Graph Structural Features

lecture-3

Course on Graph Neural Networks (Winter Term 21/22)

Prof. Dr. Holger Giese (holger.giese@hpi.uni-potsdam.de)

Christian Medeiros Adriano (christian.adriano@hpi.de) - **"Chris"**

Matthias Barkowski (matthias.barkowski@hpi.de)

The Strength of Weak Ties [Granovetter 73]

American Journal of Sociology, Volume 78, Issue 6 (May, 1973), 1360-1380.

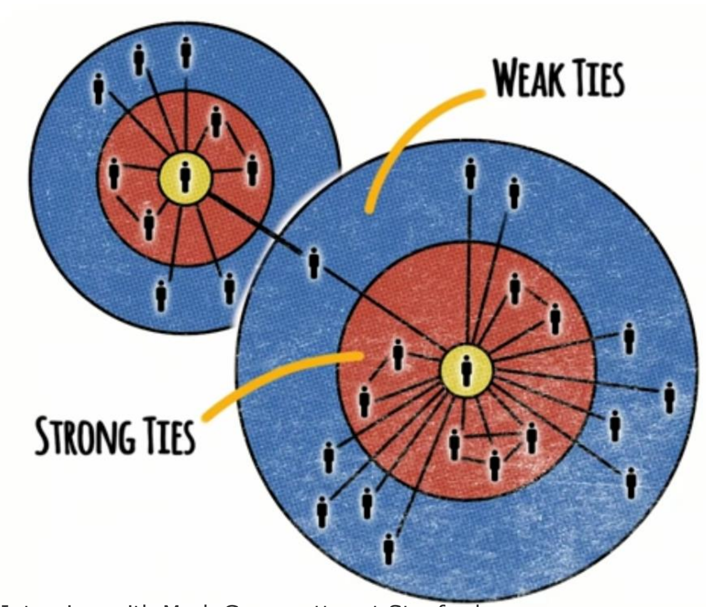
The Strength of Weak Ties[1]

Mark S. Granovetter
Johns Hopkins University

Analysis of social networks is suggested as a tool for linking micro and macro levels of sociological theory. The procedure is illustrated by elaboration of the macro implications of one aspect of small-scale interaction: the strength of dyadic ties. It is argued that the degree of overlap of two individuals' friendship networks varies directly with the strength of their tie to one another. The impact of this principle on diffusion of influence and information, mobility opportunity, and community organization is explored. Stress is laid on the cohesive power of weak ties. Most network models deal, implicitly, with strong ties, thus confining their applicability to small, well-defined groups. Emphasis on weak ties lends itself to discussion of relations *between* groups and to analysis of segments of social structure not easily defined in terms of primary groups.

A fundamental weakness of current sociological theory is that it does not relate micro-level interactions to macro-level patterns in any convincing way. Large-scale statistical, as well as qualitative, studies offer a good deal of insight into such macro phenomena as social mobility, community organization, and political structure. At the micro level, a large and increasing body of data and theory offers useful and illuminating ideas about what transpires within the confines of the small group. But how interaction in small groups aggregates to form large-scale patterns eludes us in most cases.

I will argue, in this paper, that the analysis of processes in interpersonal

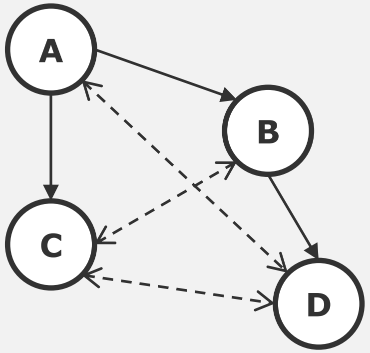


source: Interview with Mark Granovetter at Stanford - https://www.youtube.com/watch?v=g3bBajcR5fE&pbjreload=101

Interesting facts:

- Most people change jobs voluntarily
- 56% found new jobs through personal contacts
- 3/4 of the highest income categories
- More difficult to change if>5 years in a job
- People with 2 to 5 years in a job were more likely to find jobs through weak ties

Why?



Triads tend to form -> increasing the spread of information within a cluster

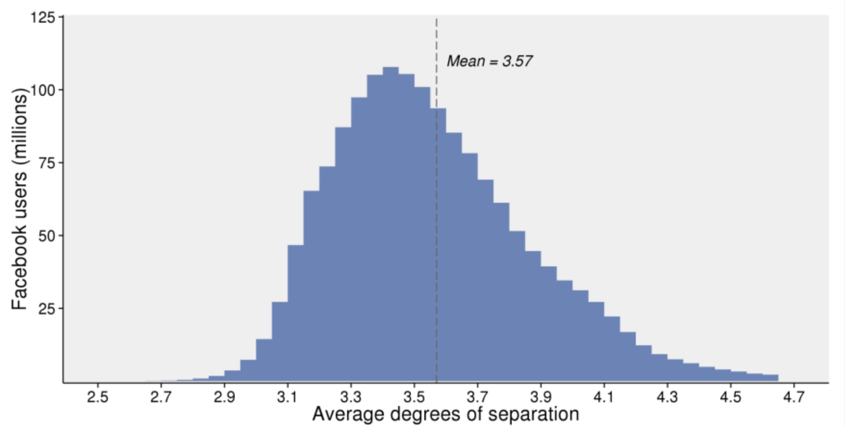
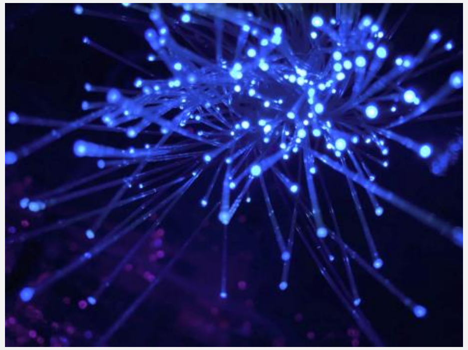


Figure 1. Estimated average degrees of separation between all people on Facebook.

Shrinking average shortest paths

https://research.fb.com/blog/2016/02/three-and-a-half-degrees-of-separation/



Instant global communication

JJ Ying https://unsplash.com/photos/8bghKxNU1j0

Lecture topics

1. Structure of Graph
   - Motifs – the building blocks
   - Graphlets – the landscape
   - Metrics
     - Centrality metrics
     - Structural membership metrics
     - Homophily metrics (Assortativity and Neighborhood)
2. Clustering
   - Community detection
   - Spectral clustering

Network metrics

IT Systems Engineering | Universität Potsdam

Network-level metrics

- Average node degree (degree distribution)
- Average clustering
- Average path length (diameter)
- Node connectivity (distribution among components)

