JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Network Metrics

3/06/24

Recap

- Last class:
  - Causal Inference with text

- Reminders:
  - HW 3 due (next) Friday
  - Midterm in 1 week

Outline

- Introduction and definitions
- Basic Network Metrics
- Advanced Network Methods
- Graph Neural Network

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Introduction and Definitions

Motivation: understand relationship

- High School Partnership Network

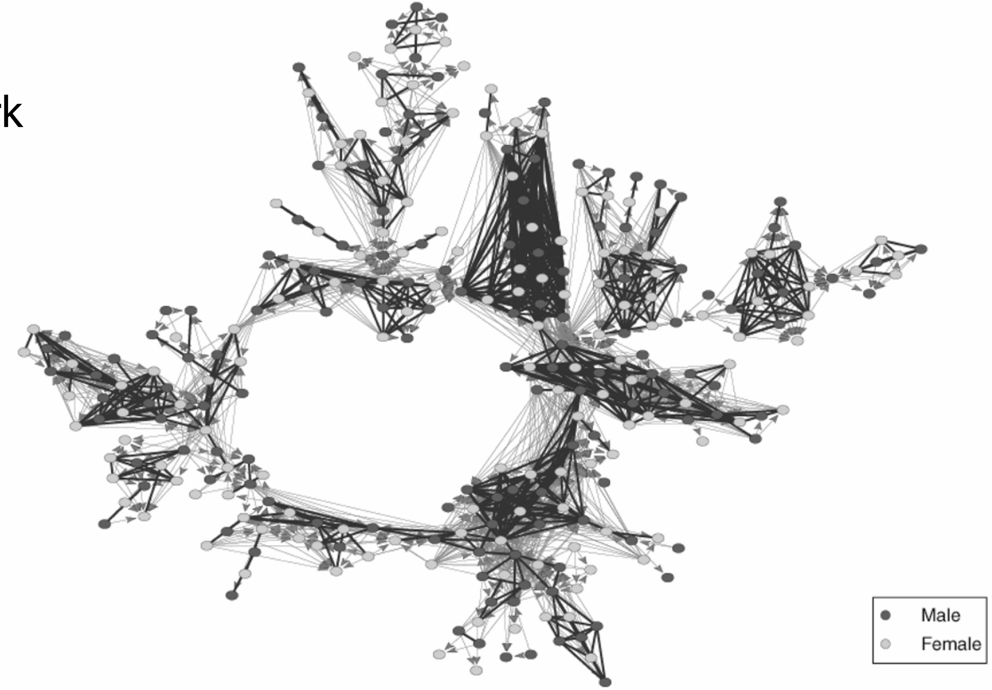


FIG. 3.—Temporally ordered ties in the Jefferson High partnership network

Bearman, P. S., Moody, J., & Stovel, K. (2004). Chains of affection: The structure of adolescent romantic and sexual networks. American journal of sociology, 110(1), 44-91.

Motivation: understand epidemic

- Sex Partner Network and HIV

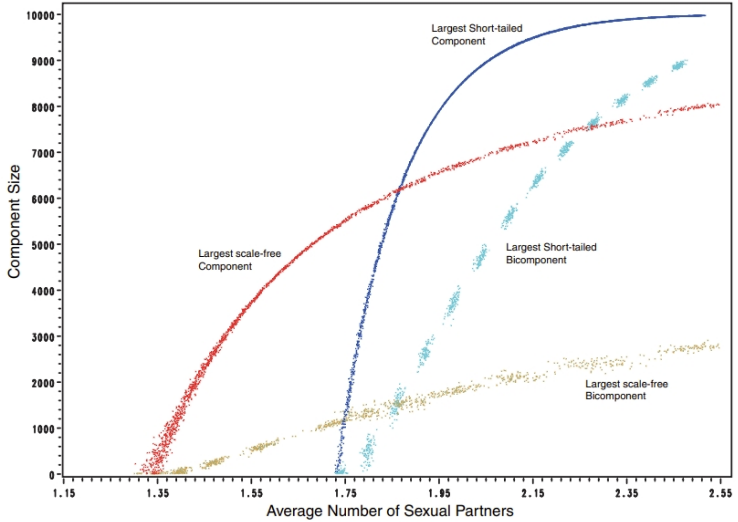


Fig. 3. Size of Largest component and bicomponent by average number of sexual partners for short-tailed and scale-free distributions. The curves plot the growth of the largest component and bicomponent as a function of the average degree, based on 100 simulations of a 10,000-node network at each degree setting. The red curve plots the analytic solution for the size of the giant component for the simulated networks with scale-free distributions, and the orange curve plots the largest bicomponent. The dark blue curve plots the analytic solution for the size of the largest component for the simulated low-degree networks, and the light blue curve plots the size of the largest bicomponent. The bicomponent curves are not continuous due to sampling.

Moody, J., Adams, J., & Morris, M. (2017). Epidemic potential by sexual activity distributions. Network science, 5(4), 461-475.

Motivation: understand online “epidemic”

- Lies spread faster than the truth

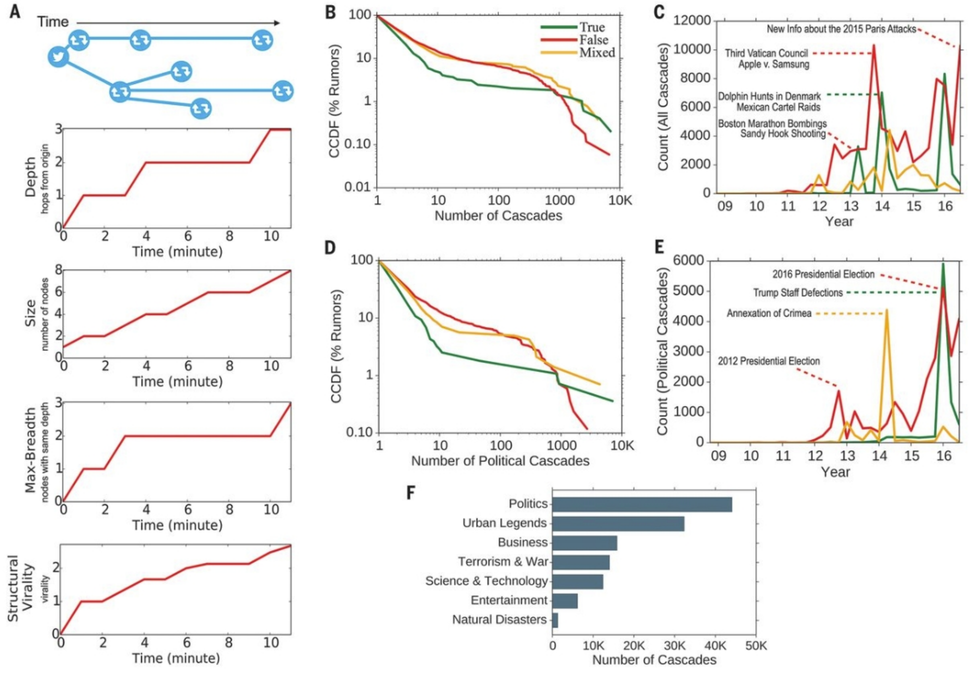


Fig. 1 **Rumor cascades.**

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. science, 359(6380), 1146-1151.

Motivation: how to succeed as individual

- Looking for a job? Making Weak Ties.
- Want to be influential? Try something new, but don’t go too far.

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.

