Boston Data Day

Skill-A-Thon 2013:

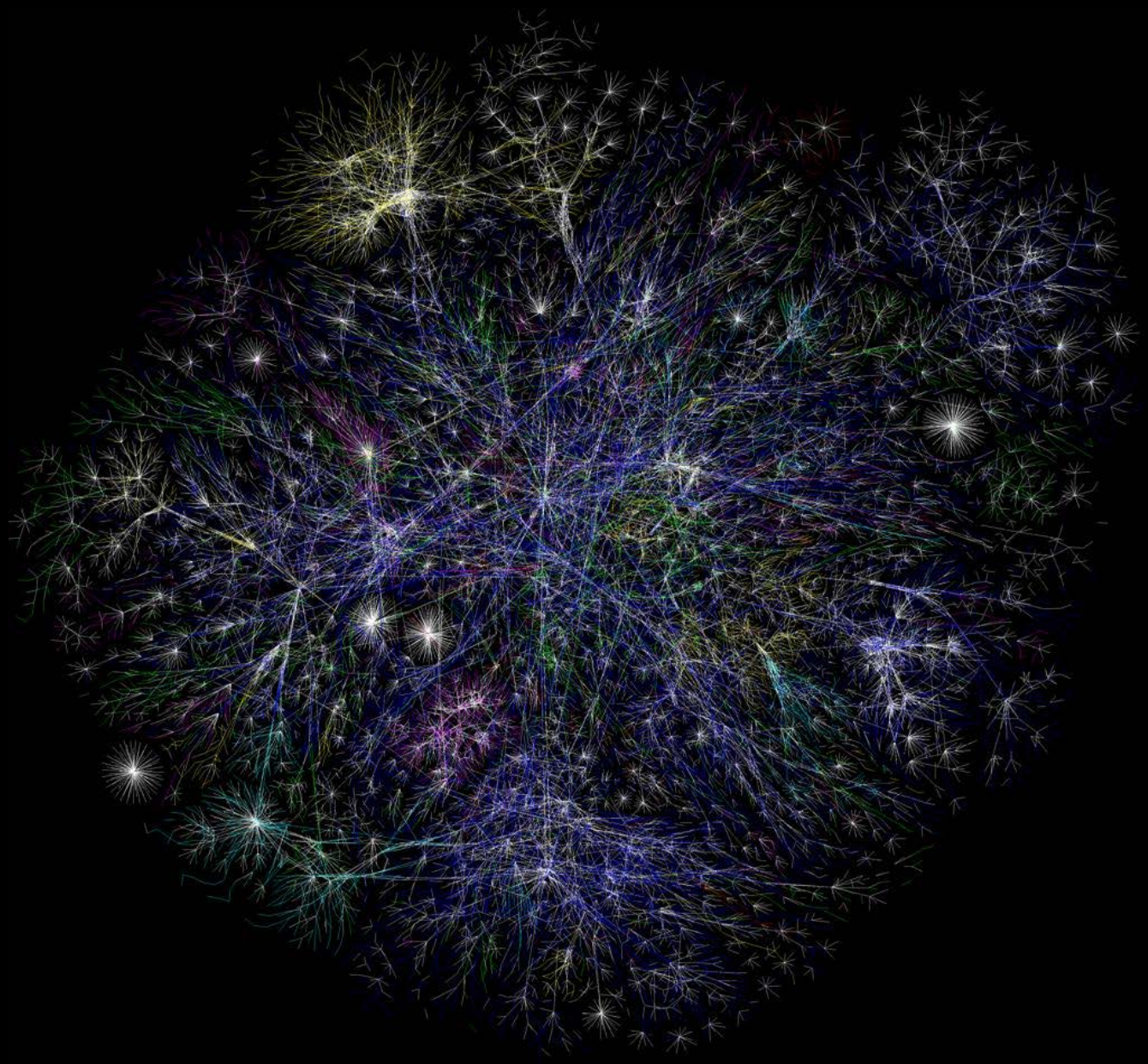
Thinking with networks

Katherine Ognyanova (Katya) • Postdoctoral Research Fellow, The Lazer Lab

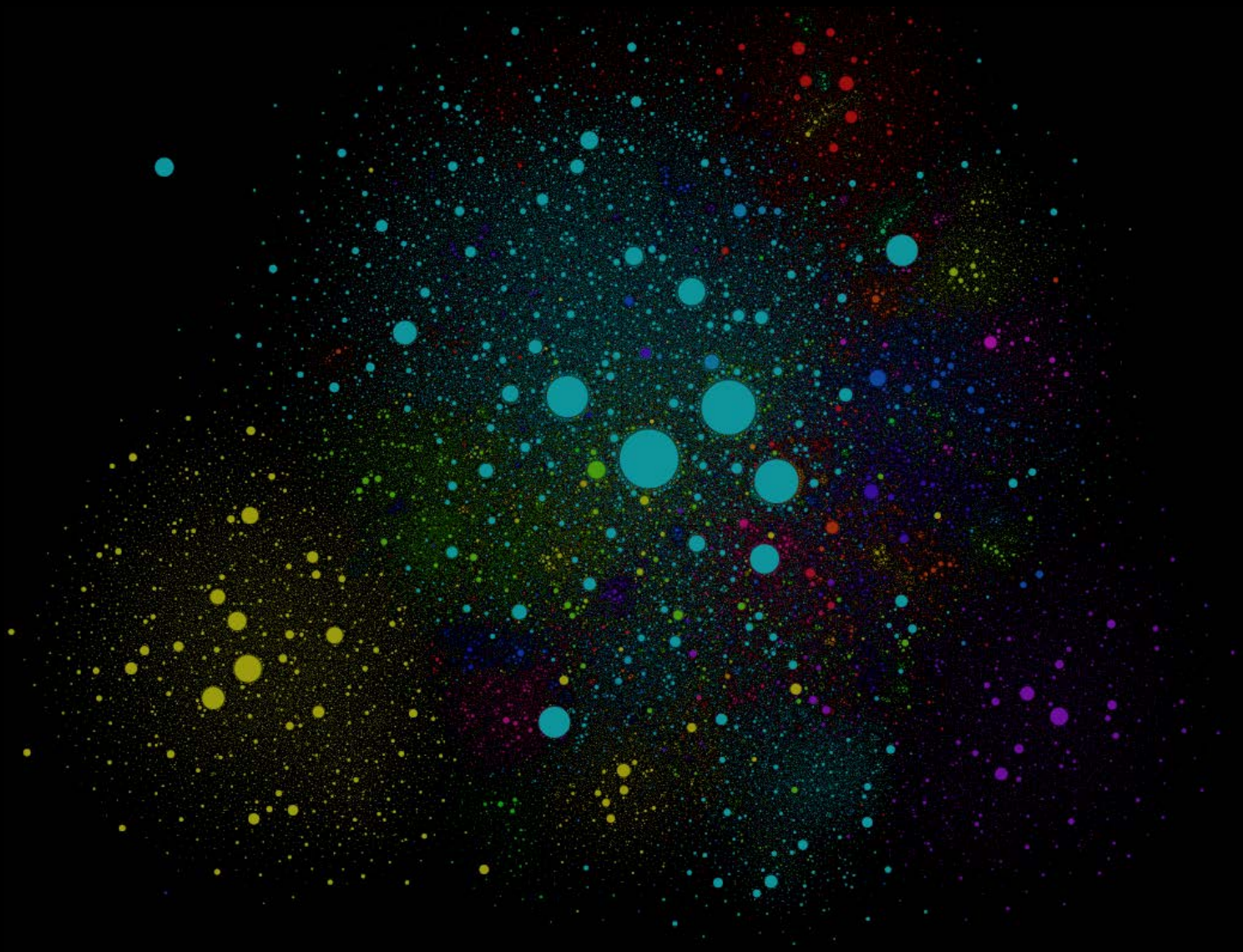
Email: katya@ognyanova.net • Website: www.kateto.net • Twitter: @ognyanova

Where do we see network data?

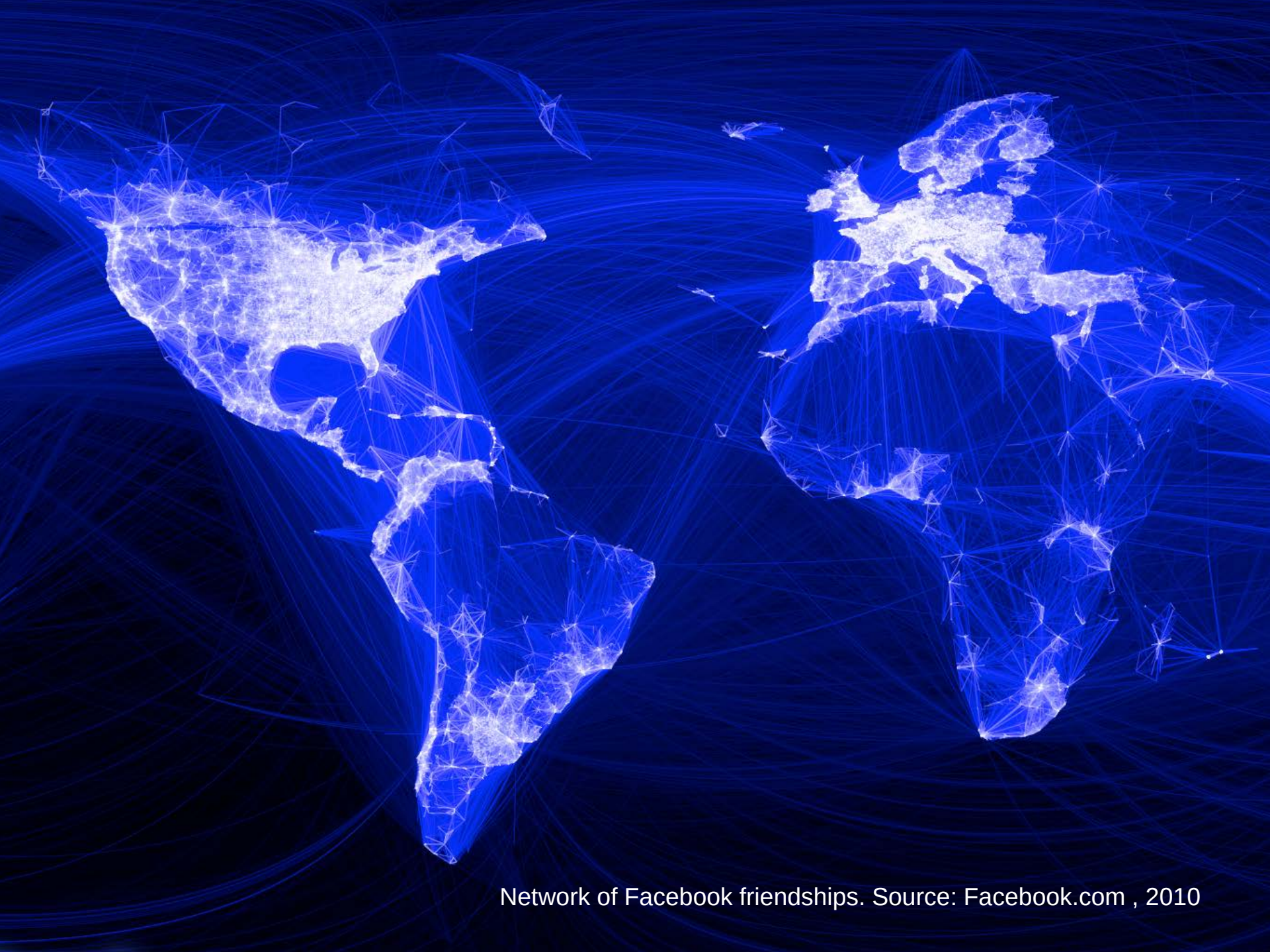
Global Flights Network. Source: www.visualizing.org, 2012