Node-level metrics

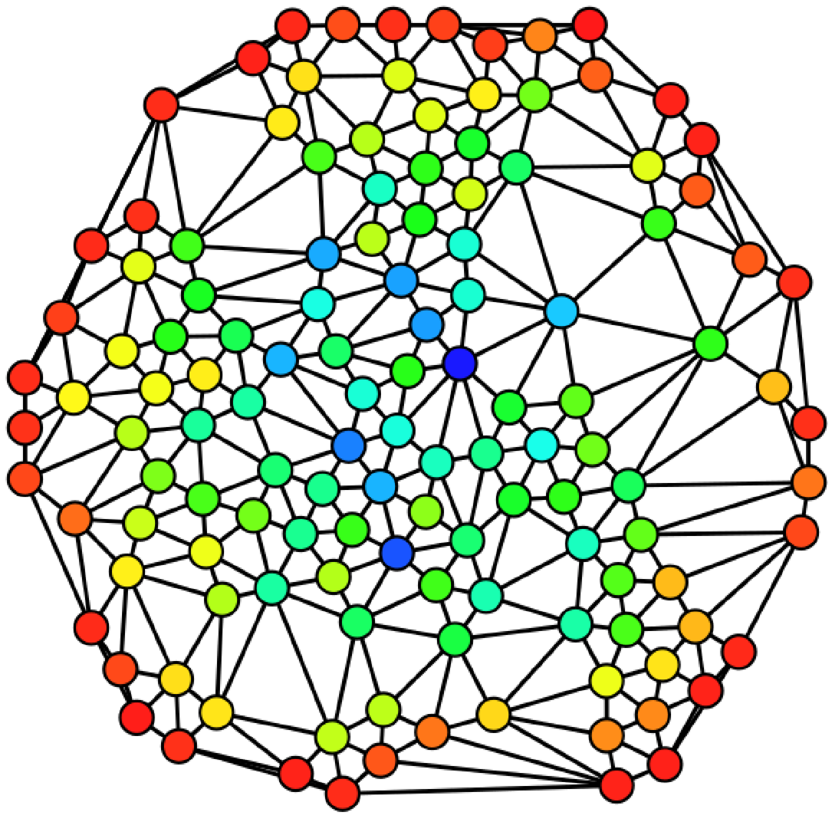
- Centrality (degree-based): Katz, PageRank
- Centrality (path-based): closeness and betweenness

Pair-wise metrics

- Edge-overlap
- Node equivalence (Regular and Structural)
- Node similarity

Graph theoretic measures

- Eccentricity $e(n_i)$: the maximum distance between nodes n_i and any other node
- Diameter (d): largest $e(n_i)$ for across all n_i
- Radius (r): smallest $e(n_i)$ for across all n_i
  - Central node: $e(n_i) = r$
  - Graph center: $\forall n_i |\ e(n_i) = r$
  - Graph periphery: $\forall n_i |\ e(n_i) = d$



source: Wikipedia

Centrality Metrics – Degree-based

Centrality: node degree
Eigenvector Centrality: importance of a node (x_i) depends on importance (x_j) of its neighbors

Centrality: node degree

source: Wikipedia

Eigenvector Centrality: importance of a node (x_i) depends on importance (x_j) of its neighbors

$$x_i = k^{-1} \sum_j x_j$$

Or use the Adjacency matrix A: $x_i = k^{-1} \sum_j A_{ij} x_j,\ x = k^{-1} A x,\ Ax = kx$

$x = \text{eigenvector of } A$

k = from the *ith* eigenvalue of the leading eigenvector x ,

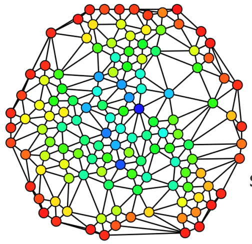
$\boldsymbol{k = \frac{1}{\lambda_2}}$, where λ_2 is be the largest non-negative eigenvalue (From the Perron-Frobenius Theorem).

Katz Centrality: $x_i = k^{-1} \sum_j A_{ij} x_j + \beta$,where $\beta = constant$

PageRank: $x_i = k^{-1} \sum_j A_{ij} \frac{x_j}{k_j^{out}} + \beta$,where $\mathrm{k_j^{out}} = \text{outdegree of node j}$

Newman, Mark. *Networks*. Oxford university press, 2018.
Borgatti, Stephen P., and Martin G. Everett. "A graph-theoretic perspective on centrality." *Social networks* 28.4 (2006): 466-484.

Centrality Metrics – Path-Based



source: Wikipedia

Closeness-Centrality:

$x_i = \frac{1}{\sum_j d(i,j)}$ or take the harmonic mean $\sum_j \frac{1}{d(i,j)}$ (solves $d = \infty$)

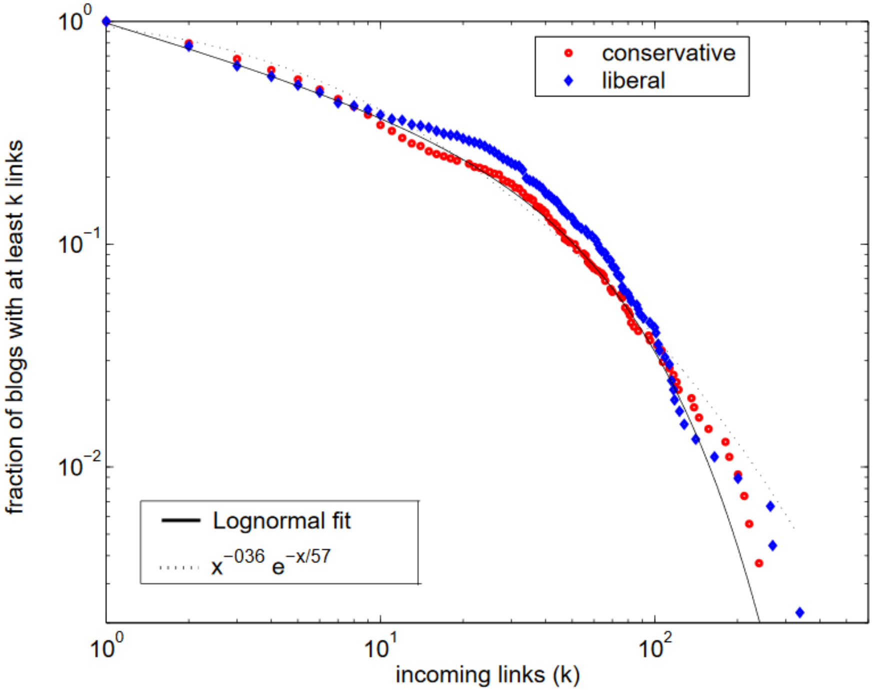
Betweenness-Centrality:

$x_i = \sum_{s \to t, s \neq t} n_{st}^i$, where n_{st}^i is the shortest path between *s* and *t* that goes through node *i* (n^i).