Atypical Combinations and Scientific Impact

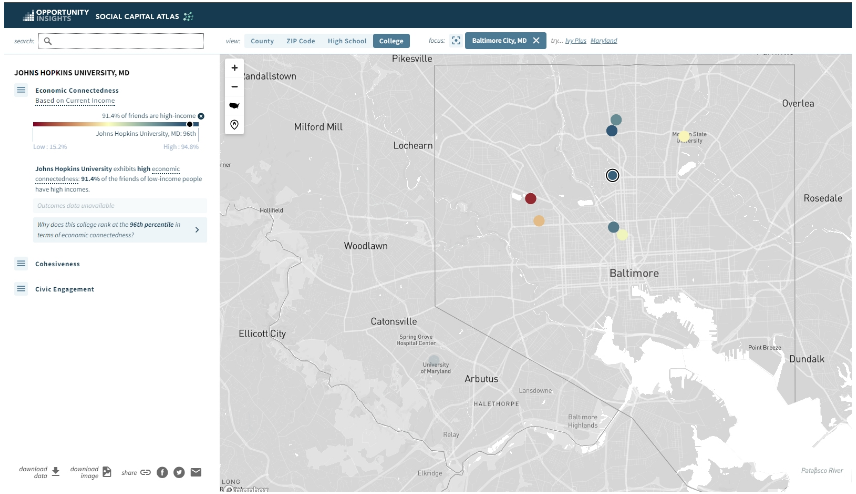
Brian Uzzi,[1,2] Satyam Mukherjee,[1,2] Michael Stringer,[2,3] Ben Jones[1,4]*

Novelty is an essential feature of creative ideas, yet the building blocks of new ideas are often embodied in existing knowledge. From this perspective, balancing atypical knowledge with conventional knowledge may be critical to the link between innovativeness and impact. Our analysis of 17.9 million papers spanning all scientific fields suggests that science follows a nearly universal pattern: The highest-impact science is primarily grounded in exceptionally conventional combinations of prior work yet simultaneously features an intrusion of unusual combinations. Papers of this type were twice as likely to be highly cited works. Novel combinations of prior work are rare, yet teams are 37.7% more likely than solo authors to insert novel combinations into familiar knowledge domains.

Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology*, *78*(6), 1360-1380.
Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. Science, 342(6157), 468-472.

Motivation: how to promote mobility as society

- https://socialcapital.org/
- Go to the right schools and make the right friends



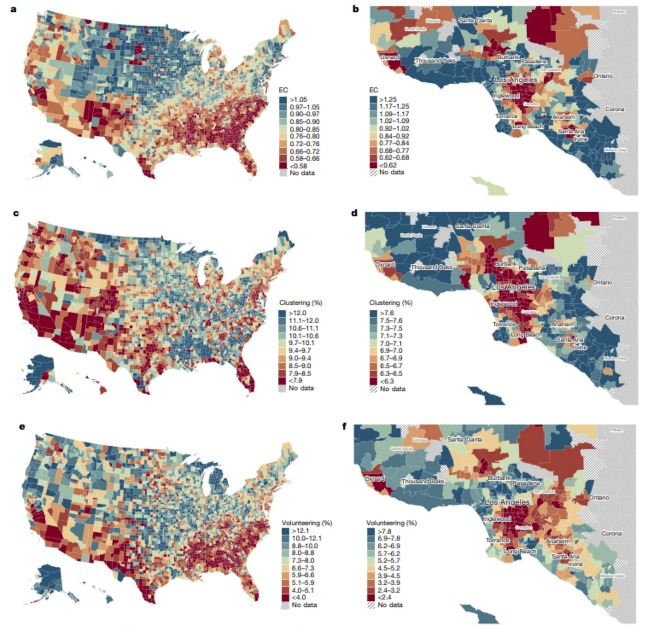


Fig. 2 | The geography of social capital in the United States. **a**, County-level map of EC, defined as twice the share of friends with above-median SES among people with below-median SES. **b**, ZIP-code-level map of EC in Los Angeles. **c**, County-level map of average clustering, defined as the share of an individual's friend pairs who are friends with each other. **d**, ZIP-code-level map of average clustering in Los Angeles. **e**, County-level map of volunteering rates, defined as the percentage of individuals who are members of volunteering or activism groups as classified by Facebook. **f**, ZIP-code-level map of volunteering rates in Los Angeles. We omit counties and ZIP codes where statistics are estimated on fewer than 100 Facebook users with below-median SES. These maps must be viewed in colour to be interpretable. Analogous maps for all ZIP codes in the United States are available at https://www.socialcapital.org. Extended Data Fig. 1 presents county-level maps of other social capital measures. Maps were made with the QGIS software package.

Chetty, R., Jackson, M. O., Kuchler, T., Stroebel, J., Hendren, N., Fluegge, R. B., ... & Wernerfelt, N. (2022). Social capital I: measurement and associations with economic mobility. *Nature*, *608*(7921), 108-121.

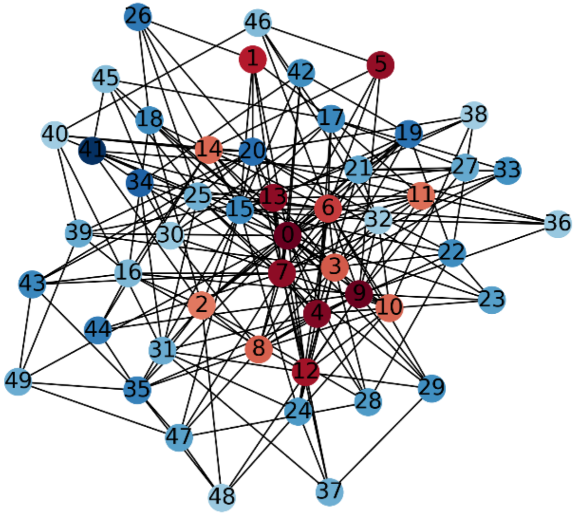
How might we represent network?

Represent connections between vertices/nodes

- Vertex: a node of the graph
- Edge: a link between two vertices

A graph consists of a set of nodes and a set of edges

- $G(V, E)$



Graph Data: Adjacency Matrix

- The matrix of vertices connections

Encode in a symmetric matrix (for undirected network)

$(n \times n)\ matrix\ A$

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

The adjacency matrix has elements

$$a_{ij} = \begin{cases} 1 & if\ i\ and\ j\ are\ connected \\ 0 & otherwise \end{cases}$$

	Mark	Peter	Bob	Jill	Aaron
Mark	0	1	0	1	0
Peter	1	0	1	0	1
Bob	0	1	0	1	0
Jill	1	0	1	0	1
Aaron	0	1	0	1	0

Example from: https://bookdown.org/markhoff/social_network_analysis/understanding-network-data-structures.html

Graph Data: Edge Lists

- Two-column matrices that directly indicate how vertices are connected

	Mark	Peter	Bob	Jill	Aaron
Mark	0	1	0	1	0
Peter	1	0	1	0	1
Bob	0	1	0	1	0
Jill	1	0	1	0	1
Aaron	0	1	0	1	0

	PersonA	PersonB
1	Mark	Peter
2	Mark	Jill
3	Peter	Bob
4	Peter	Aaron
5	Bob	Jill
6	Jill	Aaron

Netzschleuder network catalogue, repository and centrifuge

Networks Stats API Git Issues Contribute Health About

Multiple regexp terms separated by '&' Search

Tip: click on the table header to sort the list. Hover your mouse over it to obtain a legend.