Internet Backbone: Each node represents an IP address. Source: Wikipedia

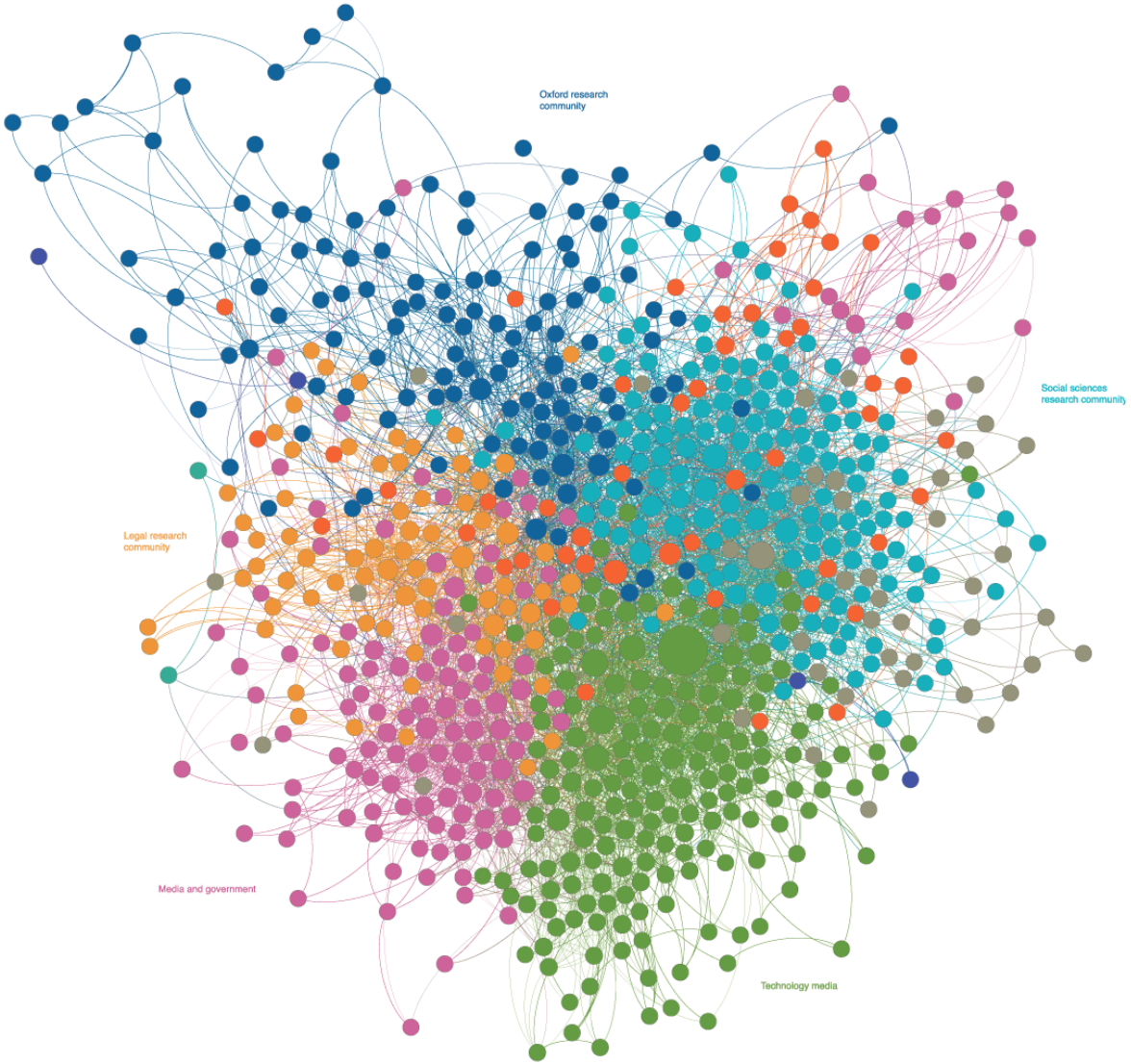


Map of the Web: Nodes represent websites. Source: internet-map.net

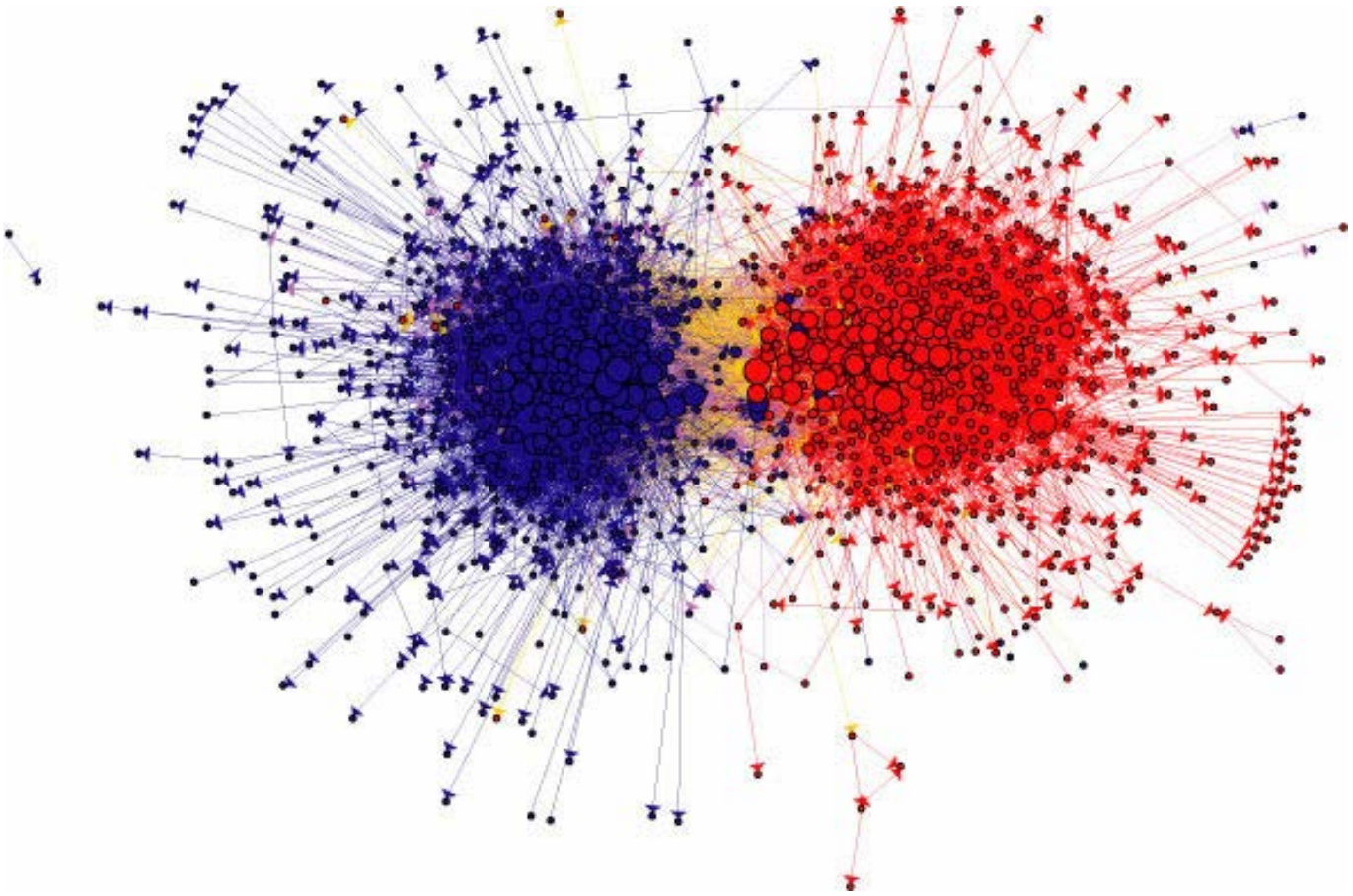


Network of Facebook friendships. Source: Facebook.com , 2010

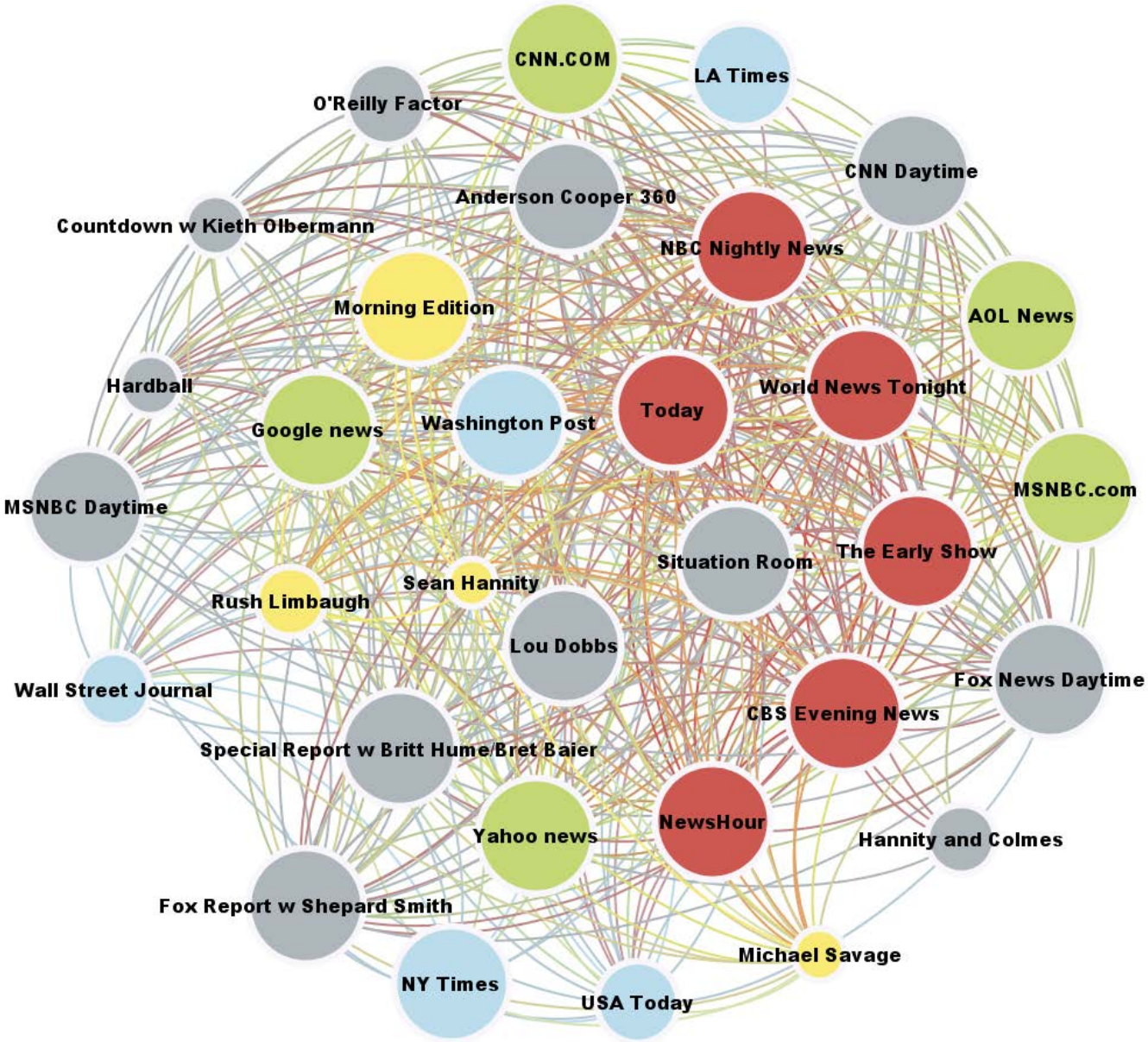
Wikipedia co-editing patterns among top editors in English. Source: Hogan, 2012



