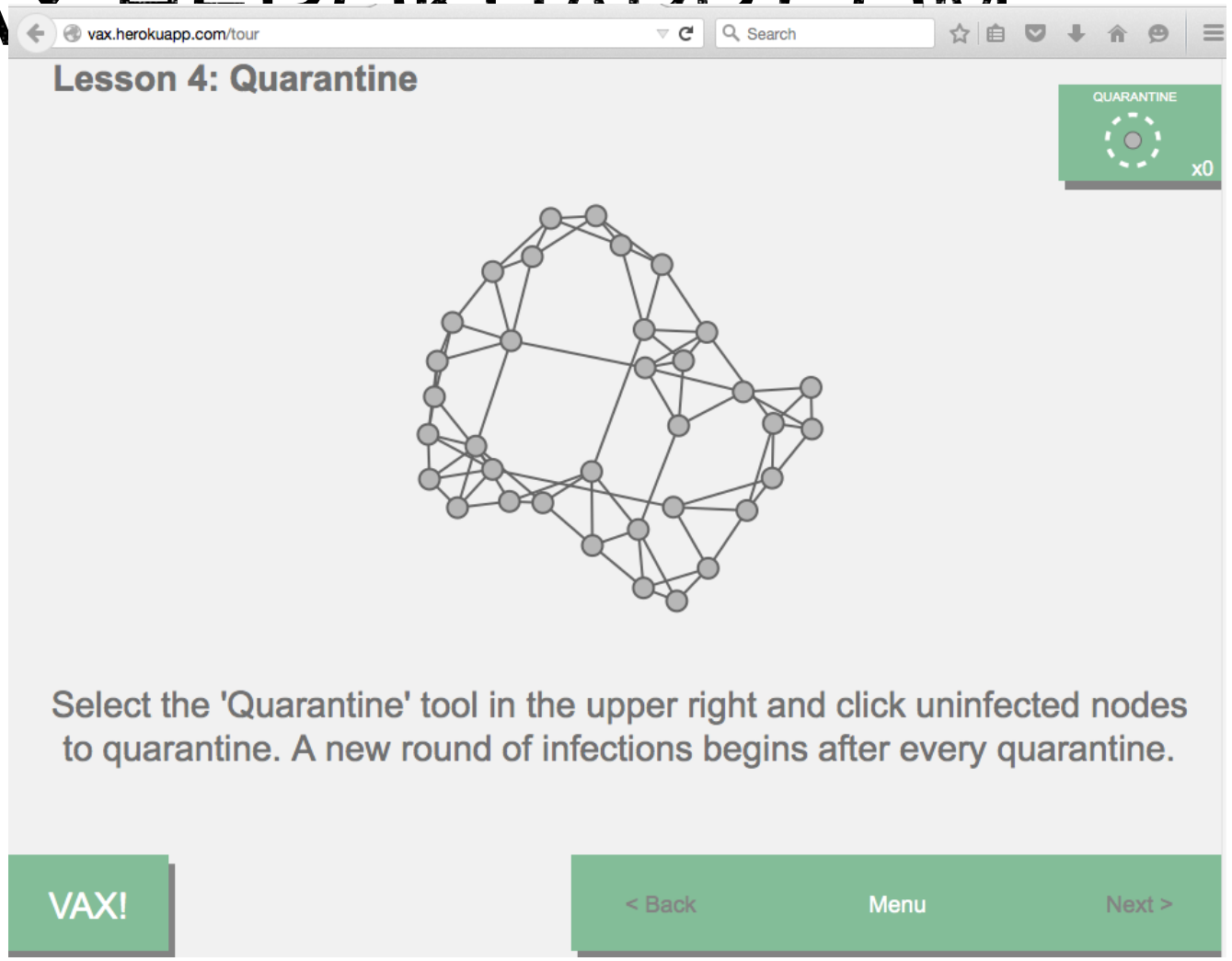
Hue (from red = 0 to blue = max) shows the node betweenness