Name	Title	Nodes	Edges	$\langle k \rangle$	σ_k	λ_h	τ	r	c	$\oslash$	S	Kind	Mode	n	Tags
7th_graders	Vickers 7th Graders (1981)	29	740	25.52	20.34	17.73	1.71	-0.01	0.76	2	1.00	Directed	Unipartite	1	Social Offline Multilayer Unweighted Metadata
academia_edu	Academica.edu (2011)	200,169	1,398,063	6.98	46.24	109.99	78.34	-0.02	0.04	16	1.00	Directed	Unipartite	1	Social Online Unweighted
add_health	Adolescent health (ADD HEALTH) (1994)	2,587	12,969	5.01	5.65	11.92	29.03	0.29	0.17	10	0.98	Directed	Unipartite	84	Social Offline Weighted
adjnoun	Word adjacencies of David Copperfield	112	425	7.59	6.85	11.54	2.27	-0.13	0.16	5	1.00	Undirected	Unipartite	1	Informational Language Unweighted
advogato	Advogato trust network (2009)	6,541	51,127	7.82	34.13	68.61	20.71	-0.05	0.11	9	0.77	Directed	Unipartite	1	Social Online Weighted
amazon_copurchases	Amazon co-purchasing network (2003)	410,236	3,356,824	8.18	16.30	40.36	1805.09	-0.01	0.25	22	1.00	Directed	Unipartite	4	Economic Commerce Unweighted
amazon_ratings	Amazon customer ratings (2010)	3,376,972	5,838,041	3.46	19.33	83.61	610.18	-0.02	0.00	28	0.86	Undirected	Bipartite	1	Economic Preferences Weighted Timestamps
ambassador	Philippines Ambassador bombing (2000)	16	19	2.38	2.23	3.23	3.96	-0.21	0.59	4	0.69	Undirected	Unipartite	15	Social Offline Weighted Temporal
anybeat	Anybeat social network (2013)	12,645	67,053	5.30	89.97	92.23	41.02	-0.12	0.02	10	1.00	Directed	Unipartite	1	Social Online Unweighted
arxiv_authors	Arxiv authors (1993-2003)	133,280	396,160	5.94	27.24	92.56	158.60	0.21	0.32	14	0.13	Undirected	Unipartite	5	Social Collaboration Unweighted Projection
arxiv_citation	arXiv citation networks (1993-2003)	34,546	421,578	12.20	30.90	74.33	63.13	-0.01	0.15	14	1.00	Directed	Unipartite	2	Informational Citation Unweighted
arxiv_collab	Scientific collaborations in physics (1995-2005)	40,421	175,692	8.69	12.73	49.17	232.91	0.19	0.25	18	0.90	Undirected	Unipartite	5	Social Collaboration Weighted Projection
as_skitter	Skitter IP graph (2005)	1,696,415	11,095,298	13.08	136.86	653.66	2574.51	-0.08	0.01	31	1.00	Undirected	Unipartite	1	Technological Communication Unweighted
at_migrations	Austrian internal migrations (2002-2022)	2,115	2,908,569	1375.21	4635.25	475.07	4.50	-0.07	0.59	3	1.00	Directed	Unipartite	1	Social Economic Travel Weighted Political Timestamps Metadata
bag_of_words	Bag of words (2008)	8,341,043	483,450,157	115.92	3196.79	3405.63	2.29	-0.16	0.00	5	1.00	Undirected	Bipartite	5	Informational Text Bipartite Weighted Metadata
baidu	Chinese online encyclopedias (2011)	2,141,300	17,794,839	8.31	171.13	431.03	319.36	-0.03	0.00	20	0.98	Directed	Unipartite	1	Informational Web graph Unweighted
baseball	Baseball steroid use (2008)	84	84	2.00	4.78	1.73	6.60	-0.45	0.00	4	0.56	Undirected	Bipartite	2	Social Offline Weighted Projection
berkstan_web	Webgraph (Berkeley-Stanford)	685,231	7,600,595	11.09	285.98	663.80	389618.19	-0.10	0.01	208	0.96	Directed	Unipartite	1	Informational Web graph Unweighted
bible_nouns	Bible noun phrases	1,773	9,131	10.30	17.87	37.62	16.49	-0.05	0.16	8	0.96	Undirected	Unipartite	1	Informational Language Weighted
bibsonomy	BibSonomy	972,120	2,555,080	5.26	201.14	167.83	9651.27	-0.05	0.00	22	0.96	Undirected	Bipartite	1	Informational Folksonomy Unweighted Multigraph Timestamps
bison	Bison dominance	26	314	12.08	8.37	17.33	0.96	-0.06	0.79	2	1.00	Directed	Unipartite	1	Social Animal Weighted
bitcoin	Bitcoin transactions (2009-2013)	6,336,770	16,057,711	2.53	396.54	358.35	231730.89	-0.07	0.00	2,053	0.99	Directed	Unipartite	1	Economic Cryptocurrency Unweighted Timestamps
bitcoin_alpha	Bitcoin Alpha trust network (2017)	3,783	24,186	6.39	34.44	45.31	25.67	-0.16	0.08	10	1.00	Directed	Unipartite	1	Social Online Weighted Signed Timestamps
bitcoin_trust	Bitcoin OTC trust network (2017)	5,881	35,592	6.05	38.30	50.83	25.68	-0.16	0.06	9	1.00	Directed	Unipartite	1	Social Online Weighted Signed Timestamps

Example from: https://bookdown.org/markhoff/social_network_analysis/understanding-network-data-structures.html
Collections of Social Network Datasets: https://networks.skewed.de/

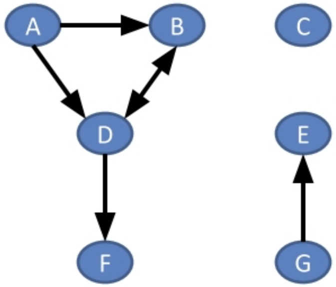
Types of Edges

- Directed vs. undirected

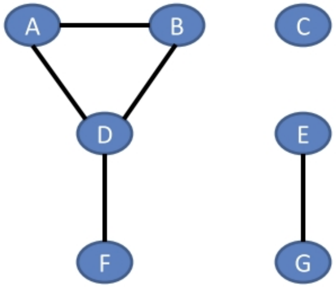
Directed & undirected

- Communication vs. friendship networks

twitter



facebook



Directed sociomatrix

	A	B	C	D	E	F	G
A	-	1	0	1	0	0	0
B	0	-	0	1	0	0	0
C	0	0	-	0	0	0	0
D	0	1	0	-	0	1	0
E	0	0	0	0	-	0	0
F	0	0	0	0	0	-	0
G	0	0	0	0	1	0	-

Unsymmetric
Symmetric
Unsymmetric

Undirected sociomatrix

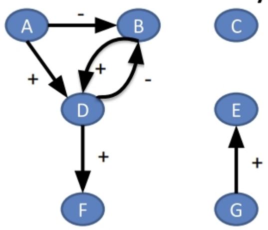
	A	B	C	D	E	F	G
A	-	1	0	1	0	0	0
B	1	-	0	1	0	0	0
C	0	0	-	0	0	0	0
D	1	1	0	-	0	1	0
E	0	0	0	0	-	0	1
F	0	0	0	1	0	-	0
G	0	0	0	0	1	0	-