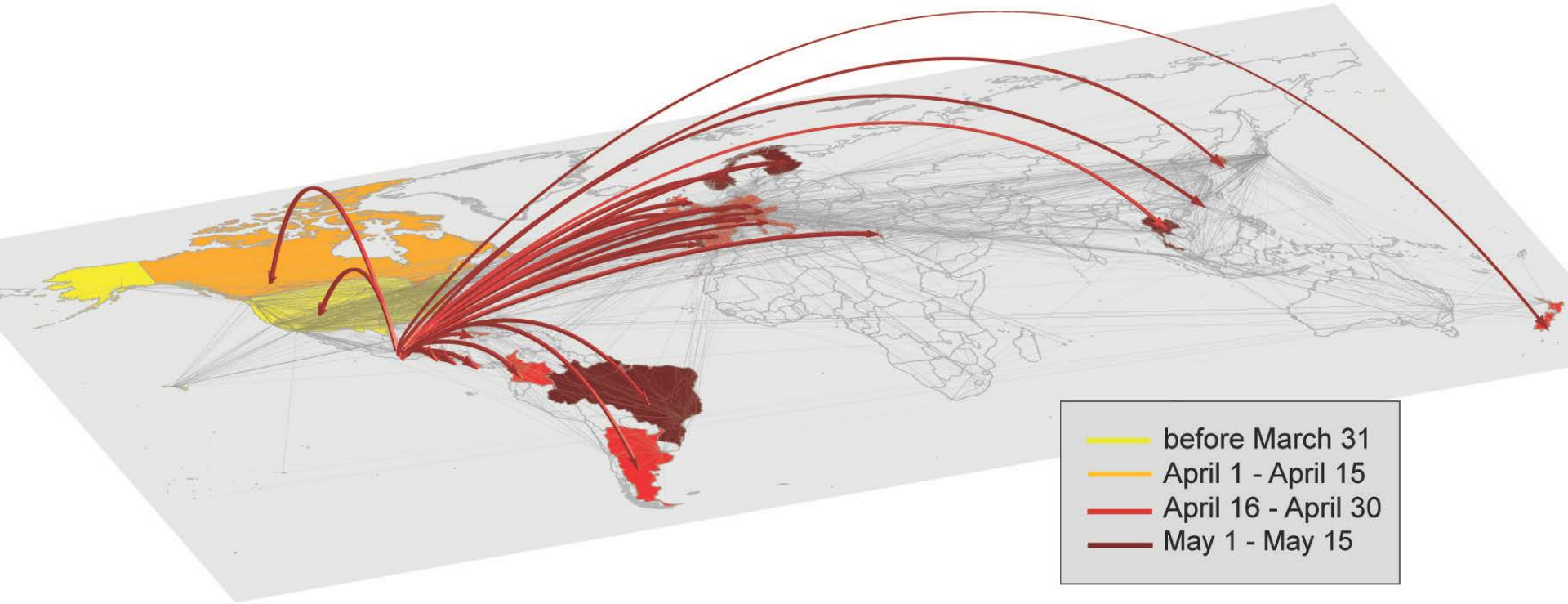
Twitter network, @OIIOxford. Source: Hogan, 2011



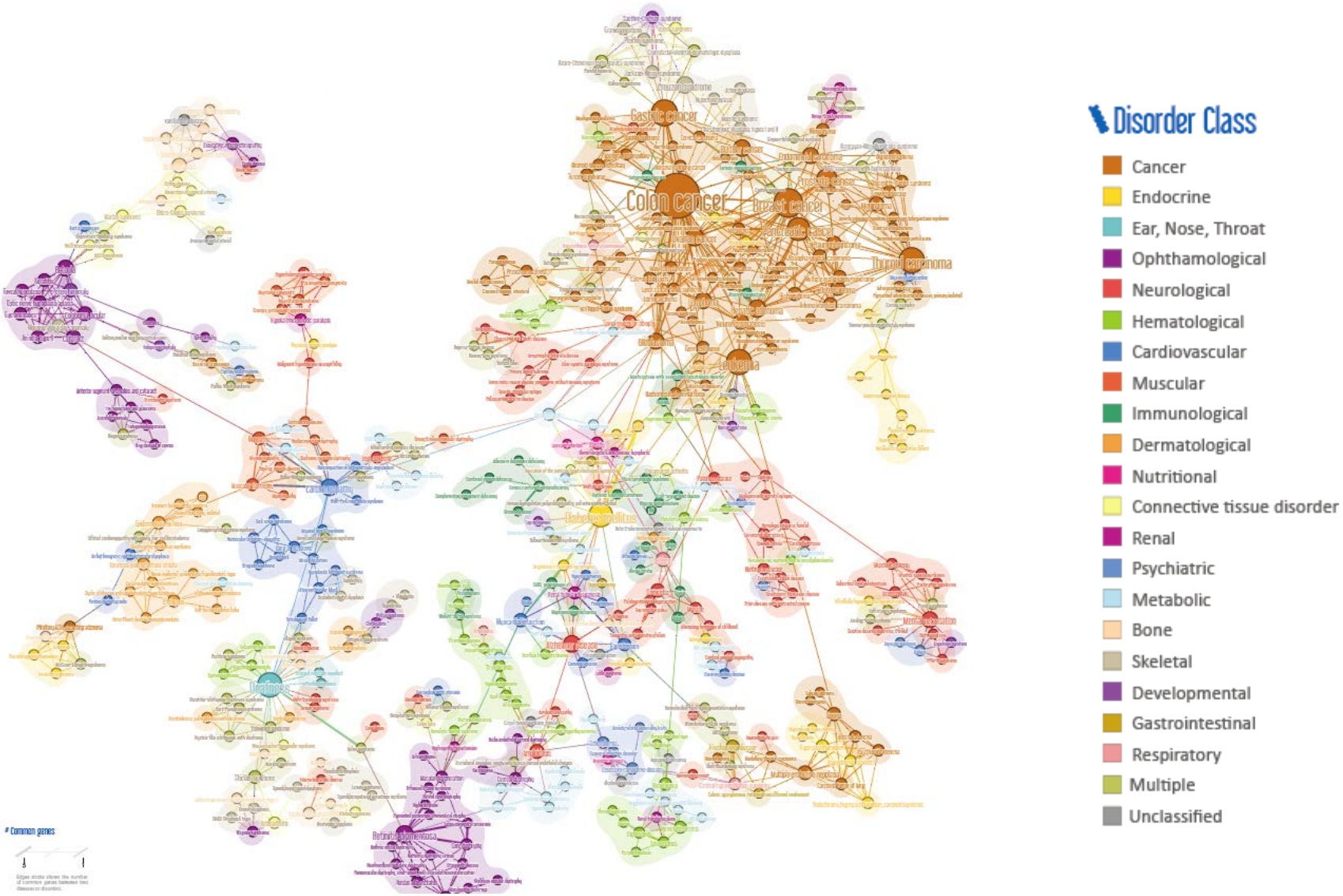
Linking patterns in the US political blogosphere. Source: Adamic & Glance, 2005



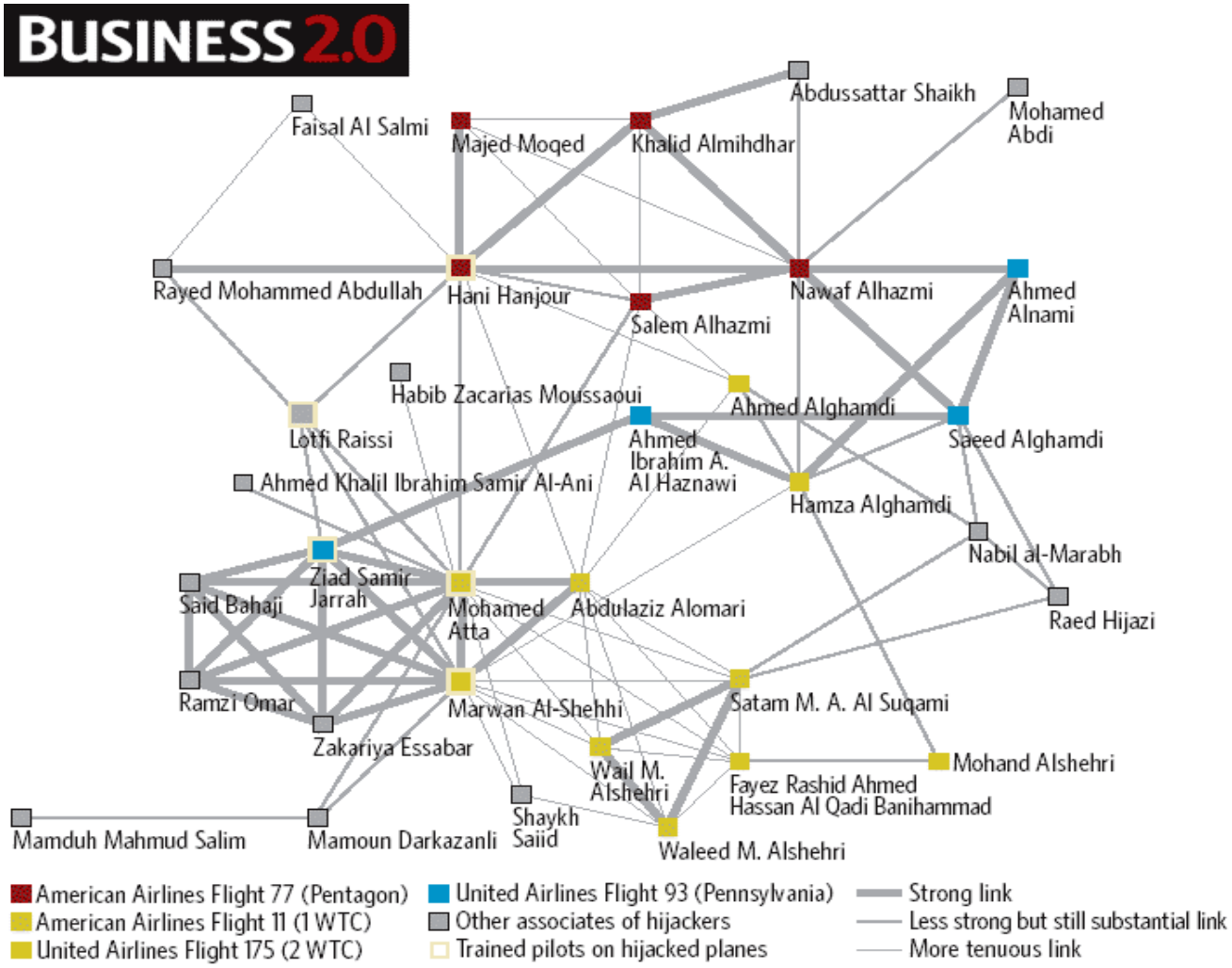
Adoption of topics in the US media system. Source: Ognyanova, 2013