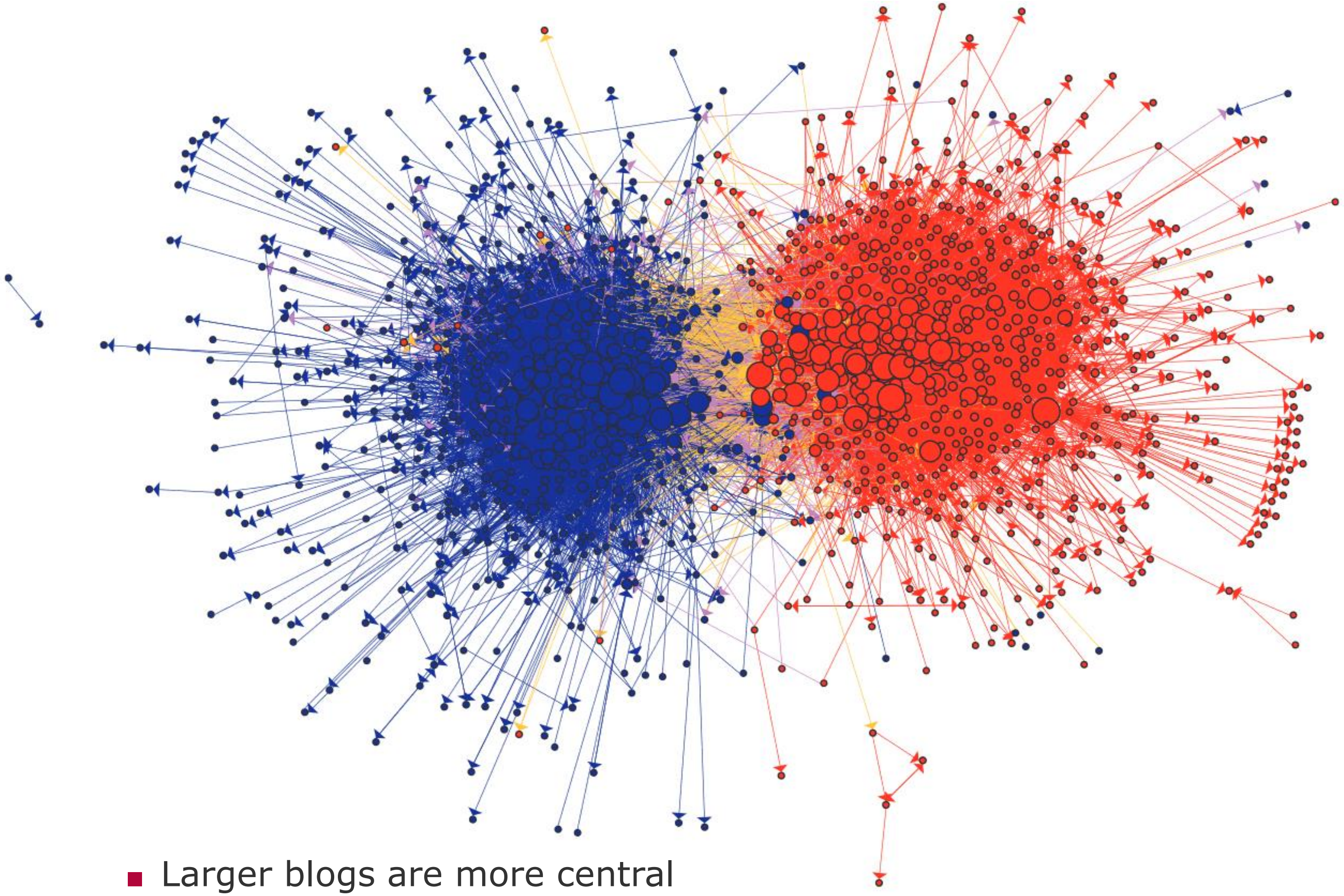
Newman, Mark. *Networks*. Oxford university press, 2018.
Borgatti, Stephen P., and Martin G. Everett. "A graph-theoretic perspective on centrality." *Social networks* 28.4 (2006): 466-484.

Political Blogosphere [Adamic & Glance 2005]

Cumulative distribution of incoming links for political blogs by category



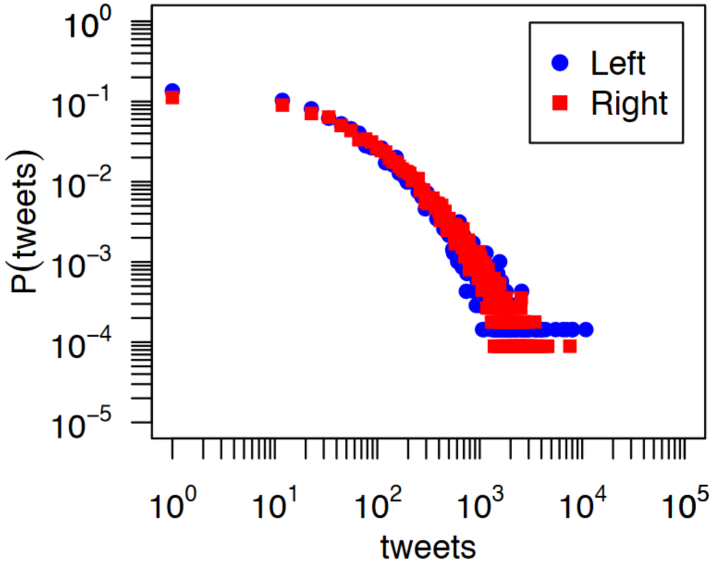
Similar distributions can be fit with lognormal or exponential functions.



- Larger blogs are more central
- Homophily
- Yellow links show blogs from opposite views commenting on each other

Source: Adamic, Lada A., and Natalie Glance. "The political blogosphere and the 2004 US election: divided they blog." *Proceedings of the 3rd international workshop on Link discovery*. 2005.

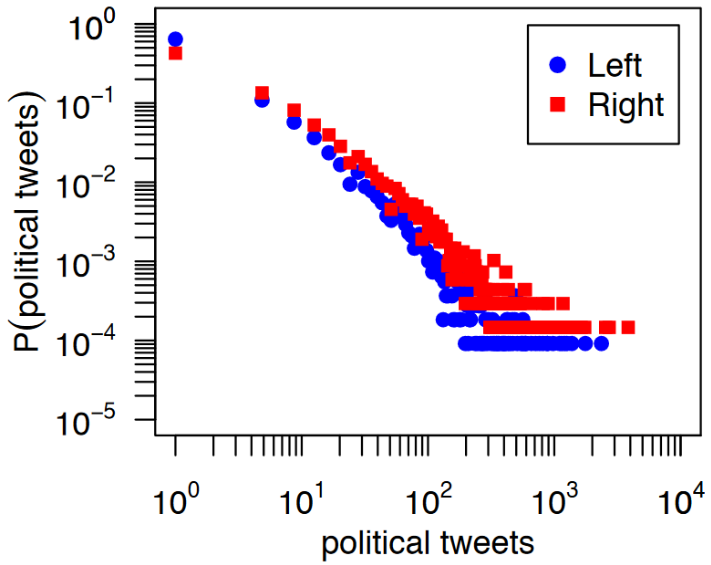
Partisan Asymmetries in Twitter [Conover et al. 2012]