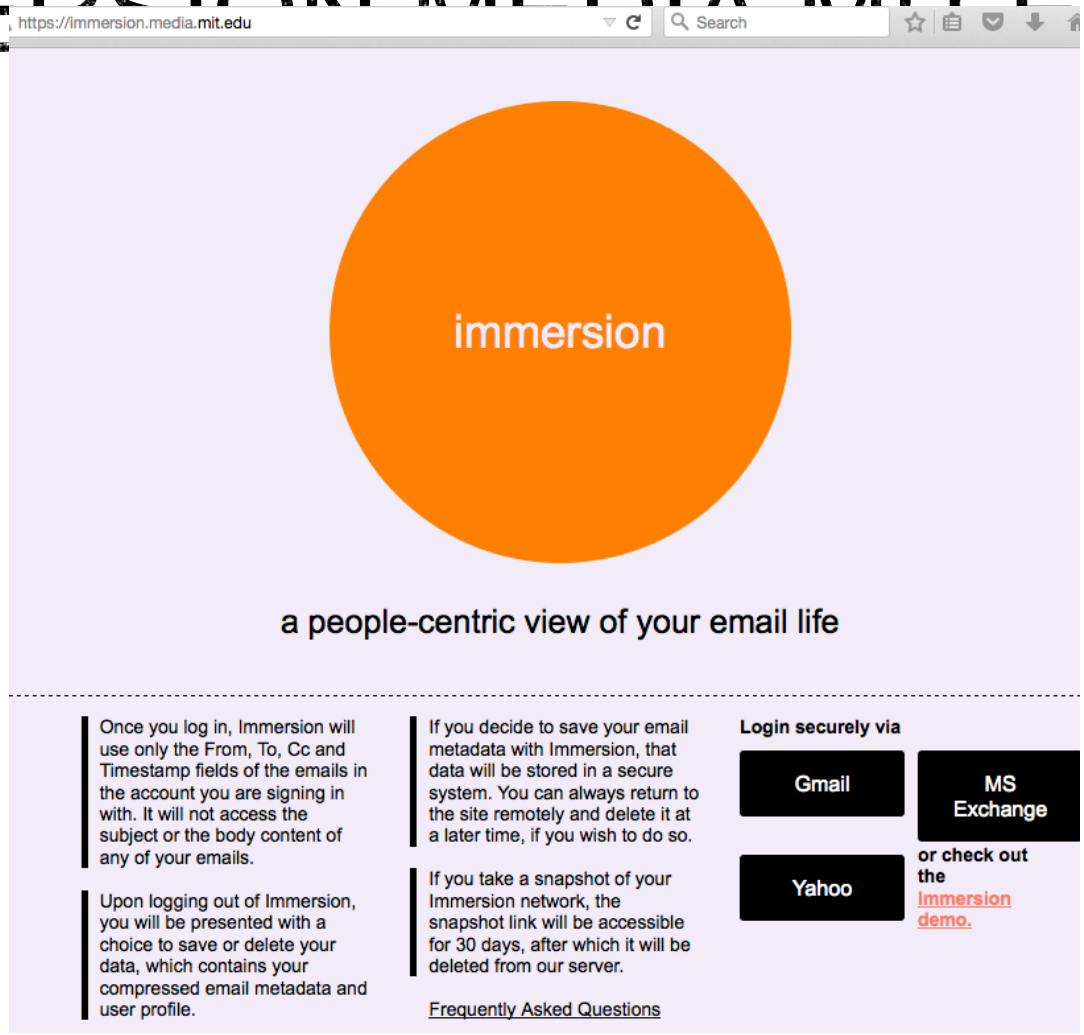
Betweenness centrality: an indicator of a node's centrality in a network. It is equal to the number of shortest paths from all nodes to all others that pass through that node.

A node with high betweenness centrality has a large influence on the transfer of information through the network.