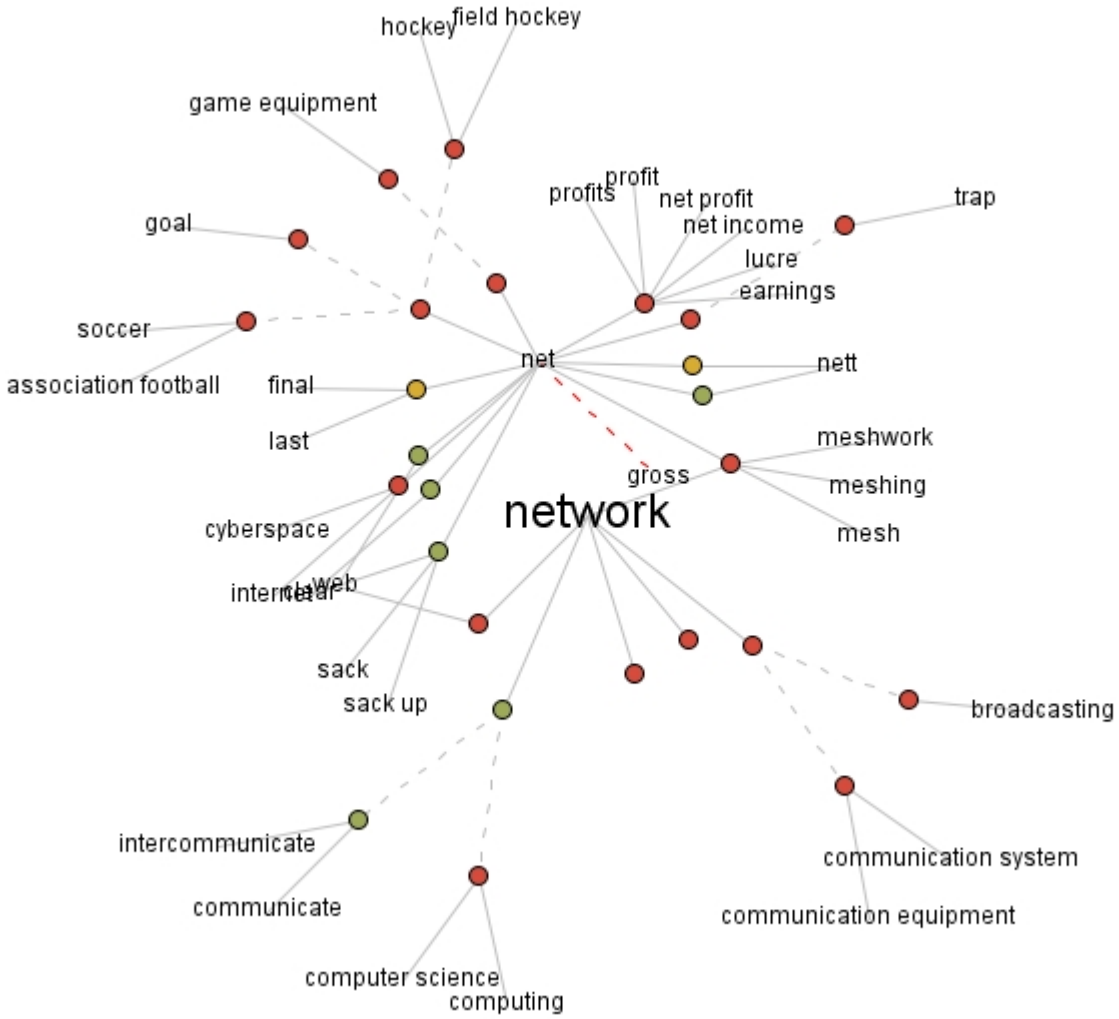
Global Spread of N1H1 and human mobility. Source: gleamviz.org, 2011



Disease network: phenotype and disease gene associations. Source: Diseasome.eu, 2013



Terrorism Network. Source: Business 2.0, December 2001. *Six Degrees of Mohamed Atta*



Semantic Networks: Relationships between concepts. Source: Thinkmap Visual Thesaurus, 2012

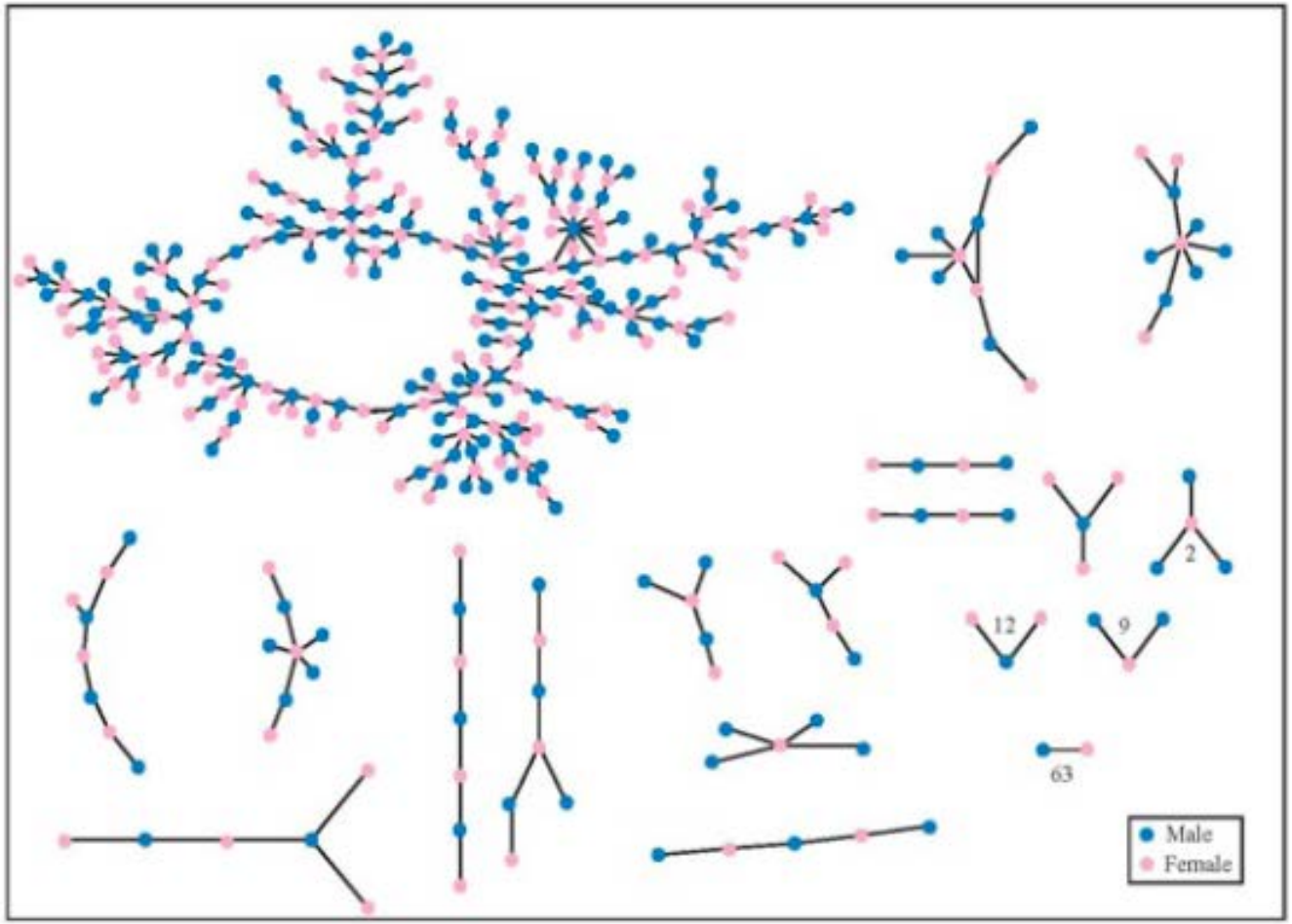


Figure 2.7: A network in which the nodes are students in a large American high school, and an edge joins two who had a romantic relationship at some point during the 18-month period in which the study was conducted [50].

High-school Romance Network. Source: Easley & Kleinberg (2010) Networks, Crowds and Markets