Both groups produce same amount of tweets per capita

tweets = number of tweets a user posts

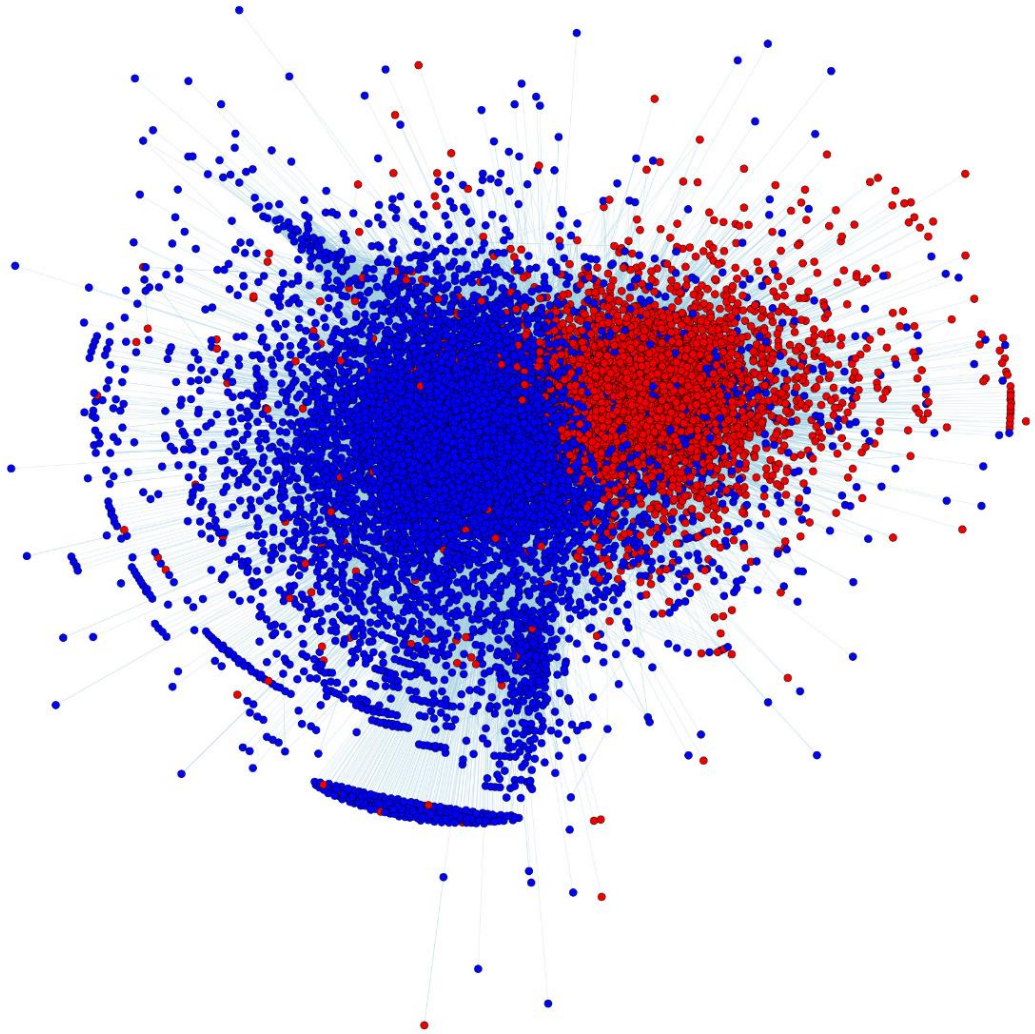
P(tweets) probability of a user



Right-leaning produce 2x more political content

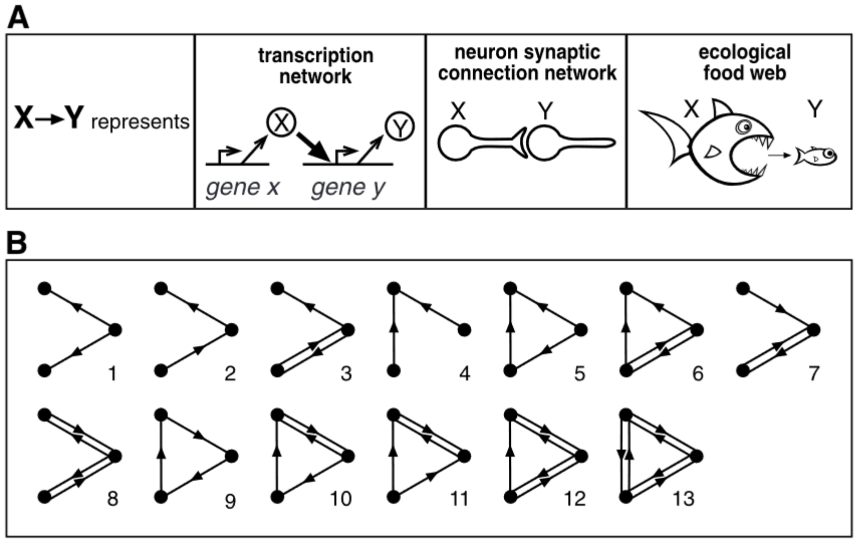
One or more political hash tags

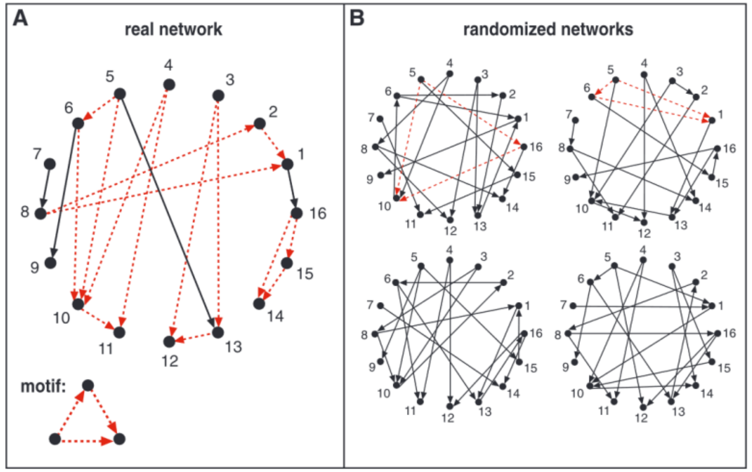
- Right-leaning= 22% of tweets
- Left-leaning= 12% of tweets



source: Conover, Michael D., et al. "Partisan asymmetries in online political activity." *EPJ Data Science* 1.1 (2012): 6.

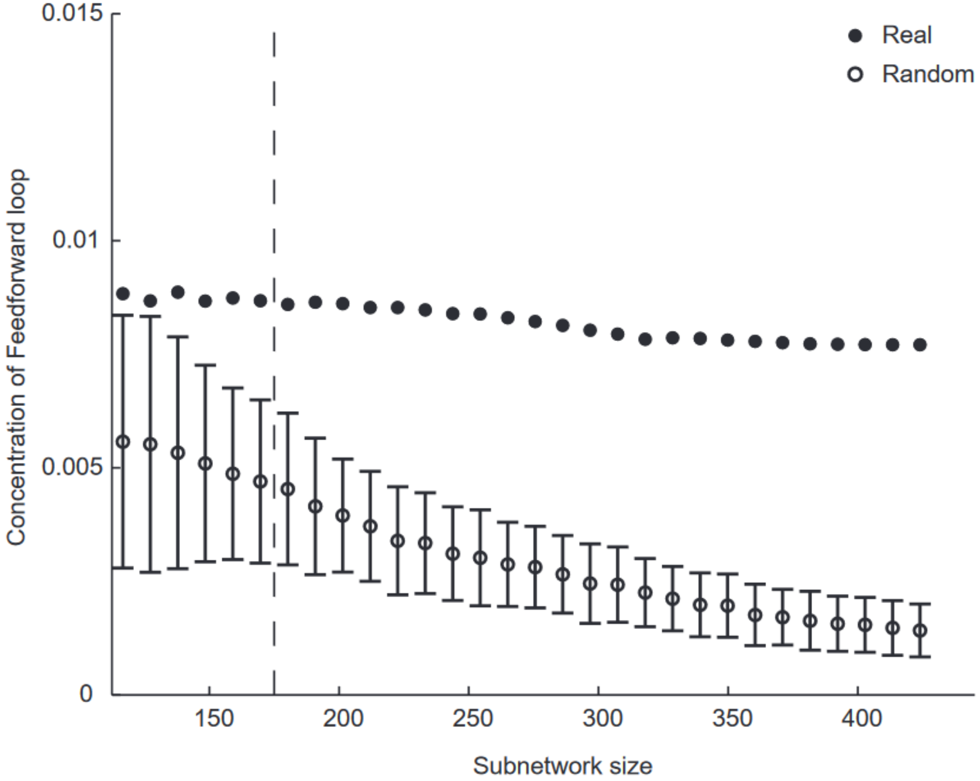
Motifs [Milo 2002]





Remarks

- Number of motifs grow exponentially with the number of edges
- Real networks tend to be rich in motifs, whereas Random networks not.
- As Real networks evolve, their motif count remains stable, whereas in Random networks, it colapses.
- This might suggest that motifs provide struture to the proper functioning of these networks



[Milo 2002] Milo, R. *et al*. Network motifs: Simple building blocks of complex networks. *Science* **298**, 824–827 (2002).