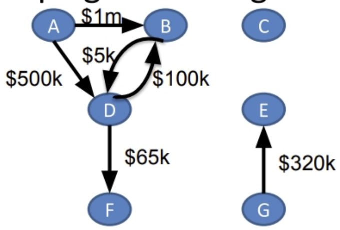
Symmetric
Symmetric
Symmetric

Types of Edges

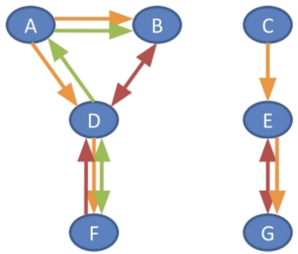
- Weighted vs. unweighted
- Multiplex

- Affect in a sorority vs. campaign financing



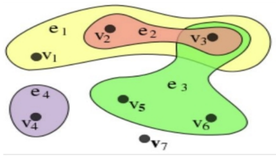


Organizations: authority, trust, & friendship



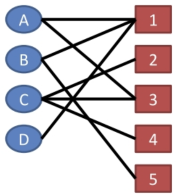
Hypergraph Incidence Matrix

	e_1	e_2	e_3	e_4
v_1	1	0	0	0
v_2	1	1	0	0
v_3	1	1	1	0
v_4	0	0	0	1
v_5	0	0	1	0
v_6	0	0	1	0
v_7	0	0	0	0



Bipartite sociomatrix

	1	2	3	4	5
A	1	0	1	0	0
B	1	0	0	0	1
C	0	1	1	1	0
D	1	0	0	0	0



Example from: https://sonic.northwestern.edu/
Example of hypergraph: Lungeanu, A., Carter, D. R., DeChurch, L. A., & Contractor, N. S. (2021). How team interlock ecosystems shape the assembly of scientific teams: A hypergraph approach. In Computational Methods for Communication Science (pp. 95-119). Routledge.

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Basic Metrics

Network Parameters

Different Dimensions to Consider

- **Entity:** Nodes vs. Edges (e.g., degree, path length)
- **Scale:** Local vs. Global (e.g., cluster, dimensions)
- **Topology:** Structure (e.g., small world network, scale-free network)
- **Quantity:** Volume (e.g., weighted edges)
- **Quality:** Classification (e.g., friends, family, …)
- …

Different combinations of dimensions create different network metrics;

You can always **create your own**.

Example 1: Network Density

Edges * Global (Ignore multiplex hypergraph topology for all examples)

- For a directed unweighted network with n nodes, the max number of possible edges is:

$n(n-1)$

- For an undirected unweighted network:

$n(n-1)/2$

- Network density: $$\frac{Number\ of\ edges}{Number\ of\ possible\ edges}$$

Americans are becoming more isolated

Table 3. Structural Characteristics of Core Discussion Networks

	1985 (N = 1,167[a])	2004 (N = 788[b])
Network Density		
<.25	9.9%	7.3%
.25–.49	18.5%	11.8%
.50–.74	37.9%	39.5%
>.74	33.7%	41.4%
Mean	.60	.66
SD	.33	.33
Mean Frequency of Contact (days per year)		
6–12	3.7%	3.0%
>12–52	15.3%	10.6%
>52–365	81.0%	86.4%
Mean	208.92	243.81
SD	117.08	114.86
Length of Association (in years)		
>0–4.5	12.1%	10.7%
>4.5–8+	87.9%	89.3%
Mean	6.72	7.01
SD	1.34	1.00

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

McPherson, M., Smith-Lovin, L., & Brashears, M. E. (2006). Social isolation in America: Changes in core discussion networks over two decades. American sociological review, 71(3), 353-375.

Example 2: Closeness Centrality

Nodes * Global

- Measuring the mean shortest distance from a node to every other nodes in a network with n nodes:

$$\frac{1}{n-1}\sum d_{ij}$$

- Where d represent the length of the shortest path between i and j. Here, the path length refers to the number of nodes between i and j (degrees of separation).

How minorities generate impact from a peripheral location

- Start from periphery and channel through emotions (sentiment analysis)

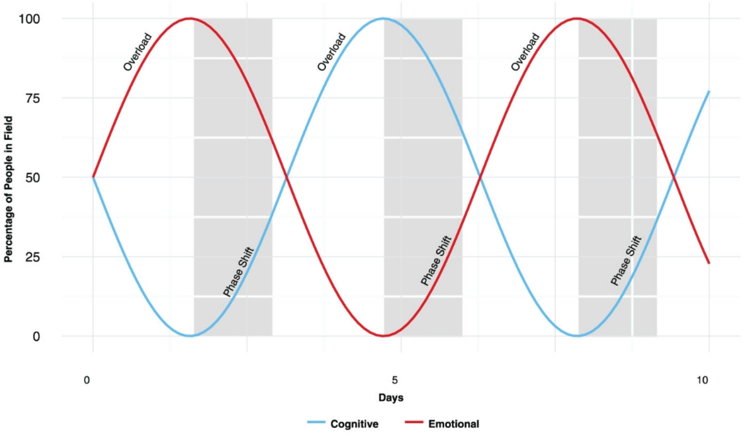


Figure 1. Idealized Opportunity Structure Created by Cognitive-Emotional Currents

Table 1. Variable Descriptions and Descriptive Statistics

Variable	Description	Mean	SD
Media Influence (Outcome)	Number of words in press release reproduced verbatim or paraphrased by six national media sources.	4.590	18.736
Fringe Media Frames	Euclidean distance between five dummy variables describing civil society organization media frames about Islam in each press release and average for all other organizations during the same year.	.913	.197
Assets	Total assets of organization sponsoring press release at year-end	27.0 (mill.)	68.3 (mill.)
Inter-organizational Networks	Closeness centrality of organization within field (constructed using interlocking directorates by year).	.188	.355
Narrowness of Mission	Dummy variable that describes whether organization's primary goal is influencing media discourse about Islam (1 = yes, 0 = no).	.493	.500
Displays of Fear or Anger	Dummy variable that describes whether civil society organization displays fear or anger in press release (1 = yes, 0 = no).	.654	.478
News Cycle	Number of hits for the term "Muslim" or "Islam" on Google News during month the press release was issued.	8,264	2,830
Previous Media Coverage	Dummy variable that describes whether civil society organization issuing the press release previously influenced media discourse about Islam.	.524	.500
U.S. Government Targeted	Dummy variable that describes whether the press release targets an individual or organization representing the U. S. government (1 = yes, 0 = no).	.283	.451
Public Interest	Dummy variable that describes whether main event described in the press release was one of the top-10 Google searches during the week it was issued (1 = yes, 0 = no).	.061	.239
Violence or Disruptive Activity	Dummy variable that describes whether main event described in the press release involved physical violence, strikes, protests, rallies, or boycotts (1 = yes, 0 = no).	.223	.416
Event in United States	Dummy variable that describes whether main event described in the press release occurred in the United States (1 = yes, 0 = no).	.572	.450

Bail, C. A. (2012). The fringe effect: Civil society organizations and the evolution of media discourse about Islam since the September 11th attacks. *American Sociological Review*, *77*(6), 855-879.
Bail, C. A., Brown, T. W., & Mann, M. (2017). Channeling hearts and minds: Advocacy organizations, cognitive-emotional currents, and public conversation. American Sociological Review, 82(6), 1188-1213.

Example 3: Quarter-Power Scaling

Topology * Volume * Scale

- Observation: Many biological scaling can be described as

$$Y = aM^b$$

Where Y is a biological variable, such as "*life span*"; a is a constant, b is a scaling exponent; M is a metabolic measurement, such as "*blood circulation time*". The value of b is usually ¼ or ¾.

We also have similar observations in economic growth, innovation, and pace of life in cities.

West, G. B., Brown, J. H., & Enquist, B. J. (1999). The fourth dimension of life: fractal geometry and allometric scaling of organisms. science, 284(5420), 1677-1679.