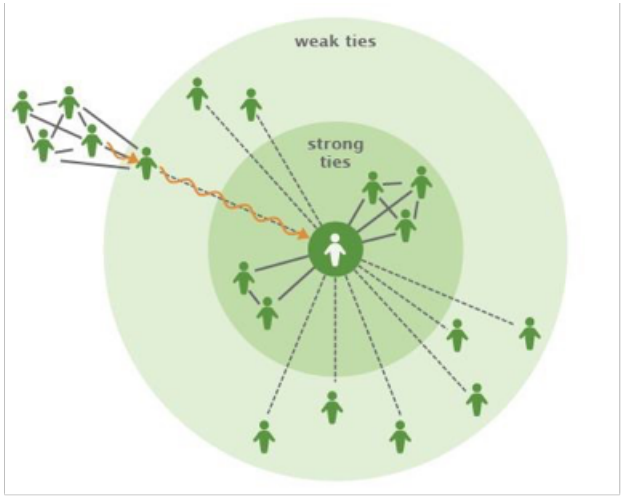
SEE HOW THINGS SPREAD OVER A NETWORK: VAX.HEROKUAPP.COM



VIEW YOUR EMAIL NETWORK: IMMERSION.MEDIA.MIT.EDU



STRENGTH OF WEAK TIES?



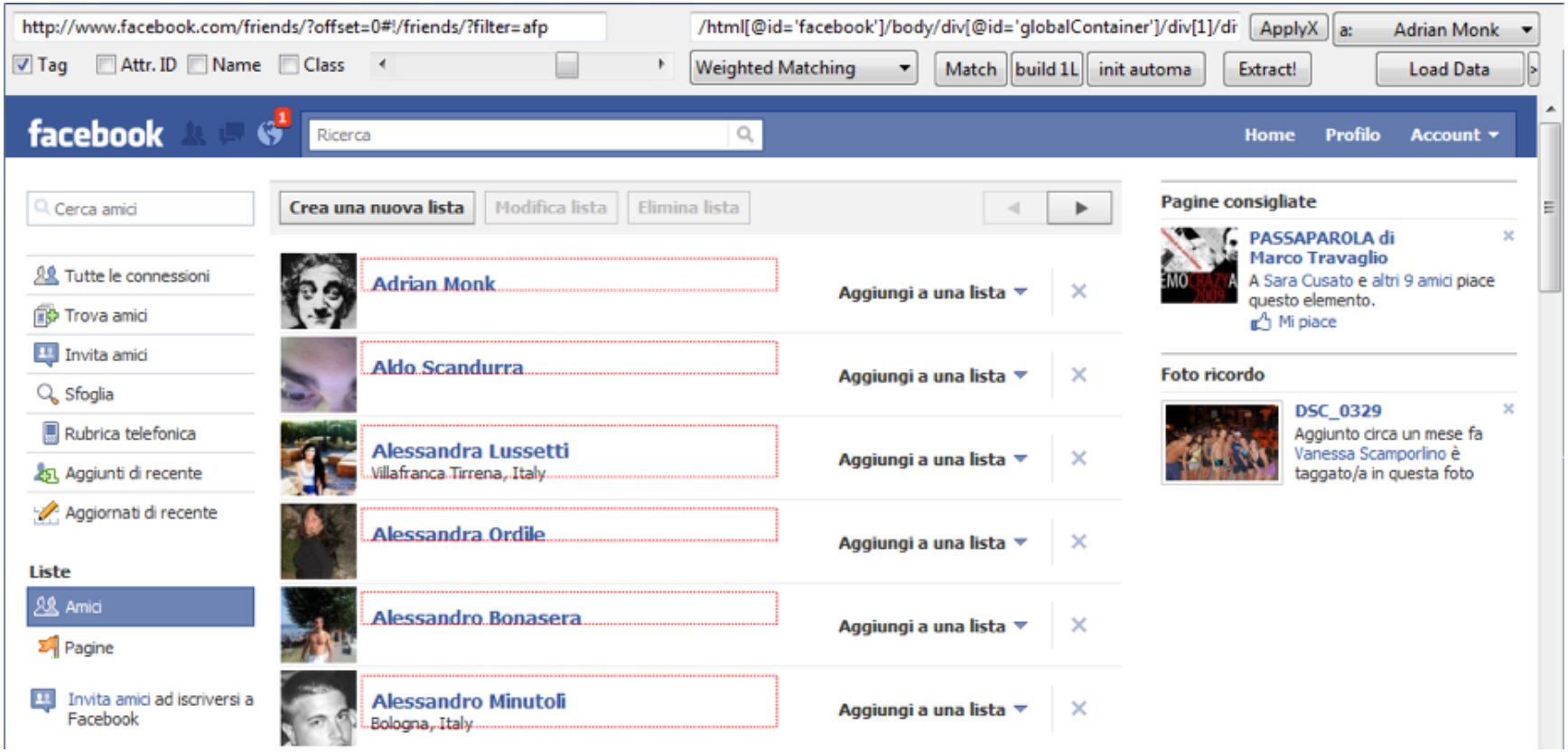
"**Linkage of micro and macro levels** … **generates paradoxes**: **weak ties**, often denounced as generative of alienation, **are** here seen as **indispensable** to **individuals' opportunities** and to their **integration into communities**; **strong ties, breeding local cohesion, lead to overall fragmentation**."