Motifs in real networks

Biological

Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
Gene regulation (transcription)					**Feed-forward loop**			**Bi-fan**			
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13			
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41			
Neurons					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
C. elegans†	252	509	125	90 ± 10	3.7	127	55 ± 13	5.3	227	35 ± 10	20
Food webs					**Three chain**			**Bi-parallel**			
Little Rock	92	984	3219	3120 ± 50	2.1	7295	2220 ± 210	25			
Ythan	83	391	1182	1020 ± 20	7.2	1357	230 ± 50	23			
St. Martin	42	205	469	450 ± 10	NS	382	130 ± 20	12			
Chesapeake	31	67	80	82 ± 4	NS	26	5 ± 2	8			
Coachella	29	243	279	235 ± 12	3.6	181	80 ± 20	5			
Skipwith	25	189	184	150 ± 7	5.5	397	80 ± 25	13			
B. Brook	25	104	181	130 ± 7	7.4	267	30 ± 7	32			

Technological

Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
Electronic circuits (forward logic chips)					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
s15850	10,383	14,240	424	2 ± 2	285	1040	1 ± 1	1200	480	2 ± 1	335
s38584	20,717	34,204	413	10 ± 3	120	1739	6 ± 2	800	711	9 ± 2	320
s38417	23,843	33,661	612	3 ± 2	400	2404	1 ± 1	2550	531	2 ± 2	340
s9234	5,844	8,197	211	2 ± 1	140	754	1 ± 1	1050	209	1 ± 1	200
s13207	8,651	11,831	403	2 ± 1	225	4445	1 ± 1	4950	264	2 ± 1	200
Electronic circuits (digital fractional multipliers)					**Three-node feedback loop**			**Bi-fan**			**Four-node feedback loop**
s208	122	189	10	1 ± 1	9	4	1 ± 1	3.8	5	1 ± 1	5
s420	252	399	20	1 ± 1	18	10	1 ± 1	10	11	1 ± 1	11
s838‡	512	819	40	1 ± 1	38	22	1 ± 1	20	23	1 ± 1	25
World Wide Web					**Feedback with two mutual dyads**			**Fully connected triad**			**Uplinked mutual dyad**
nd.edu§	325,729	1.46e6	1.1e5	2e3 ± 1e2	800	6.8e6	5e4±4e2	15,000	1.2e6	1e4 ± 2e2	5000

Milo, R. *et al*. Network motifs: Simple building blocks of complex networks. *Science* **298**, 824–827 (2002).

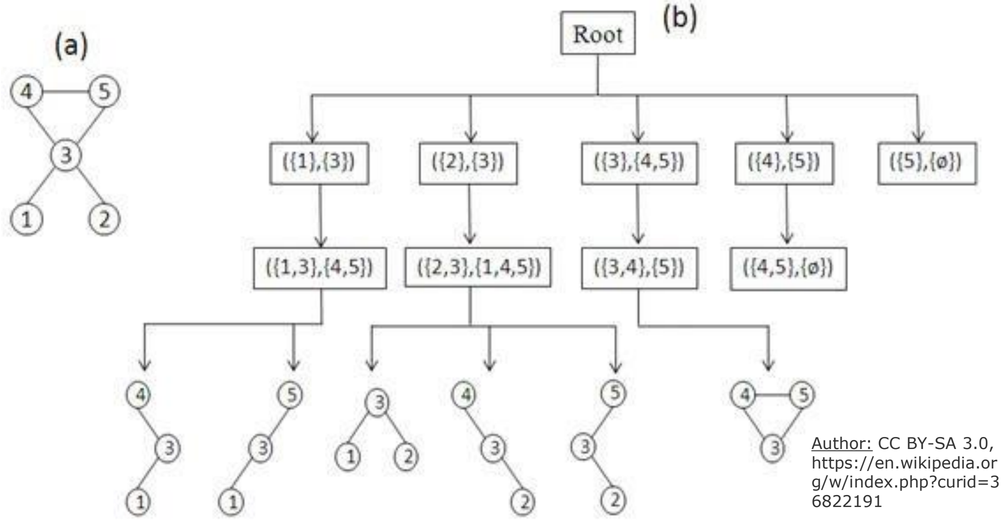
Motif counting – ESU (FANMOD) algorithm [Wernicke 2006]

Algorithm: ENUMERATESUBGRAPHS(G, k) (ESU)
Input: A graph $G = (V, E)$ and an integer $1 \le k \le |V|$.
Output: All size-k subgraphs in G.

for each vertex $v \in V$ **do**
  $V_{Extension} \leftarrow \{u \in N(\{v\}) : u > v\}$
  call EXTENDSUBGRAPH$(\{v\}, V_{Extension}, v)$
return

EXTENDSUBGRAPH$(V_{Subgraph}, V_{Extension}, v)$
if $|V_{Subgraph}| = k$ **then output** $G[V_{Subgraph}]$ **and return**
while $V_{Extension} \neq \emptyset$ **do**
  Remove an arbitrarily chosen vertex w from $V_{Extension}$
  $V'_{Extension} \leftarrow V_{Extension} \cup \{u \in N_{excl}(w, V_{Subgraph}) : u > v\}$
  call EXTENDSUBGRAPH$(V_{Subgraph} \cup \{w\}, V'_{Extension}, v)$
return

[Wernicke 2006] Wernicke S (2006). "Efficient detection of network motifs". *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. **3** (4): 347–359.