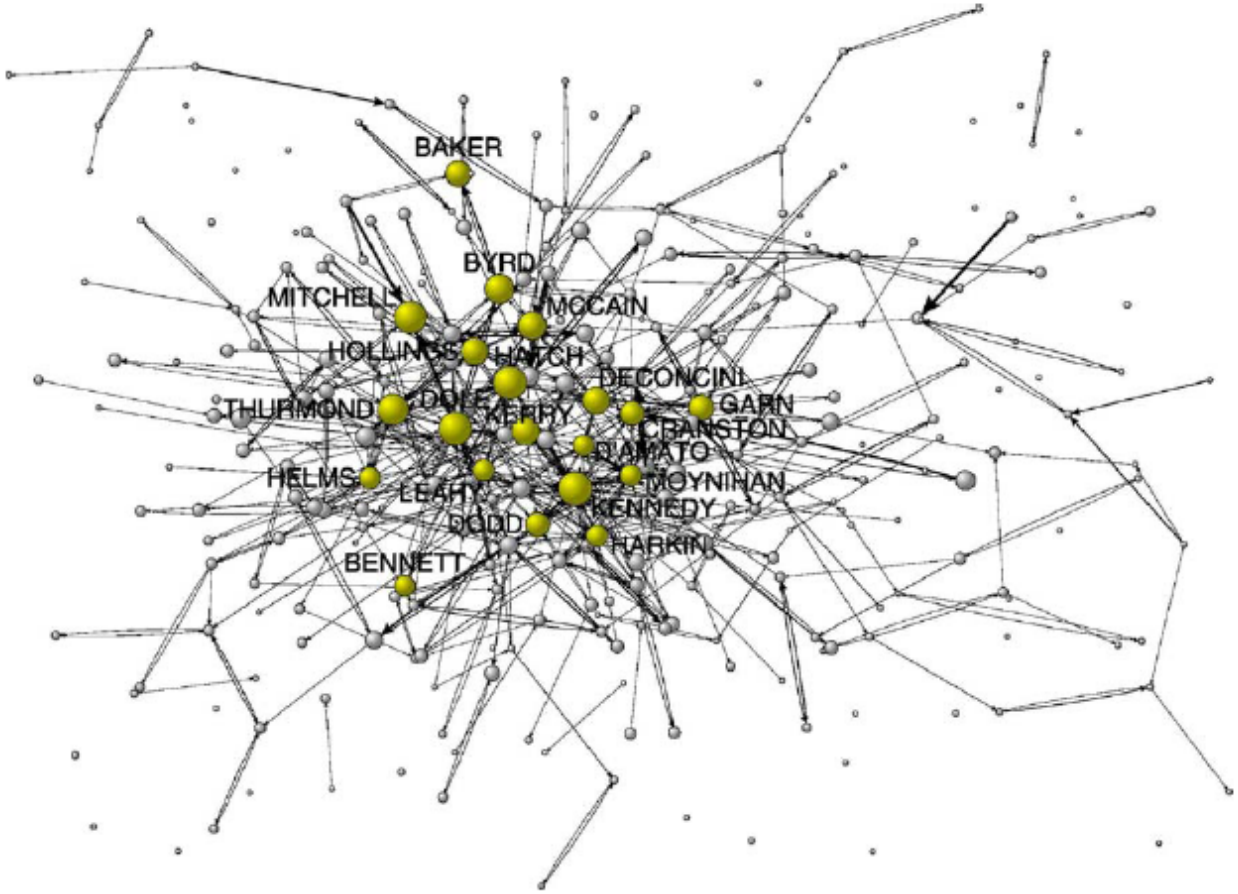
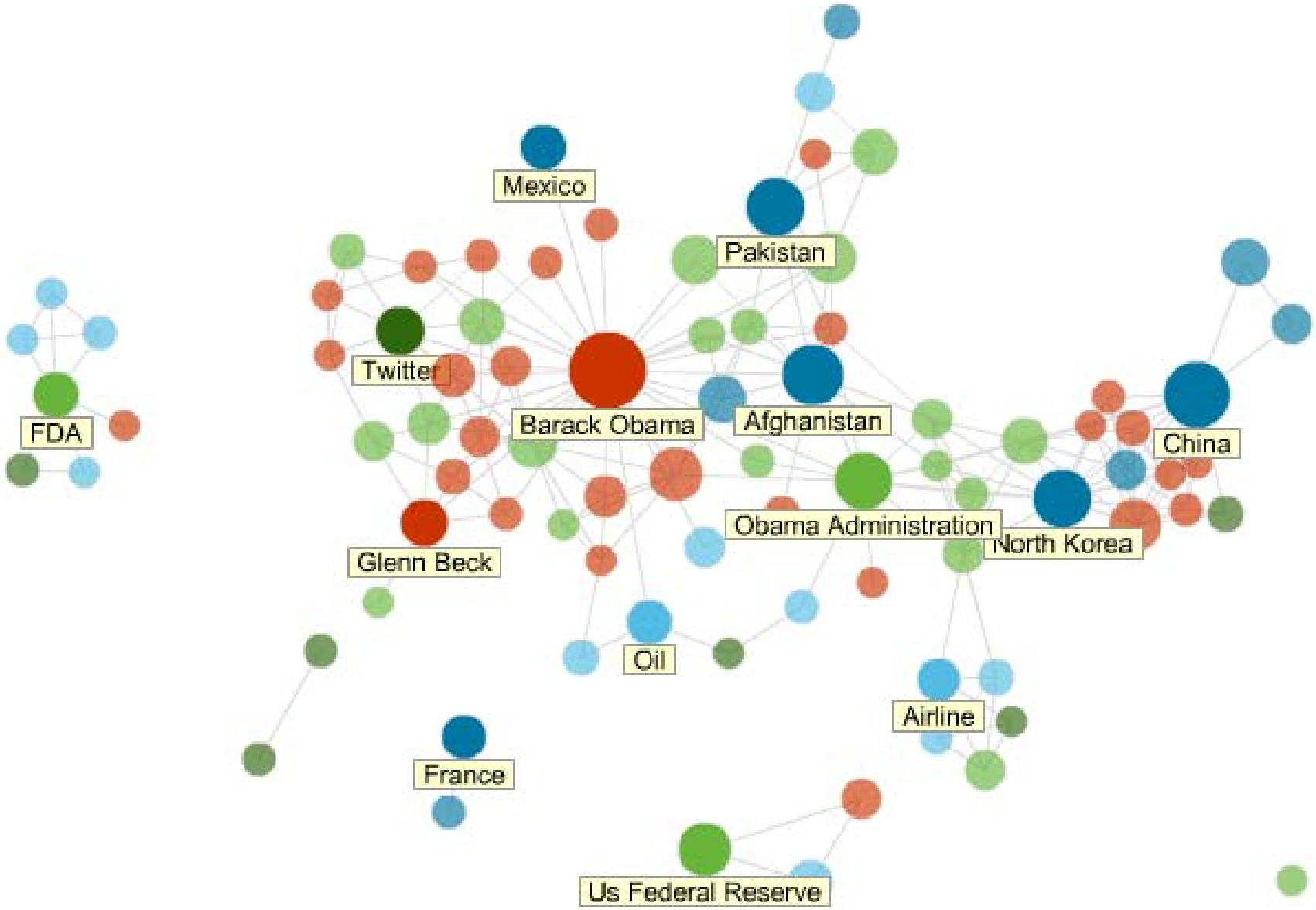
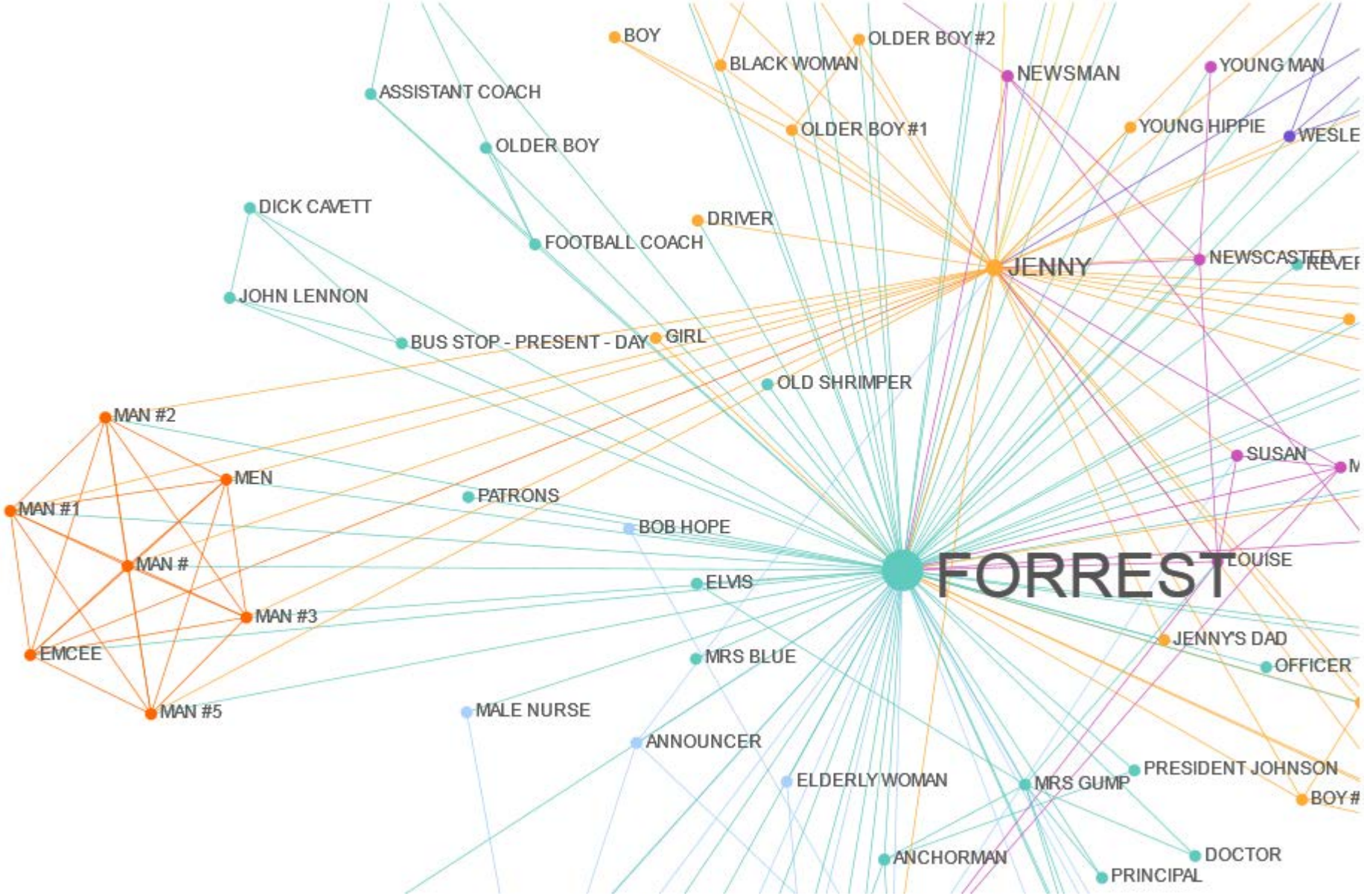


Fig. 4. Strongest weighted cosponsorship ties in the full Senate network, 1973–2004. Note: Size of each vertex is proportional to the Senator's connectedness score, the width of each arrow is proportional to the weighted quantity of bills cosponsored w_{ij} (values of $w_{ij} < 10$ not shown) and vertices that represent the top 20 Senators are identified by name. Figure drawn using Kamada–Kawai algorithm in Pajek (de Nooy et al., 2005).

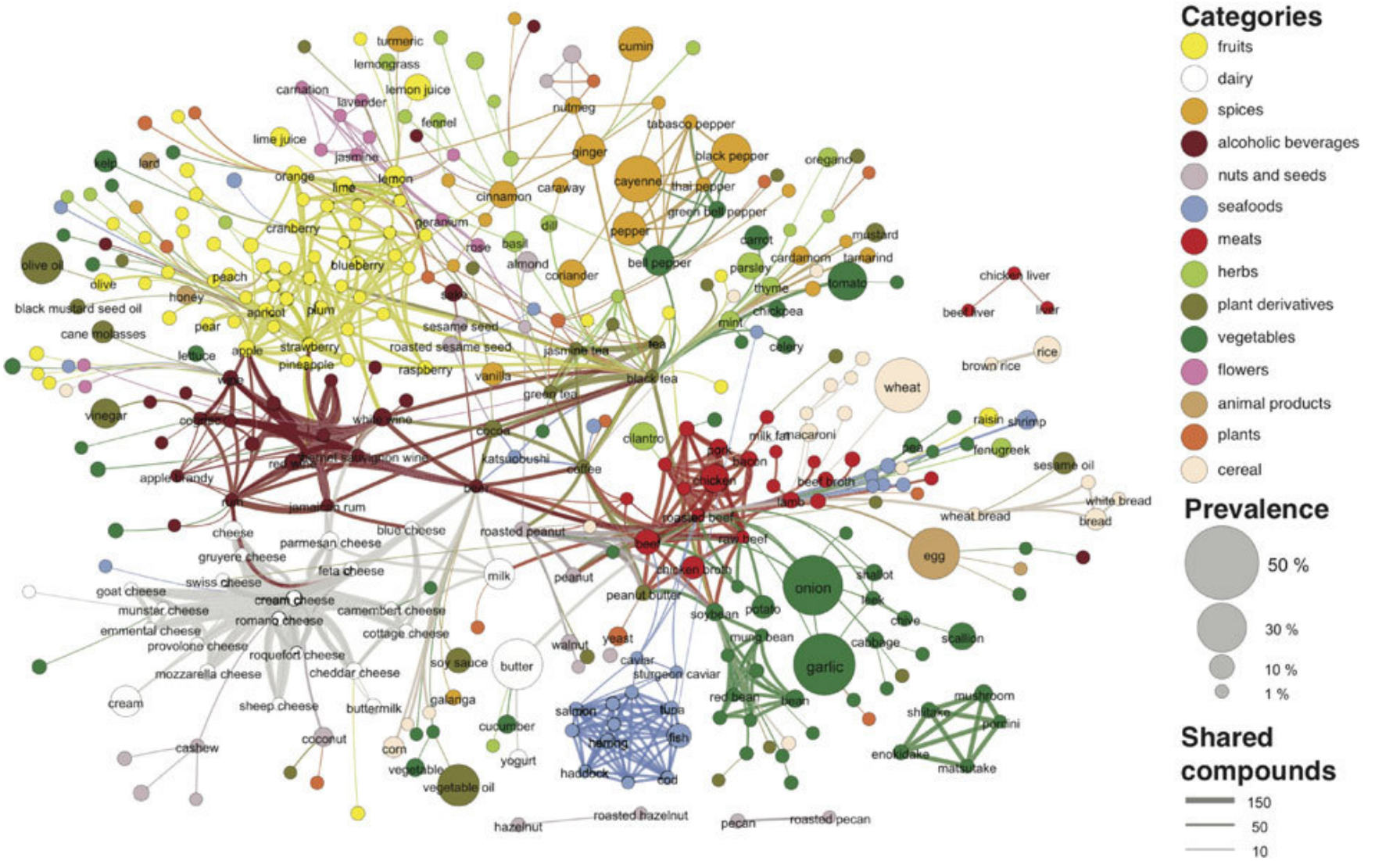
US Senate co-sponsorship network 1973-2004. Source: Fowler, 2006



Network of news topics. Source: Slate News Dots, 2010



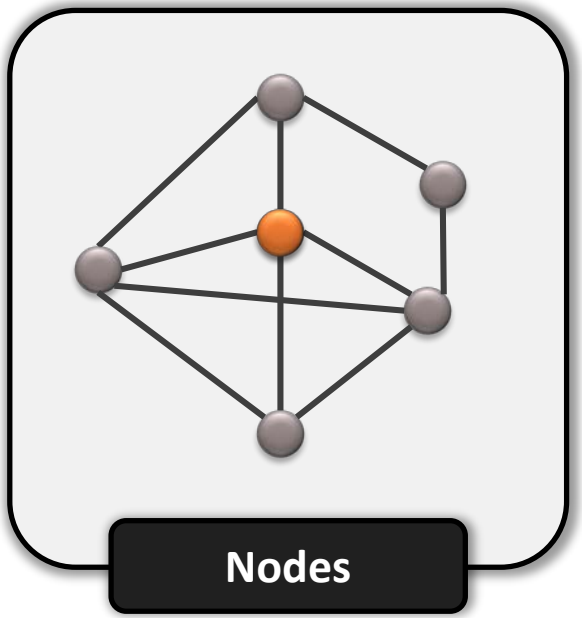
Forrest Gump movie characters social network. Source: moviegalaxies.com, 2013

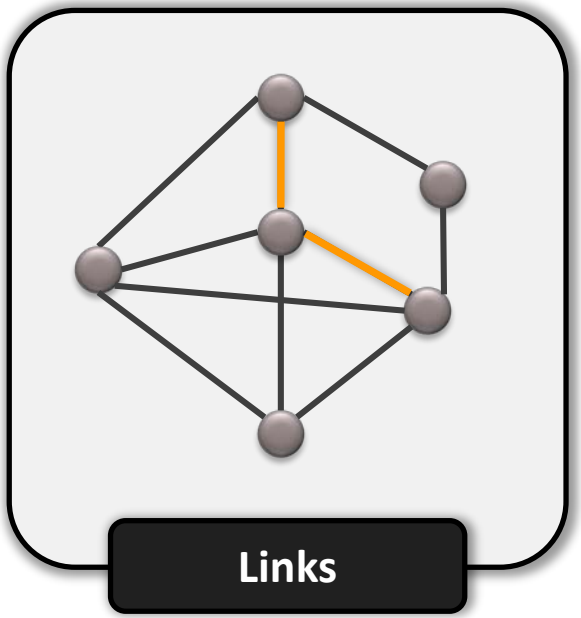


Flavor network and the principles of food preparing. Source: Ahn, Ahnert, Bagrow & Barabasi , 2012

What do we do with network data?