Bettencourt, L. M., Lobo, J., Helbing, D., Kühnert, C., & West, G. B. (2007). Growth, innovation, scaling, and the pace of life in cities. Proceedings of the national academy of sciences, 104(17), 7301-7306.

- Theory: maximize metabolic capacity - transportation through space-filling fractal networks of branching tubes

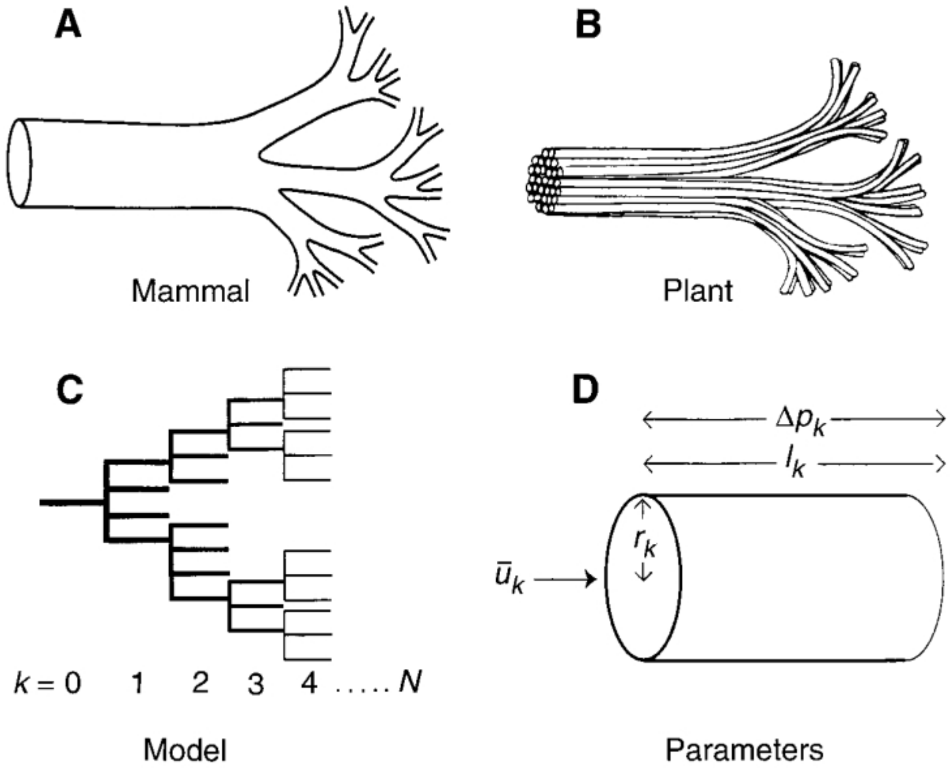


Fig. 1. Diagrammatic examples of segments of biological distribution networks: (**A**) mammalian circulatory and respiratory systems composed of branching tubes; (**B**) plant vessel-bundle vascular system composed of diverging vessel elements; (**C**) topological representation of such networks, where k specifies the order of the level, beginning with the aorta ($k = 0$) and ending with the capillary ($k = N$); and (**D**) parameters of a typical tube at the kth level.

West, G. B., Brown, J. H., & Enquist, B. J. (1997). A general model for the origin of allometric scaling laws in biology. *Science*, *276*(5309), 122-126.

Table 1. Values of allometric exponents for variables of the mammalian cardiovascular and respiratory systems predicted by the model compared with empirical observations. Observed values of exponents are taken from (*2*, *3*); ND denotes that no data are available.

Cardiovascular				Respiratory			
Variable	Exponent			Variable	Exponent		
	Predicted		Observed		Predicted		Observed
Aorta radius r_0	3/8 =	0.375	0.36	Tracheal radius	3/8 =	0.375	0.39
Aorta pressure Δp_0	0 =	0.00	0.032	Interpleural pressure	0 =	0.00	0.004
Aorta blood velocity u_0	0 =	0.00	0.07	Air velocity in trachea	0 =	0.00	0.02
Blood volume V_b	1 =	1.00	1.00	Lung volume	1 =	1.00	1.05
Circulation time	1/4 =	0.25	0.25	Volume flow to lung	3/4 =	0.75	0.80
Circulation distance l	1/4 =	0.25	ND	Volume of alveolus V_A	1/4 =	0.25	ND
Cardiac stroke volume	1 =	1.00	1.03	Tidal volume	1 =	1.00	1.041
Cardiac frequency ω	−1/4 =	−0.25	−0.25	Respiratory frequency	−1/4 =	−0.25	−0.26
Cardiac output $\dot{E}$	3/4 =	0.75	0.74	Power dissipated	3/4 =	0.75	0.78
Number of capillaries N_c	3/4 =	0.75	ND	Number of alveoli N_A	3/4 =	0.75	ND
Service volume radius	1/12 =	0.083	ND	Radius of alveolus r_A	1/12 =	0.083	0.13
Womersley number α	1/4 =	0.25	0.25	Area of alveolus A_A	1/6 =	0.083	ND
Density of capillaries	−1/12 =	−0.083	−0.095	Area of lung A_L	11/12 =	0.92	0.95
O_2 affinity of blood P_{50}	−1/12 =	−0.083	−0.089	O_2 diffusing capacity	1 =	1.00	0.99
Total resistance Z	−3/4 =	−0.75	−0.76	Total resistance	−3/4 =	−0.75	−0.70
Metabolic rate B	3/4 =	0.75	0.75	O_2 consumption rate	3/4 =	0.75	0.76

West, G. B., Brown, J. H., & Enquist, B. J. (1997). A general model for the origin of allometric scaling laws in biology. *Science*, *276*(5309), 122-126.

List of Other Metrics

Node Degree (in-degree; out-degree)

Degree distribution

Betweenness centrality

Eigenvector centrality

Page Rank (Google)

Constraint (Structure hole)

Hubs and Authorities (HITS)

Clustering coefficient

Components

Subgraphs

N-cliques

N-clans

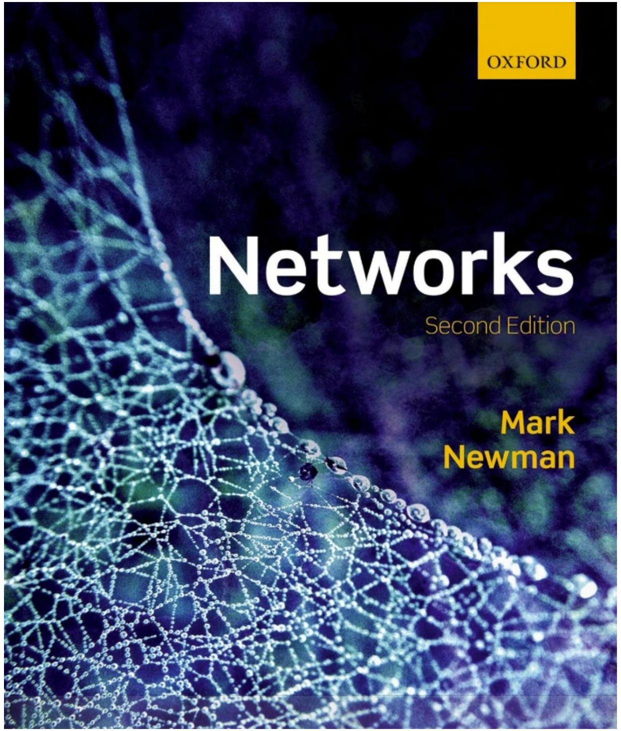
K-plexes

K-cores

Structural Equivalence

Shortcut

…



For more information, refer to textbooks, Wikipedia or python/R packages (e.g. NetworkX https://networkx.org/)