EXT set of all the nodes that :

1. are adjacent to at least one of the nodes in SUB (guarantees expansion is a connected graph)
2. their numerical labels must be larger than the label of first element in SUB. (guarantees termination)

Many other algorithms have bee developed and tested

Networks ↓	Size → / Algorithms ↓	3	4	5	6	7	8	9	10
E. coli	Kavosh	0.30	1.84	14.91	141.98	1374.0	13173.7	121110.3	1120560.1
	FANMOD	0.81	2.53	15.71	132.24	1205.9	9256.6	-	-
	Mavisto	13532	-	-	-	-	-	-	-
	Mfinder	31.0	297	23671	-	-	-	-	-
Electronic	Kavosh	0.08	0.36	8.02	11.39	77.22	422.6	2823.7	18037.5
	FANMOD	0.53	1.06	4.34	24.24	160	967.99	-	-
	Mavisto	210.0	1727	-	-	-	-	-	-
	Mfinder	7	14	109.8	2020.2	-	-	-	-
Social	Kavosh	0.04	0.23	1.63	10.48	69.43	415.66	2594.19	14611.23
	FANMOD	0.46	0.84	3.07	17.63	117.43	845.93	-	-
	Mavisto	393	1492	-	-	-	-	-	-
	Mfinder	12	49	798	181077	-	-	-	-

https://en.wikipedia.org/wiki/Network_motif#Motif_discovery_algorithms

Motif caveats

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Caveats

- Stability of a network structure not necessary imply function stability [Igram et al. 2006]
- Network structure not always imply function [Voigt et al. 2005]
- Static analysis might hide the relationship between structure and function
  - Temporal network motifs have been investigated [Braha & Bar-Yam 2009][Holme & Saramäki 2012]

[Braha & Bar-Yam 2009] Braha D., Bar-Yam Y. (2009). “Time-Dependent Complex Networks: Dynamic Centrality, Dynamic Motifs, and Cycles of Social Interactions.” In: Gross T., Sayama H. (eds) *Adaptive Networks. Understanding Complex Systems*. Springer, Berlin, Heidelberg
[Holme & Saramäki 2012] Holme, P., & Saramäki, J. (2012). Temporal networks. *Physics reports*, *519*(3), 97-125.
[Ingram et al. 2006] Ingram PJ, Stumpf MP, Stark J (2006). "Network motifs: structure does not determine function". *BMC Genomics.* 7: 108.
[Voigt et al. 2005] Voigt CA, Wolf DM, Arkin AP (2005). "The Bacillus subtilis sin operon: an evolvable network motif". *Genetics.* 169 (3): 1187–202.

Motif-based spectral clustering

Goal: find clusters of motifs,

How: find clusters that minimize the number of motif that are cut

Definitions

Motif-cuts: number of motifs whose at least one edge is cut by the clustering

Motif-volume: number of nodes participating in a certain motif type

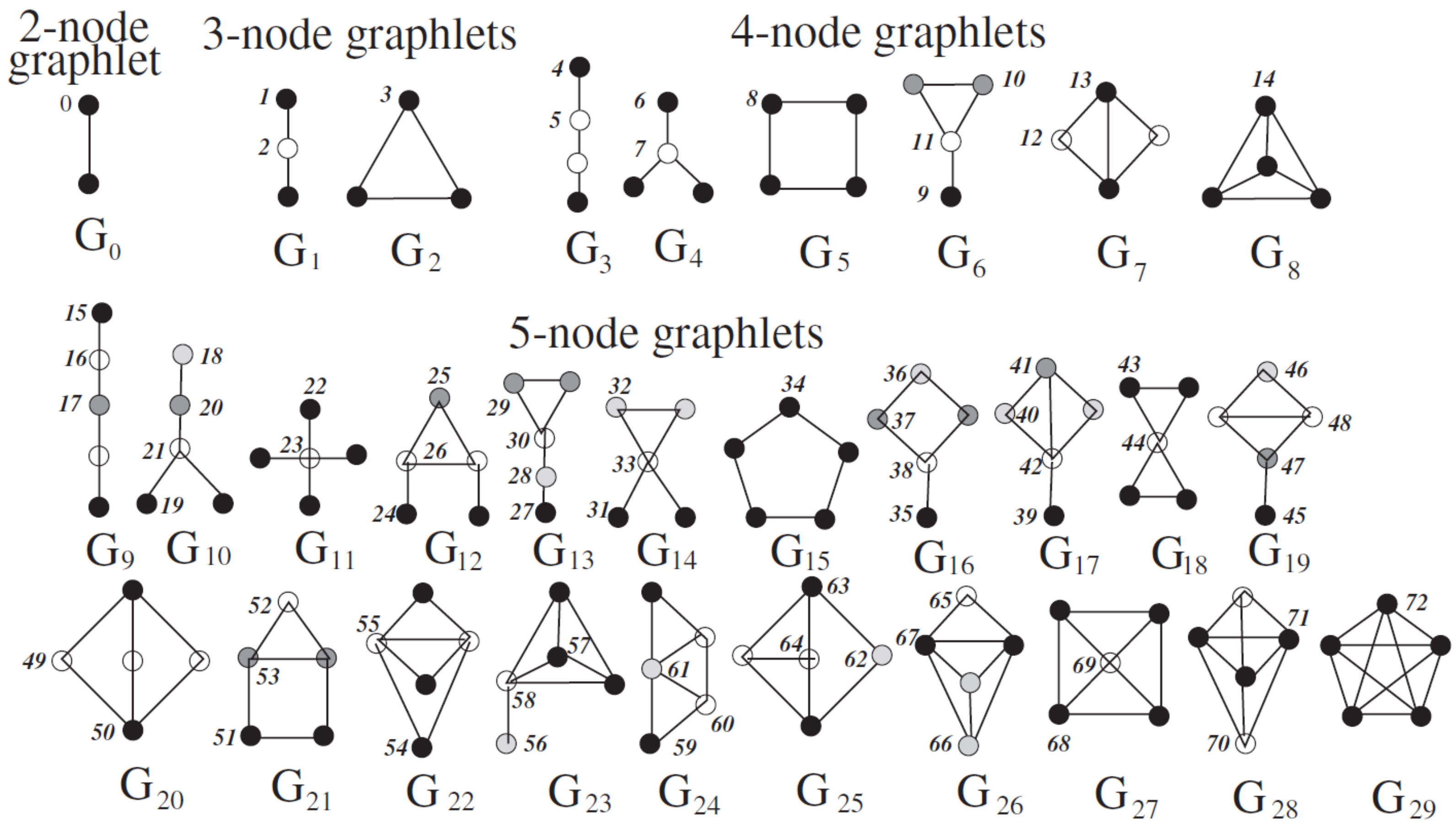
Motif Conductance: $\phi_M(S) = \frac{MotifCuts}{MotifVolume}$, where S is the set of nodes that minizes conductance.

This is an NP-Hard problem... One approximation is the **Motif-based Spectral Clustering**:

- Step-1 Preprocessing – create a weighted graph *W* by counting the number of times a motif edge appears
- Step-2 Decomposition – use standard special clustering on the Weighted graph
- Step-3 Grouping – same as spectral clustering

Lu, Zhenqi, Johan Wahlström, and Arye Nehorai. "Community detection in complex networks via clique conductance." *Scientific reports* 8.1 (2018): 1-16.