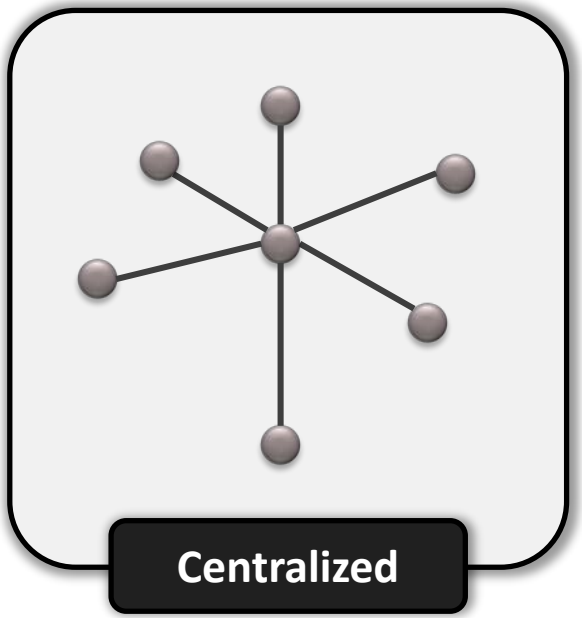
Identification of key elements

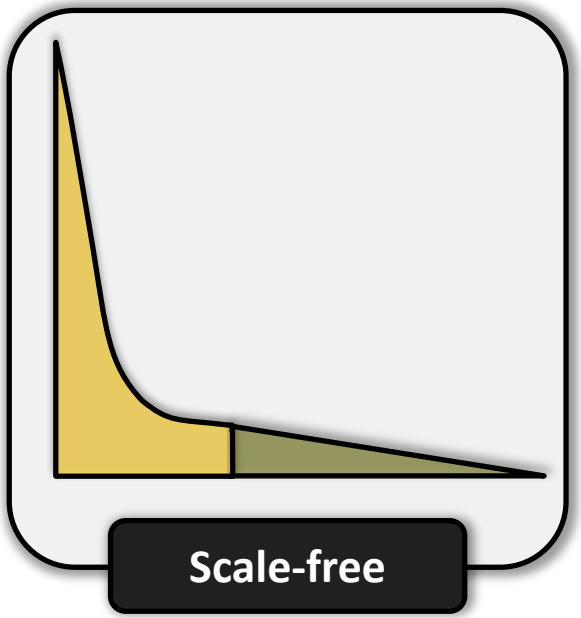


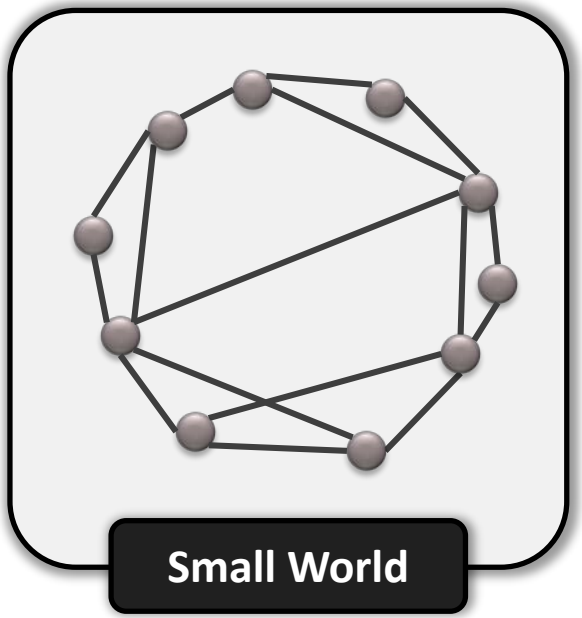


Examples

Variation in global network structure

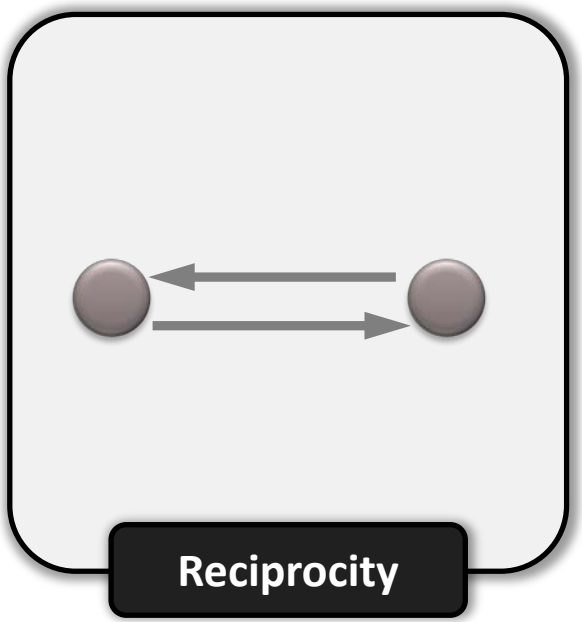


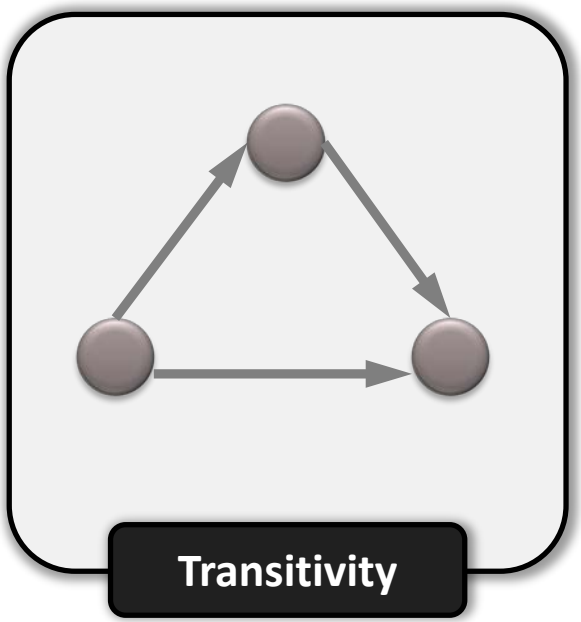


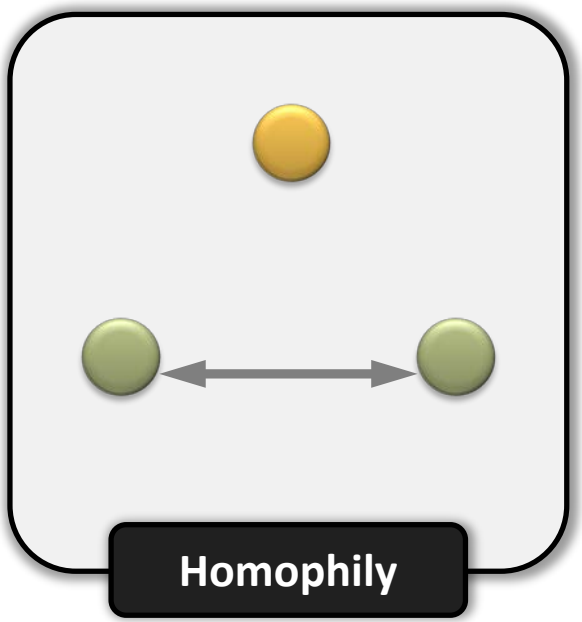


Examples

Drivers of link formation and dissolution

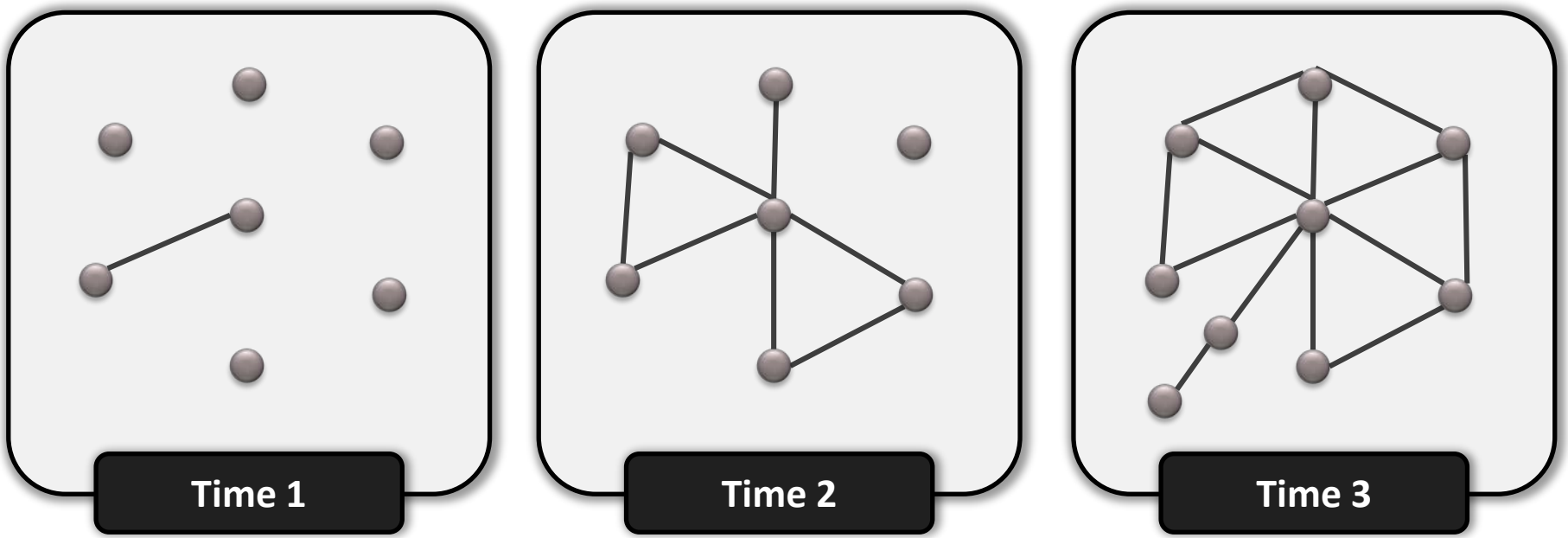






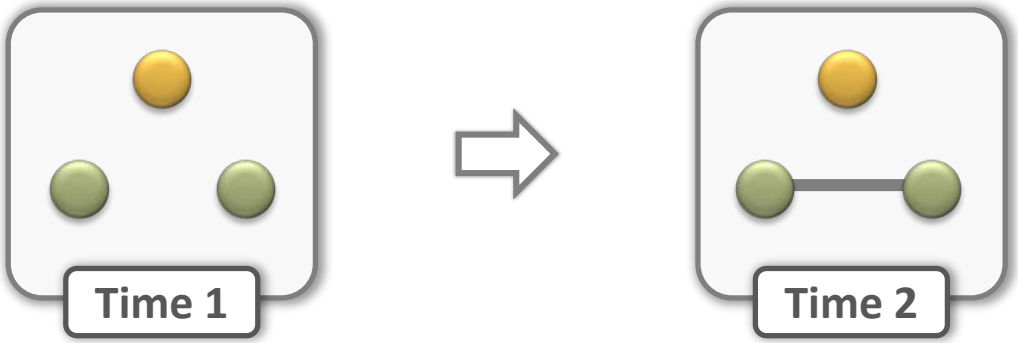
Examples

Dynamic networks and evolution over time

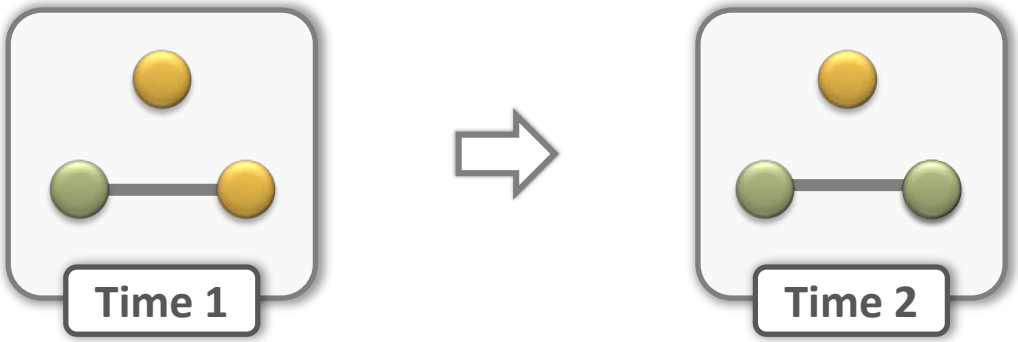


Interdependence of structure and attributes / behavior

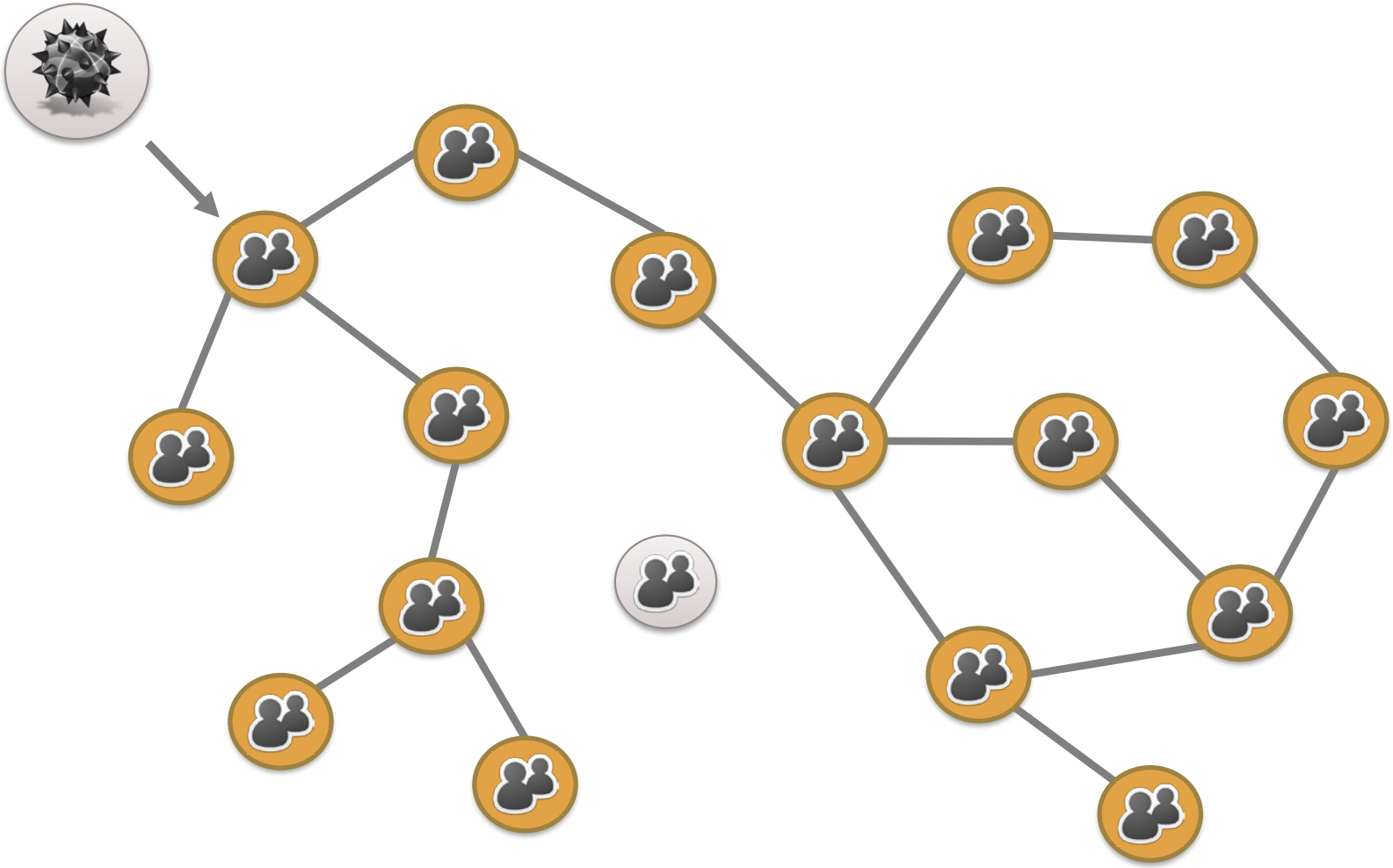
Selection: attributes or behavior affect structure



Influence: structure affects attributes or behavior



Diffusion and contagion in networks



Tools for network analysis & visualization

Try these if you like writing scripts

R programming environment
User interface: RStudio
Network packages:

- **Statnet** (includes the packages *network*, *sna*, *ergm, latentnet*, *degreenet*, *networksis)*
- **igraph** (available for R and Python)
- **RSiena** (modeling network evolution)

Python programming language
User interface: Anaconda, Spyder
Network packages:

- **NetworkX** (complex networks)
- **igraph** (available for R and Python)

Try these if you like clicking on buttons

NodeXL is a Microsoft Excel add-on for network visualization (Windows only).
It has importer plugins for network data collection from Facebook, Twitter, and Flickr.

Ucinet is a classic software for network analysis that has been around since the 1980s but is still regularly updated. It comes with a basic visualization tool called NetDraw.

Pajek is a software for large-scale network analysis and visualization.

PNet is a suite of tools for the simulation & estimation of network models.

Gephi is an advanced tool for graph visualization and manipulation.