"Paradoxes are a welcome antidote to theories which explain everything all too neatly."

–Mark Granovetter

The strength of weak ties. American J. of Sociology (1973)

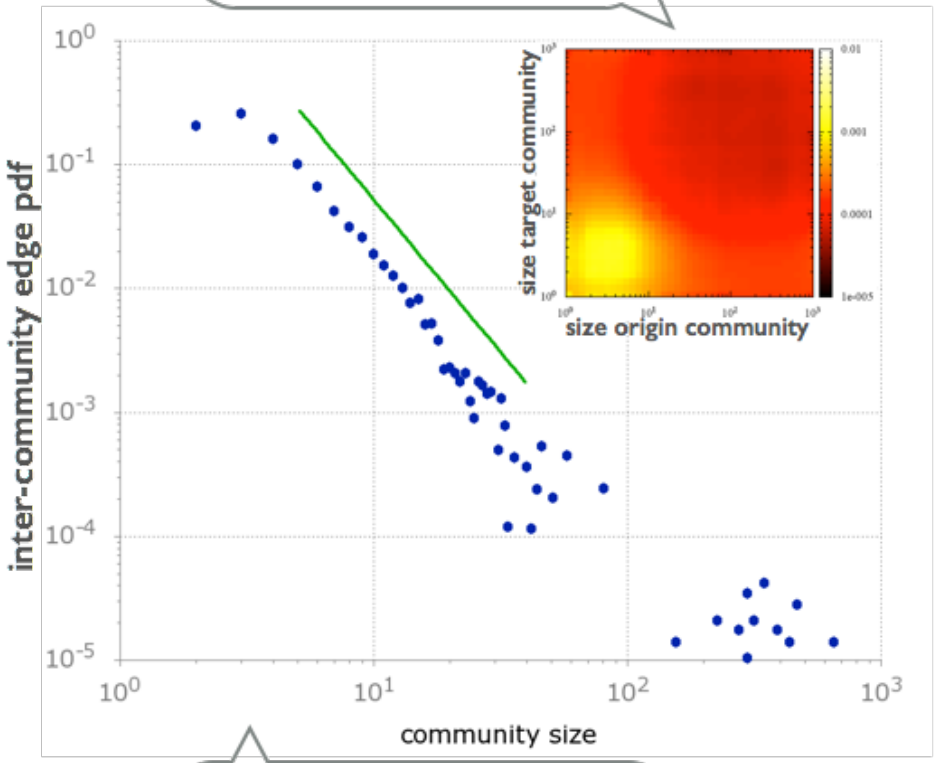
MINING FACEBOOK



Catanese, De Meo, Ferrara, Fiumara, Provetti. **Crawling Facebook for social network analysis purposes.**
ACM WIMS '11: International Conference on Web Intelligence, Mining and Semantics, 2011