Graphlets



Automorphism orbits 0,1,2, . . . , 72 for the thirty 2, 3, 4, and 5-node graphlets G_0,G_1, . . . ,G_{29}. In a graphlet G_i, i 2 {0,1,. . .29}, nodes belonging to
the same orbit are of the same shade

Modularity (Q) – quality of clustering

Modularity (Q) fraction of the edges within the given groups minus the expected such fraction if edges were distributed at random. It is defined as [Newman 2004]:

$$Q = \frac{1}{2m} \sum_{i,j} \left(\boxed{A_{i,j}} - \boxed{\frac{k_i k_j}{2m}} \right) \delta(c_i, c_j),$$

Actual connection between *i* and *j*

If nodes were randomly connected (null model)

<u>where:</u>

m = total number of edges

$A_{i,j}$ = weight of the edge between i and j

$k_i = \sum_j A_{i,j}$ is sum of the weighted edges attached to node i

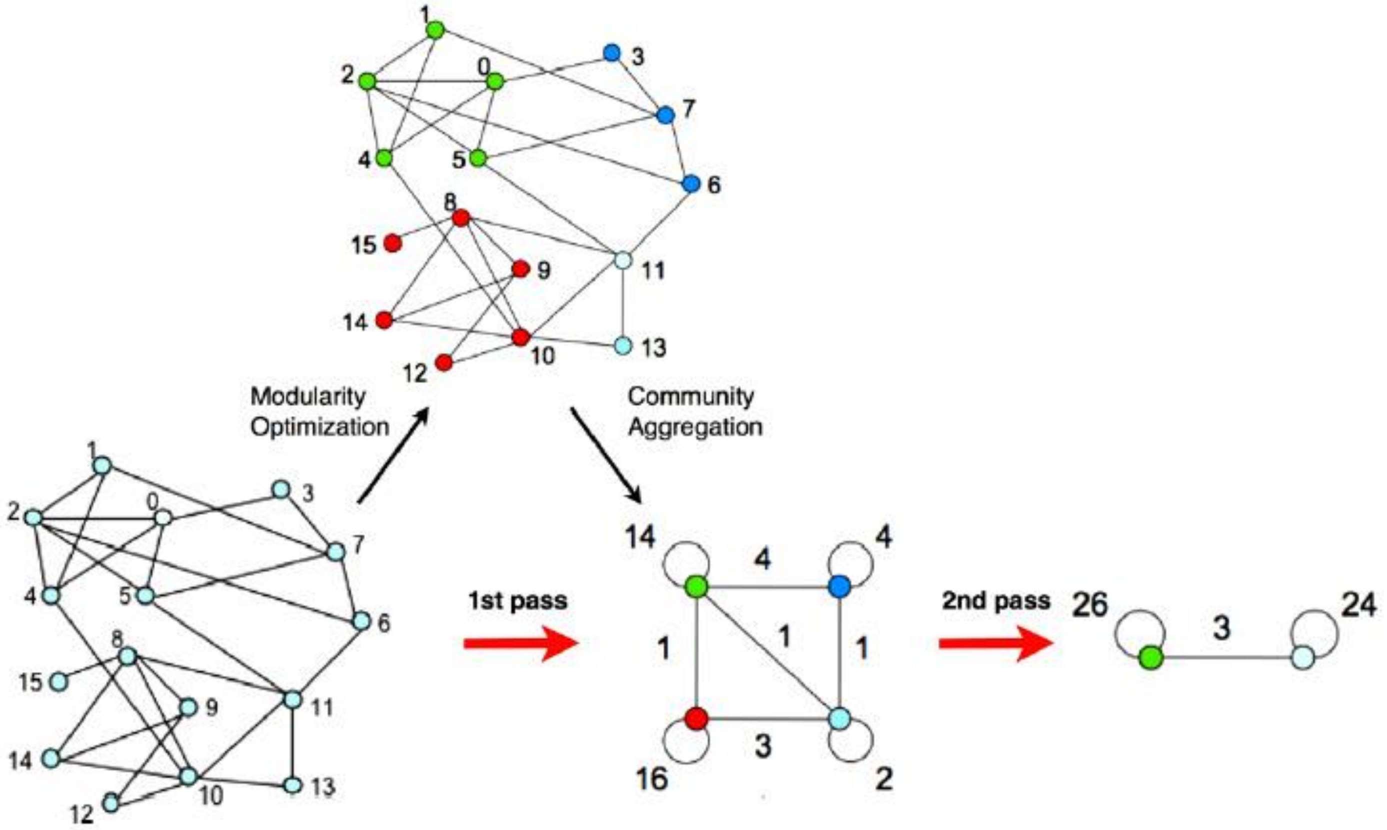
c_i = community of node i

$\delta(c_i, c_j) = 1 \; if \; i \; and \; j \in same \; community \; c$

[Newman 2004] Newman, Mark EJ. "Analysis of weighted networks." *Physical review E* 70.5 (2004): 056131.
[Blondel et al. 2008] Blondel, Vincent D., et al. (2008), "Fast unfolding of communities in large networks.", *Journal of statistical mechanics: theory and experiment, 2008(9), pp.1-13.*

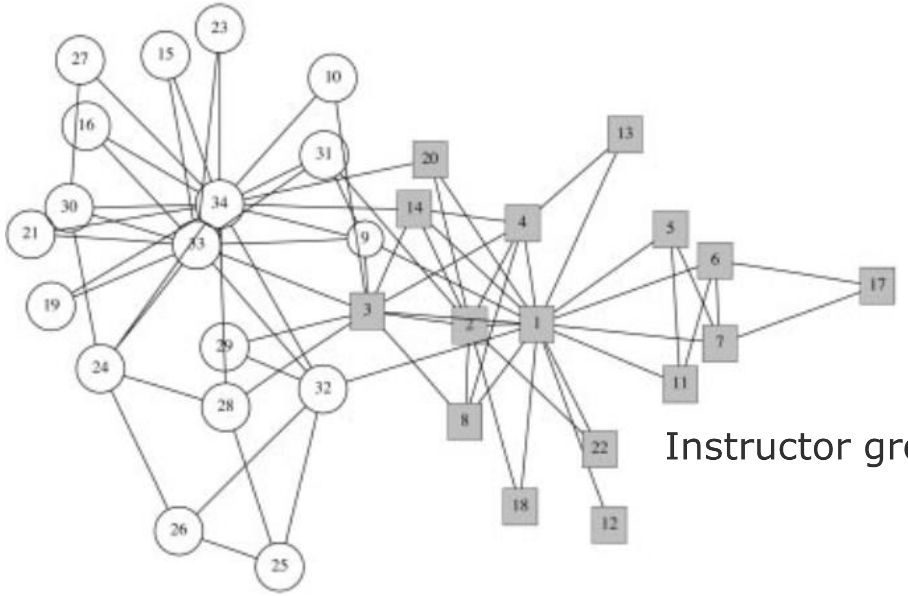
Louvain Algorithm

Iterate until find the lowest modularity value

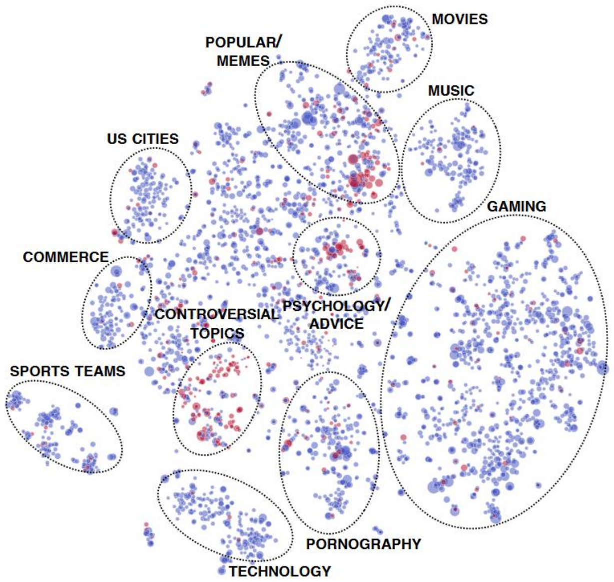


Community structure

Zachary's Karate Club Split
[Girvan et al. 2002]



Reddit Communities
[Kumar, et al. 2018]



[Girvan et al. 2002] Girvan, Michelle, and Mark EJ Newman. "Community structure in social and biological networks." *Proceedings of the national academy of sciences* 99.12 (2002): 7821-7826.
[Kumar et al. 2018] Kumar, S., Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2018, April). Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference* (pp. 933-943).