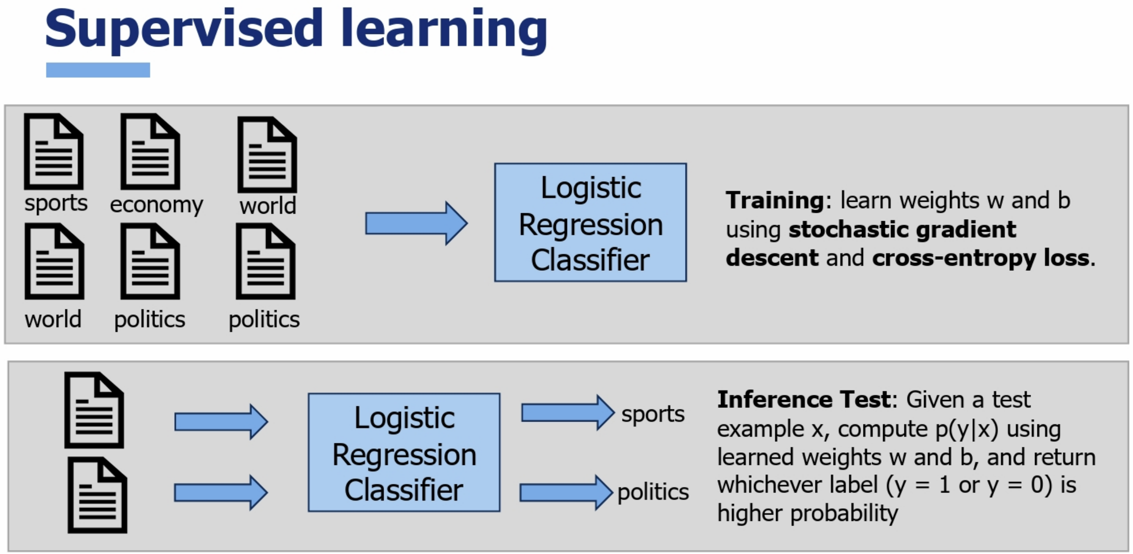
Newman, M. (2018). *Networks*. Oxford university press.

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Advanced Network Methods

Call back

- **Logistic Regression** (Feb 14) assume **independence of errors**, linearity in the logit for continuous variables, absence of multicollinearity, and lack of strongly influential outliers



Stoltzfus, J. C. (2011). Logistic regression: a brief primer. *Academic emergency medicine*, *18*(10), 1099-1104.

Network “regression”

Problem:

- Analogous to logistic regression: if we want to **predict** the probability that a pair of nodes in a network will **have a tie** between them (0,1).
- Ties between nodes in real social networks are not **independent**.

Solution

- Exponential Random Graph Model (ERGM)
- Through simulation, ERGMs allow dyadic and higher-order dependencies to be modeled. Then it can describe how **interdependent structures** shape a network.

https://eehh-stanford.github.io/SNA-workshop/ergm-intro.html#what-is-an-ergm
Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., & Morris, M. (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of statistical software*, *24*(3), nihpa54860.

ERGM Model

- Observe the **distribution of structural features** of interest in simulated networks

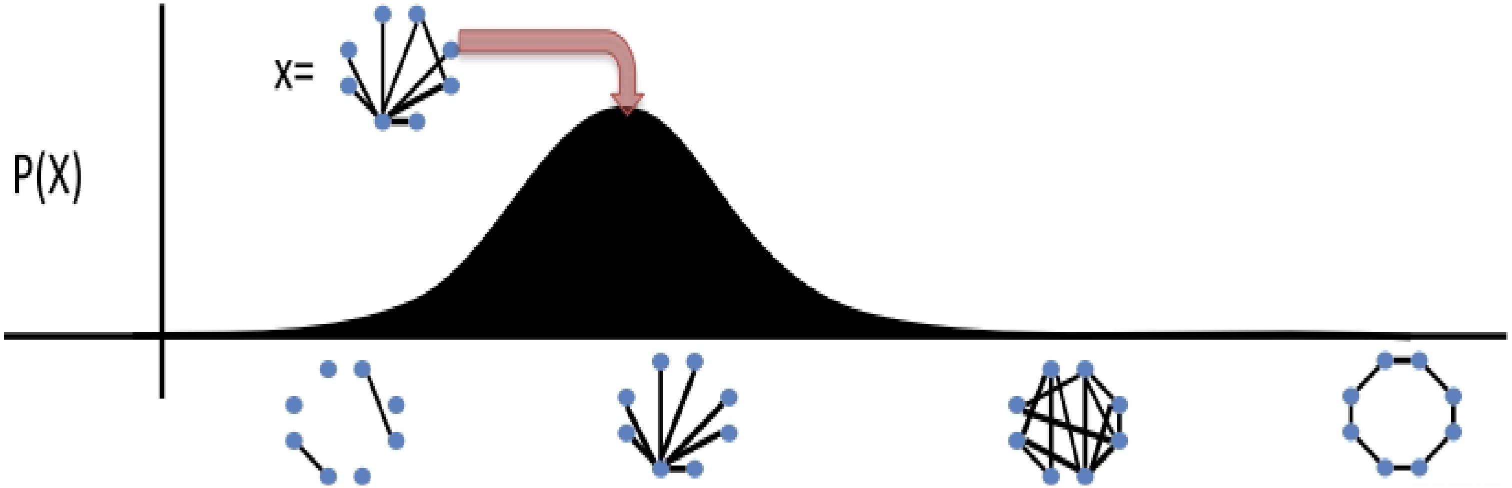


Figure from: https://sonic.northwestern.edu/

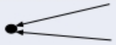
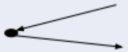
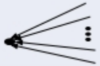
ERGM Model

- Adding different **structural metrics as X** into a "regression".

Network Statistics: Undirected

Parameter	statnet name		Parameter	statnet name	
Edge	edges		Isolates	isolates	
2-Star	kstar(2)		3-Star	kstar(3)	
Triangle	triangle		K-Star	kstar(k)	

Network Statistics: Directed

Parameter	statnet name		Parameter	statnet name	
Arc	edges		Reciprocity	mutual	
2-In-Star	istar(2)		2-Out-Star	ostar(2)	
Mixed-2-Star (two-path)	m2star				
3-In-Star	istar(3)		3-Out-Star	ostar(3)	
K-In-Star	istar(k)		K-Out-Star	ostar(k)	

See more network statistics: https://cran.r-project.org/web/packages/ergm/vignettes/ergm.pdf

ERGM Model

Let $\mathbf{Y}$ denote an $n \times n$ sociomatrix where $y_{ij} = 1$ if individuals $y_{ij} = i$ and j have a tie. Let $\mathbf{X}$ denote a matrix of covariates, which includes structural measures of the network as well as nodal and possibly edge-level attributes. A generic ERGM can be written as:

$$P_{\theta,\mathcal{Y}}(\mathbf{Y} = \mathbf{y}|\mathbf{X}) = \frac{exp\{\theta^{\top} g(y, X)\}}{\kappa(\theta, \mathcal{Y})}$$

where θ is a vector of coefficients, $g(y, \mathbf{X})$ is a vector of sufficient statistics, $\mathcal{Y}$ is the space of possible graphs, and $\kappa(\theta, \mathcal{Y})$ is a normalizing constant. That is, it's the numerator summed across all possible graphs $\mathcal{Y}$. For even moderate-sized graphs, $\kappa(\theta, \mathcal{Y})$ can be enormous, so closed-form solutions are unfeasible. The number of labeled, undirected graphs of n vertices is $2^{n(n-1)/2}$, which can get big fast. For example, for a network of $n > 7$, there are over two million undirected graphs, which means that you would need to calculate the likelihood for each one of these in order to compute κ. This is generally not practical.