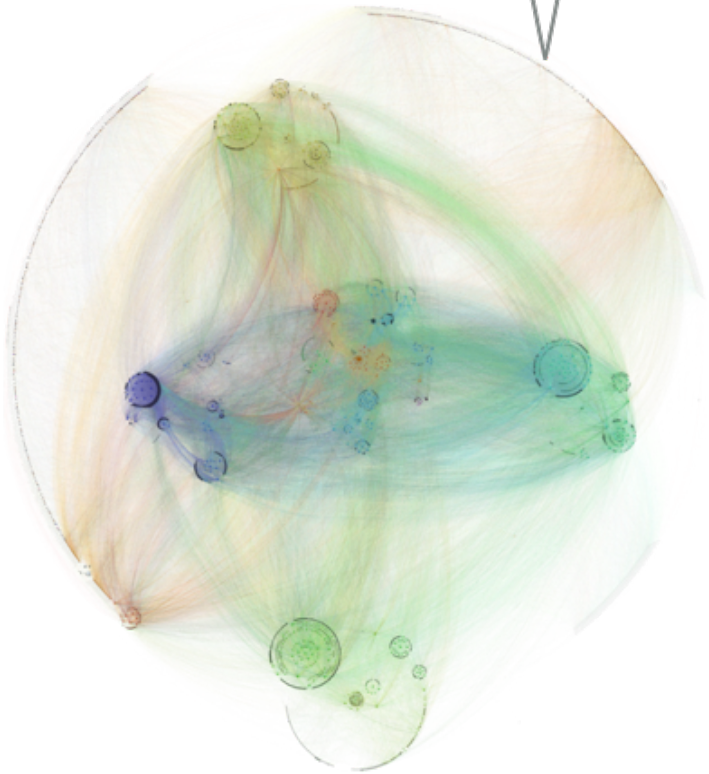
LARGE-SCALE COMMUNITY STRUCTURE

Most weak ties connect the **smaller communities** among each other



Results from label propagation clustering algorithm

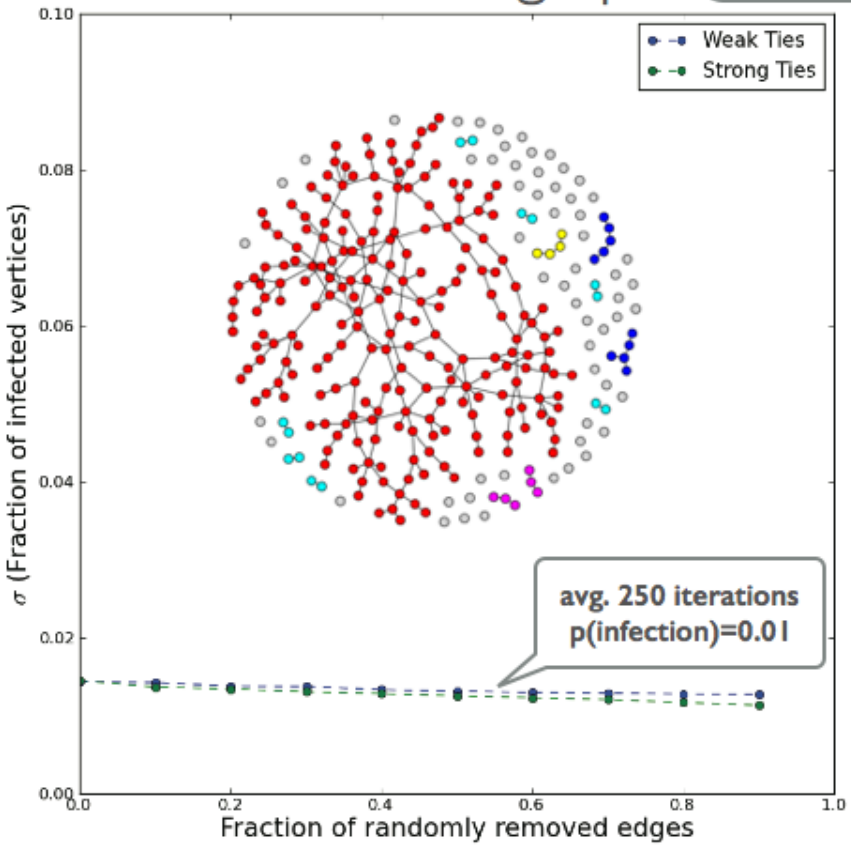
~7M nodes ~35K communities

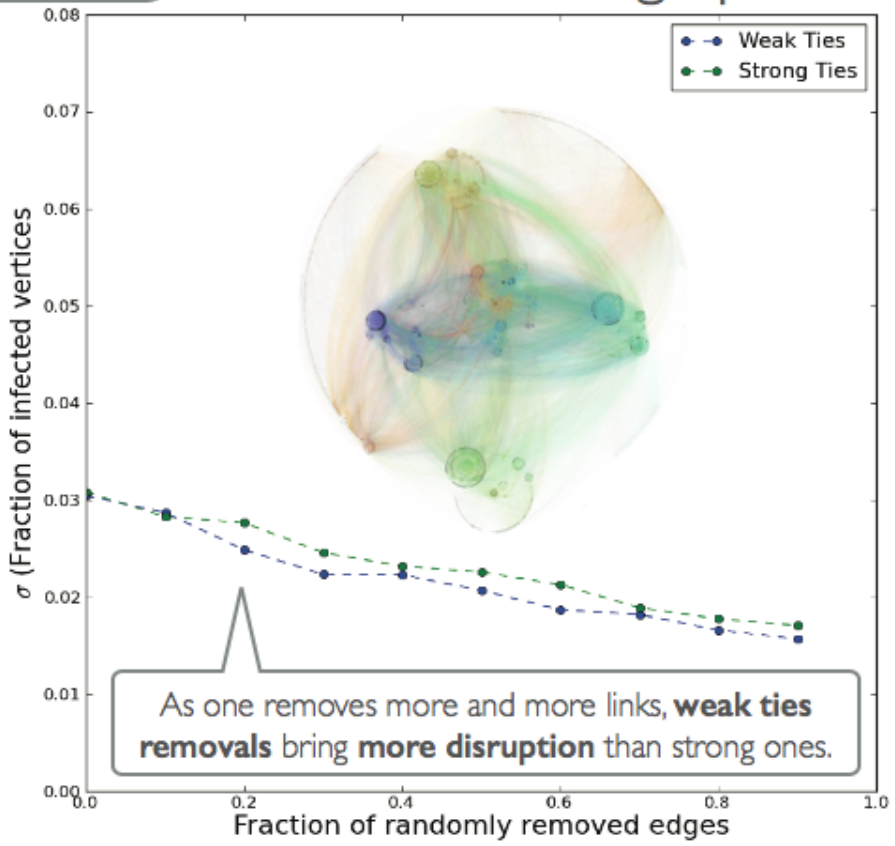


Ferrara. **A large-scale community structure analysis in Facebook.** EPJ Data Science 2012

WEAK TIES AND INFORMATION DIFFUSION

ER random graph | independent cascade model | Facebook social graph





De Meo, Ferrara, Fiumara, Provetti. On Facebook, most ties are weak. Communications of the ACM 2014

GEPHI: INTERACTIVE NETWORK VISUALIZATION

Live tutorial & Resources

USEFUL RESOURCES

- Learn how to use Gephi: https://gephi.org/users/
  - Download: https://gephi.org/users/download/
  - Installation: https://gephi.org/users/install/
  - Tutorial: https://gephi.org/users/tutorial-visualization/
  - Layouts: https://gephi.org/users/tutorial-layouts/

- Data Scientist toolbox + Datasets (curated by me): http://www.emilio.ferrara.name/code/useful-resources/
- Mining the Social Web (curated by me): http://www.emilio.ferrara.name/i400-590-mining-the-social-web/

Thanks!

Questions? @jabawack / emiliofe@usc.edu / ask me!