My favorite network analysis platform:

RStudio

File Edit Code View Plots Session Project Build Tools Help

Console C:/Users/Katya/Desktop/DOCS/Projects/BG Protests DWM/

Workspace History

Import Dataset

Data	
net.1	10000x10000 double matrix
Values	
cred	oAuth[1]
dwm.bg	list[12]
first.day	1
last.day	12
month	"october"
sleepTime	10
tag	"%D0%94%D0%90%D0%9D%D0%A1withme"

-TwitteR.R × ESF-Data-Merge-Clean.R × IA-Congress-Scripts.R × CDF-WIP-R-Syntax.R × PEJ-DATA-2008-Weeks-Syntax-for-H1-H7.R ×

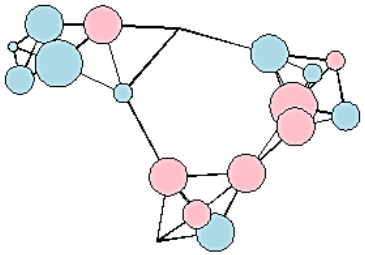
Source on Save Run Source

```
643
644
645 m3 <- ergm( pej.dich.net.nd ~ edges + mutual + gwesp + gwidegree(.5) + nodeocov('Frequency') )
646 m3; summary(m3);
647
648
649
650 m4 <- ergm(pej.dich.net.nd ~ edges + mutual + gwesp + gwidegree(.5) + nodematch("OwnerID", diff = FAL
651          + nodematch("Sector", diff = FALSE) + nodeocov('Frequency') + absdiff("Age") + absdiff("In
652          + absdiff("Education") + absdiff("Political") )
653 m4; summary(m4);
654
655
656
657 m5 <- ergm(pej.dich.net.nd ~ edges + receiver + nodematch("OwnerID", diff = FALSE) + nodeocov('Freque
658   absdiff("Age") + absdiff("Income") + absdiff("Education") + absdiff("Female") + absdiff("Political"
659 m5; summary(m5);
660
661
662 # triadcensus(c(8:9,12:15))
663 m6 <- ergm(pej.dich.net.nd ~ edges + triadcensus(c(8,12,15)) + nodematch("OwnerID", diff = FALSE) + n
664   absdiff("Age") + absdiff("Income") + absdiff("Education") + absdiff("Female") + absdiff("Political"
665 m6; summary(m6);
666
667
668 detach(package:ergm)
669
670
671
```

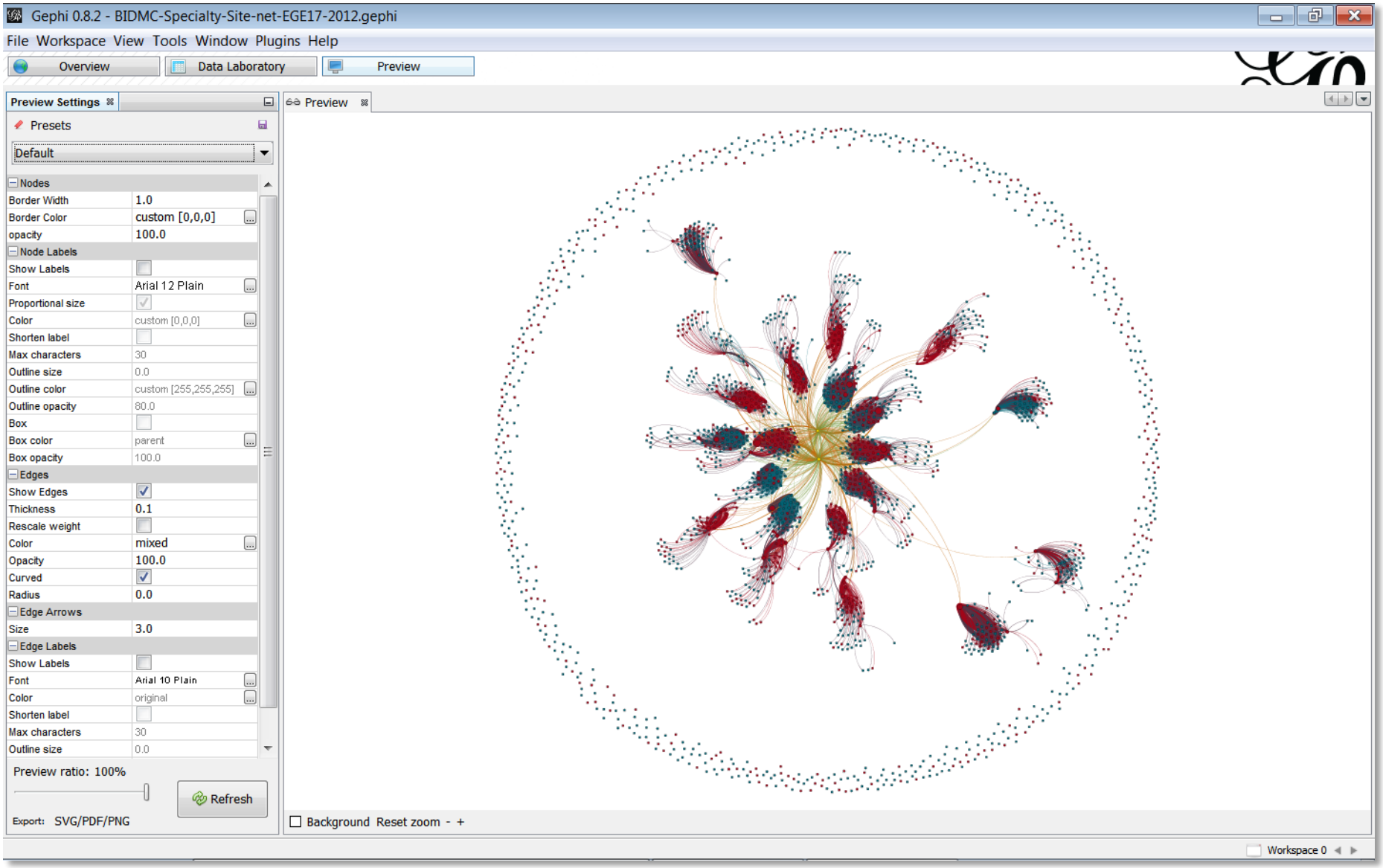
1:1 (Top Level) R Script

Files Plots Packages Help

Zoom Export Clear All

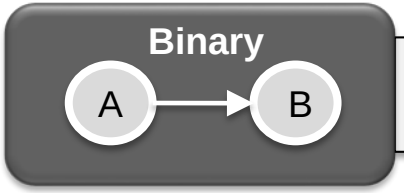


My favorite network visualization platform:

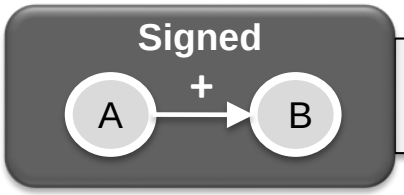


What do network formats look like?