ERGM Model

Some Definitions and Notation

- y_{ij} denotes the ij th dyad in graph y. If $y_{ij} = 1$, then i and j are connected by an edge, if $y_{ij} = 0$, they are not.
- y_{ij}^c is the status of all other pairs of vertices in y other than (i, j).
- y_{ij}^+ is the same network as y except that $y_{ij} = 1$.
- y_{ij}^- is the same network as y except that $y_{ij} = 0$.
- $\delta(y_{ij})$ is the *change statistic*: $\delta(y_{ij}) = g(y_{ij}^+) - g(y_{ij}^-)$. This is a measure of how the graph statistic $g(y)$ changes if the ijth vertex is toggled on or off.

The ergm equation can be re-written in terms of change statistics. The log-odds of a tie y_{ij} is:

$$logit(Y_{ij} = 1 | y_{ij}^c) = \theta^T \delta(y_{ij})$$

Example of ERGM

- How **reciprocal edges** and **number of edge** influence guarantee network in financial crisis and stimulus program?

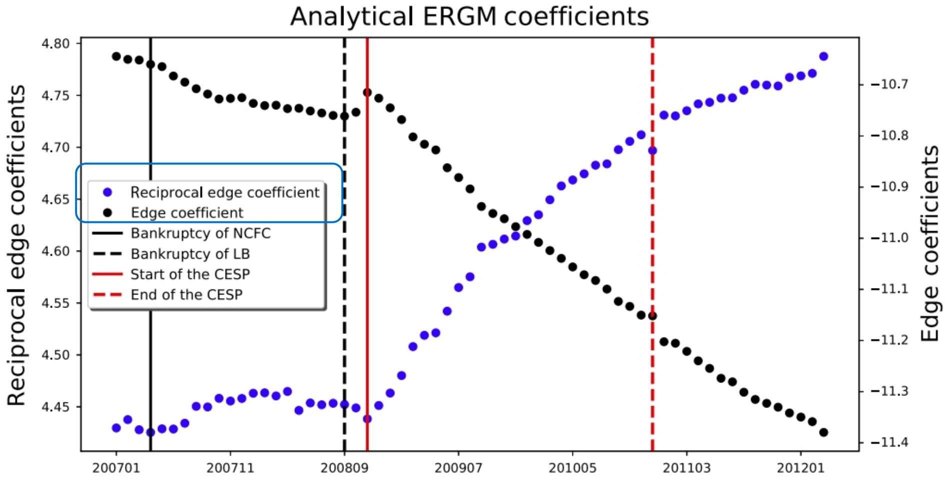
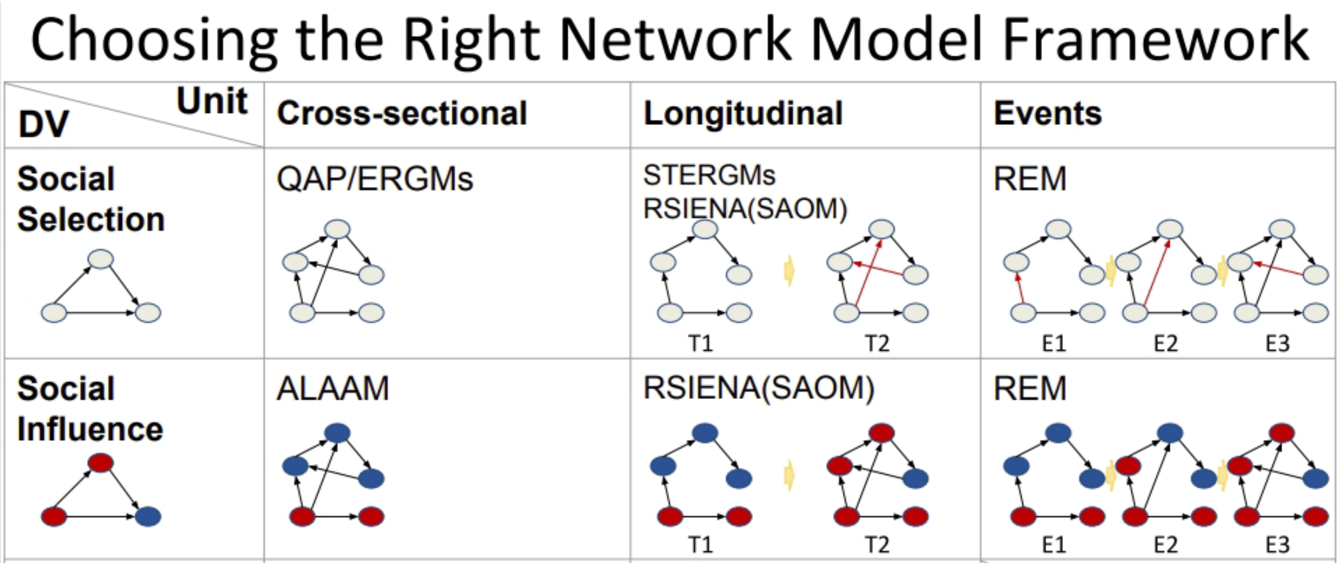


Fig. 4 Dynamic changes of coefficients in ERGM. Source data are provided as a Source Data file.

Wang, Y., Zhang, Q., & Yang, X. (2020). Evolution of the Chinese guarantee network under financial crisis and stimulus program. Nature Communications, 11(1), 2693.

Extended ERGM family and other Relevant Inference models

- Social selection: **predict ties**
- Social influence: **predict attributes of nodes**



Choosing the Right Network Model Framework

DV \ Unit	Cross-sectional	Longitudinal	Events
Social Selection	QAP/ERGMs	STERGMs RSIENA(SAOM) T1 T2	REM E1 E2 E3
Social Influence	ALAAM	RSIENA(SAOM) T1 T2	REM E1 E2 E3

Example from: https://sonic.northwestern.edu/

Problem of ERGM family

- **Not practical** for a large graph (typically within 3k-5k nodes)
- One solution is **network sampling**, sample a small graph from the large graph (another solution is Graph Neural Network)

	Static graph patterns								Temporal graph patterns				
	in-deg	out-deg	wcc	scc	hops	sng-val	sng-vec	clust	diam	cc-sz	sng-val	clust	**AVG**
RN	0.084	0.145	0.814	0.193	0.231	0.079	0.112	0.327	0.074	0.570	0.263	0.371	0.272
RPN	**0.062**	**0.097**	0.792	0.194	**0.200**	0.048	0.081	0.243	0.051	0.475	0.162	0.249	0.221
RDN	0.110	0.128	0.818	0.193	0.238	0.041	0.048	0.256	0.052	0.440	**0.097**	0.242	0.222
RE	0.216	0.305	**0.367**	0.206	0.509	0.169	0.192	0.525	0.164	0.659	0.355	0.729	0.366
RNE	0.277	0.404	0.390	0.224	0.702	0.255	0.273	0.709	0.370	0.771	0.215	0.733	0.444
HYB	0.273	0.394	0.386	0.224	0.683	0.240	0.251	0.670	0.331	0.748	0.256	0.765	0.435
RNN	0.179	0.014	0.581	0.206	0.252	0.060	0.255	0.398	0.058	0.463	0.200	0.433	0.258
RJ	0.132	0.151	0.771	0.215	0.264	0.076	0.143	**0.235**	0.122	0.492	0.161	**0.214**	0.248
RW	0.082	0.131	0.685	0.194	0.243	0.049	**0.033**	**0.243**	**0.036**	**0.423**	**0.086**	0.224	**0.202**
FF	0.082	0.105	0.664	0.194	**0.203**	**0.038**	0.092	**0.244**	0.053	0.434	0.140	**0.211**	**0.205**

Table 1: Scale-down sampling criteria. On average RW and FF perform best.