Spectral Clustering

1. Build Graph Laplacian representation:

$$Laplacian\ matrix\ (L) = \ Adjacency\ matrix\ (A) - \ Degree\ matrix\ (D)$$

2. Reduce Dimensionality:

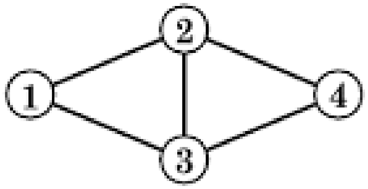
Compute eigenvalues and eigenvectors of the L matrix

Use eigenvectors to map nodes to a reduced representation

3. Categorize:

Use the new representation, assign nodes to clusters

$$D = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$



$$A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \qquad L = \begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix}$$

L = A-D

Algorithm: Spectral graph partitioning - normalized cuts

Input: adjacency matrix $\mathbf{A}$

Output: class indicator vector $\mathbf{v}$

compute $\mathbf{D} = diag(deg(\mathbf{A}))$;

compute $\mathbf{L} = \mathbf{D} - \mathbf{A}$;

solve for second smallest eigenvector:

min cut: $\mathbf{Lx} = \lambda\mathbf{x}$;

normalized cut : $\mathbf{Lx} = \lambda\mathbf{Dx}$;

set $\mathbf{s} = sign(\mathbf{x}_2)$

Further resources

Short Tutorial on Graph Laplacian:
https://csustan.csustan.edu/~tom/Clustering/GraphLaplacian-tutorial.pdf

Lecture Slides: http://graphics.stanford.edu/courses/cs468-10-fall/LectureSlides/16_spectral_methods1.pdf

Tutorial on Spectral Clustering:
nhttps://www.cs.cmu.edu/~aarti/Class/10701/readings/Luxburg06_TR.pdf

Homophily metrics

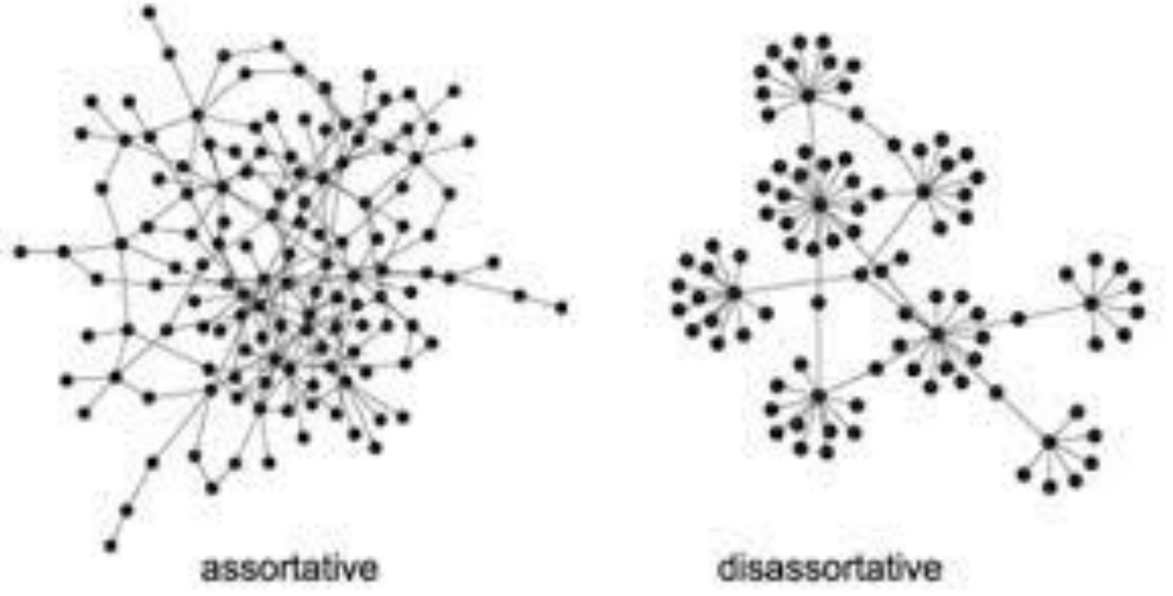
Definition: Preference to attach to nodes that are similar [Newman 2003]

Assortativity Coefficient (Pearson Correlation)

$$r(\alpha,\beta) = \frac{\sum_i (j_i^\alpha - \bar{j^\alpha})(k_i^\beta - \bar{k^\beta})}{\sqrt{\sum_i (j_i^\alpha - \bar{j^\alpha})^2}\sqrt{\sum_i (k_i^\beta - \bar{k^\beta})^2}}.$$

Neighborhood Connectivity

$$\langle k_{nn} \rangle = \sum_{k'} k' P(k'|k), \text{ where } P(k'|k)$$



[Newman 2003] *Newman, M. E. J. (2003). "Mixing patterns in networks". Physical Review E. American Physical Society (APS).* ***67*** *(2):*

Node equivalence

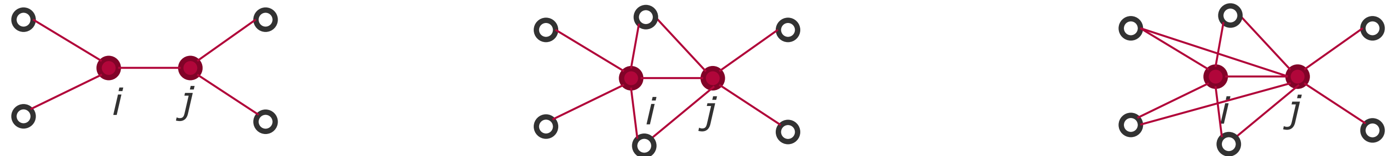
Structural equivalence

- Jaccard Similarity (intersection)
- Cosine Similarity (distance)
- Pearson Correlation
- Euclidean Distance (dissimilarity)
- Node overlap

$$j_{ij} = \frac{n_{ij}}{k_i + k_j - n_{ij}}$$

$$\sigma_{ij} = \frac{n_{ij}}{\sqrt{k_i k_j}_i}$$

$$O_{i,j} = \frac{|(N_i \cap N_j) \setminus \{i,j\}|}{|(N_i \cup N_j) \setminus \{i,j\}|}$$



$O_{ij} = 0$ $\quad$ $O_{ij} = 1/3$ $\quad$ $O_{ij} = 2/3$

Regular equivalence

Nodes that do not necessarily share neighbors, but have similar features

Next Tasks

Please enroll (if you did not yet).

Initial understanding about your graph data

1. What are possible the edges and nodes of your network?
2. What types of networks w.r.t. edge directionality?
3. What type of networks w.r.t. to node types?
4. Is your network possibly assortative, dissortative, or neither?
5. Which centrality metrics might make sense for you?
6. Are there communities/clusters that can be used as features?
7. Do you believe your graph will present high or low cluster coefficient?
8. What about the network diameter (low or high)?

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

END