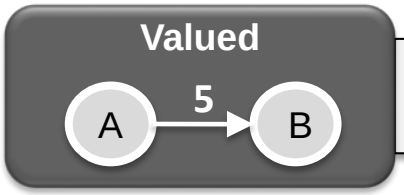
A few network types you will meet in the wild



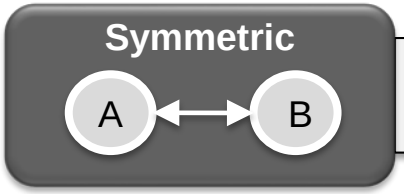
- **Link or no link** (1 or 0)



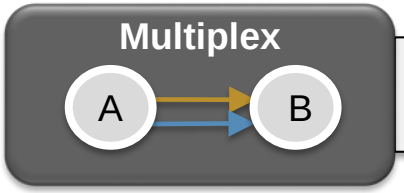
- **Positive or negative** (+, - or 0)



- **Weighted links** (each link is assigned a value)

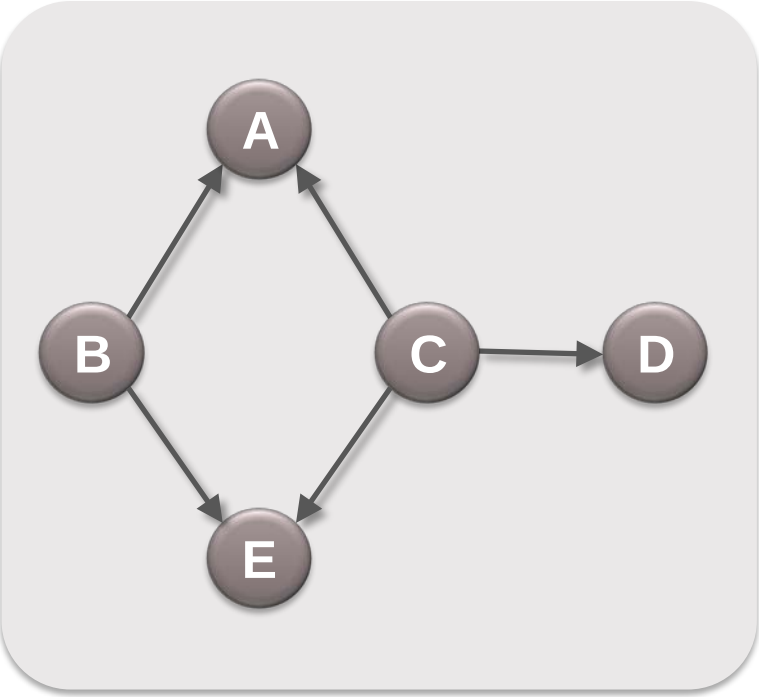


- **Directed vs. undirected** (undirected = symmetric)



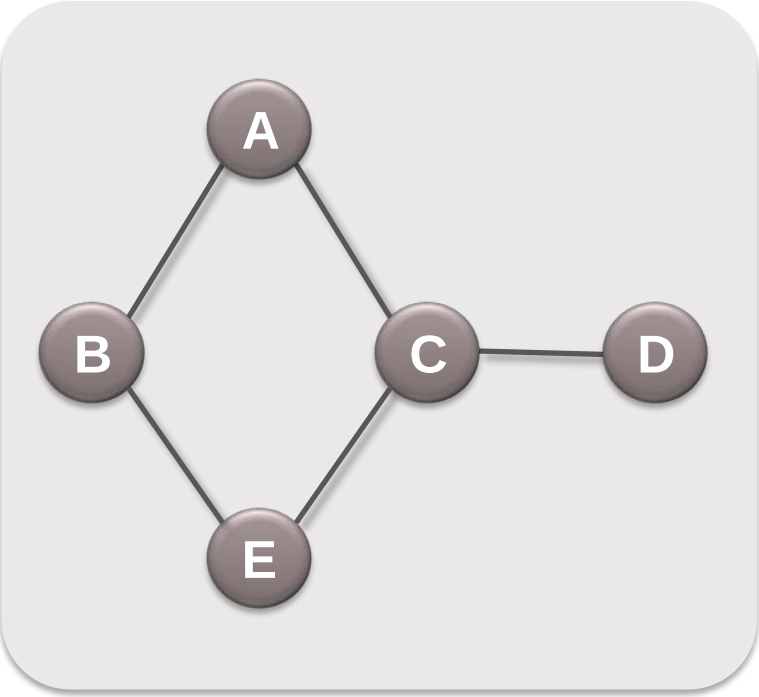
- **Multiplex** (more than one link type)

The matrix: directed networks



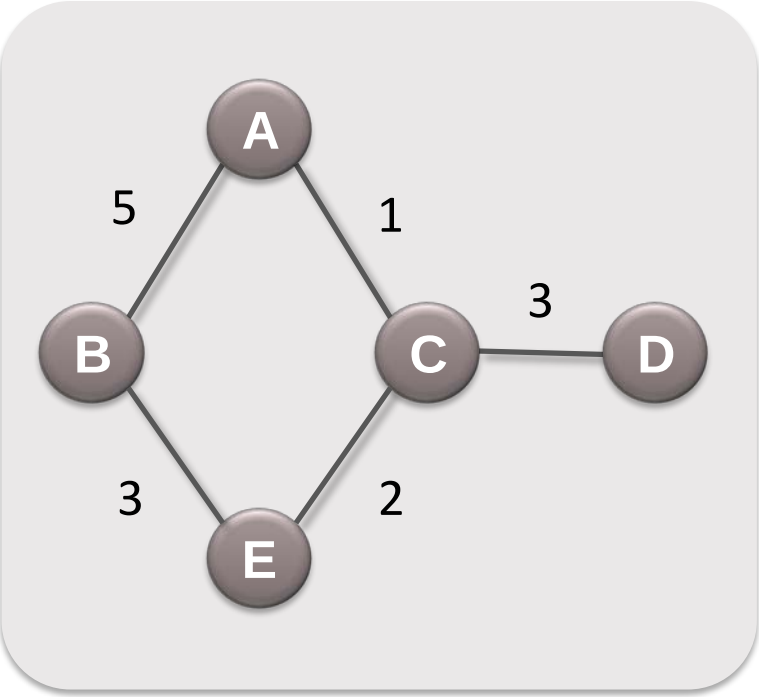
	A	B	C	D	E
A	0	0	0	0	0
B	1	0	0	0	1
C	1	0	0	1	1
D	0	0	0	0	0
E	0	0	0	0	0

The matrix reloaded: symmetric networks



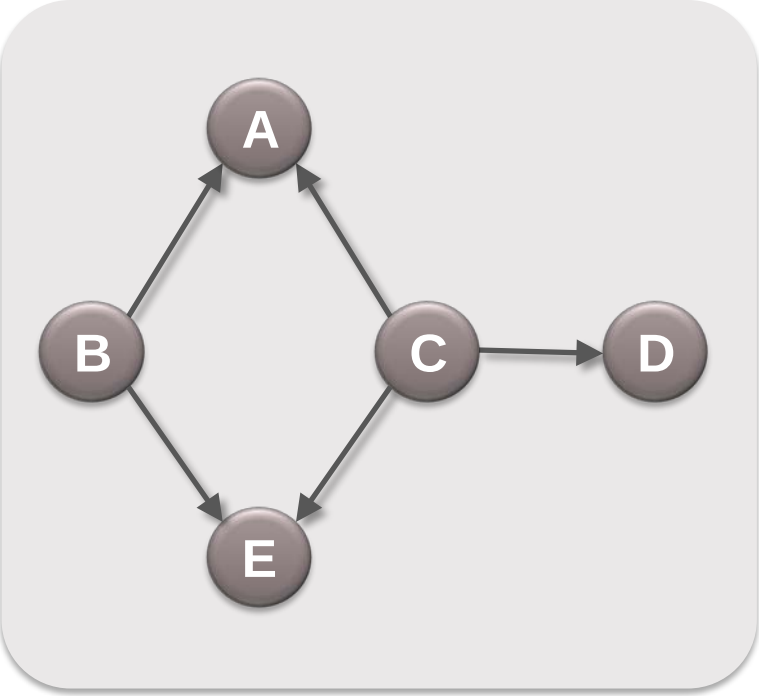
	A	B	C	D	E
A	0	1	1	0	0
B	1	0	0	0	1
C	1	0	0	1	1
D	0	0	1	0	0
E	0	1	1	0	0

The matrix revolutions: valued networks



	A	B	C	D	E
A	0	5	1	0	0
B	5	0	0	0	3
C	1	0	0	3	2
D	0	0	3	0	0
E	0	3	2	0	0

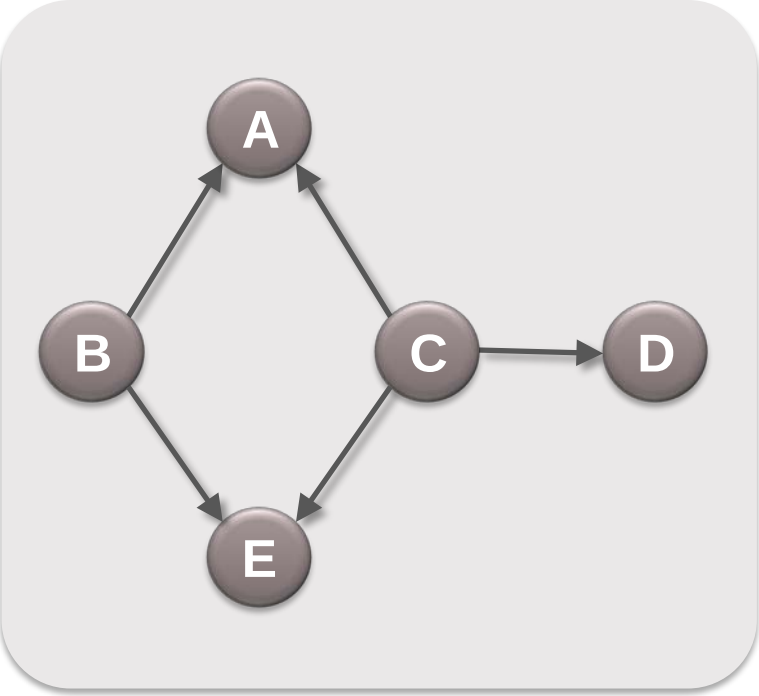
Save some bits: edgelist formats



Source	Destination	Weight
B	A	1
B	E	1
C	A	1
C	E	1
C	D	1

Note: Weights are optional.

Save some bits: nodelist formats



Source	Destinations
B	A E
C	A D E

Frequently used network file formats

- GEXF (Graph Exchange XML Format)
- GraphML, another XML format for graphs
- GML (Graph Modelling Language)
- DL, edgelist & matrix formats used by UCINET
- Net and Paj, Pajek network and project files

Coming up next:
Show-and-tell, dynamic network visualization.

Download this presentation & supporting files:
kateto.net/DataSwap

Contact Information:
Katherine Ognyanova
E-mail: katya@ognyanova.net
Website: www.kateto.net

DATA SWAP