Leskovec, J., & Faloutsos, C. (2006, August). Sampling from large graphs. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 631-636).

Call back

- Causal inference (Feb 14)
- How to conduct causal inference in network analysis?

How can we measure ATE without this problem?

- Randomized control trial (RCT)
- More realistic scenario:
  - We'll probably study effects of medicine on someone who is sick
  - If we survey people, there still might be differences: lower income person may not be able to afford medicine and may also have worse nutrition that leads to more severe illness: income is a confounder (X)
- Instead of surveying people, we take a group of people and randomly assign them to "treatment" or "control" group

Stoltzfus, J. C. (2011). Logistic regression: a brief primer. *Academic emergency medicine*, *18*(10), 1099-1104.

Example 1: Simulation + Matching

- Remove matched nodes and see what happens

Malfeasance and the Foundations for Global Trade: The Structure of English Trade in the East Indies, 1601–1833[1]

Emily Erikson
University of Massachusetts, Amherst

Peter Bearman
Columbia University

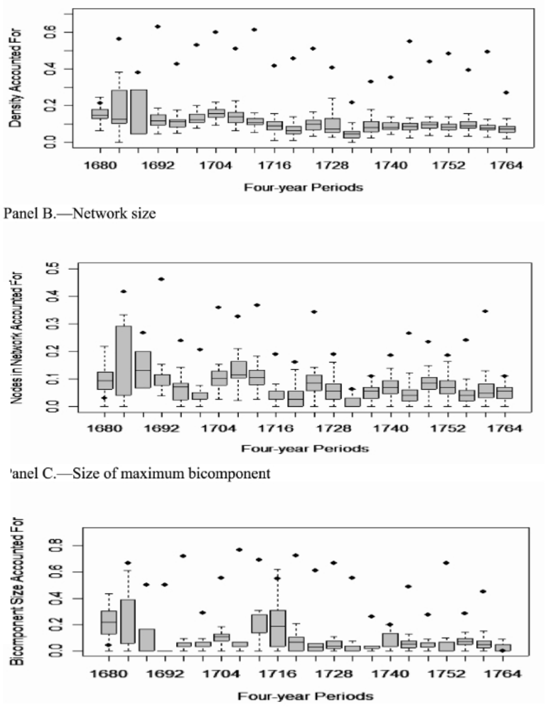


FIG. 8.—Simulations of data presented in fig. 6

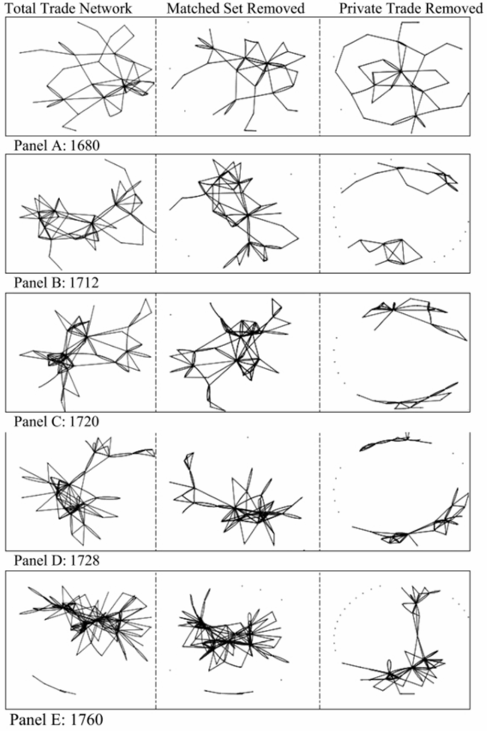


FIG. 4.—Network visualizations of the EIC's Eastern trade

Erikson, E., & Bearman, P. (2006). Malfeasance and the foundations for global trade: The structure of English trade in the East Indies, 1601–1833. American Journal of Sociology, 112(1), 195-230., J. C. (2011). Logistic regression: a brief primer. *Academic emergency medicine*, *18*(10), 1099-1104.

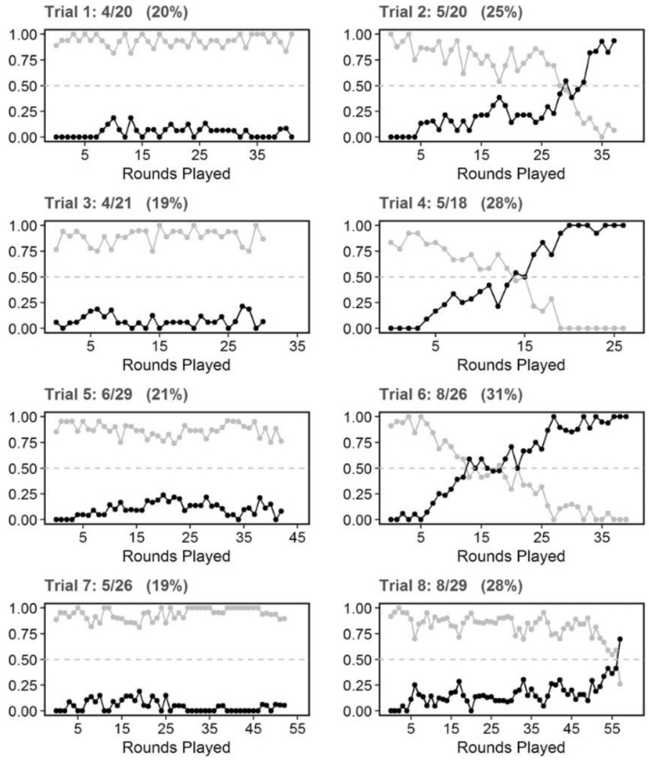
Example 2: Experiment

- Recruit people and allocate them into different networks.

Experimental evidence for tipping points in social convention

DAMON CENTOLA, JOSHUA BECKER, DEVON BRACKBILL, AND ANDREA BARONCHELLI Authors Info & Affiliations

SCIENCE • 8 Jun 2018 • Vol 360, Issue 6393 • pp. 1116-1119 • DOI: 10.1126/science.aas8827



Centola, D., Becker, J., Brackbill, D., & Baronchelli, A. (2018). Experimental evidence for tipping points in social convention. *Science*, *360*(6393), 1116-1119.

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Graph Neural Network

Call back: Large Graph Issue for ERGM

- Solution 1: **Network sampling**.
- Solution 2: **Transform** graph information to other data structures (e.g., node embedding).
- Solution 3: Analyzing the graph at the local neural level and then aggregating the neurons together (e.g., **Graph Neural Network**).

- These 3 solutions are actually intertwined in practice:

You can use network sampling methods (e.g., random walk) to calculate node embeddings;

You can also use node embedding results as input for Graph Neural Networks (GNN).

Node Embedding

- Logic of Node Embedding

1. Define a function that maps node u, v to vectors z_u, z_v
2. Define a node similarity function for u, v
3. Optimize parameters so that:

$$similarity(u, v) \approx z_v^T z_u$$

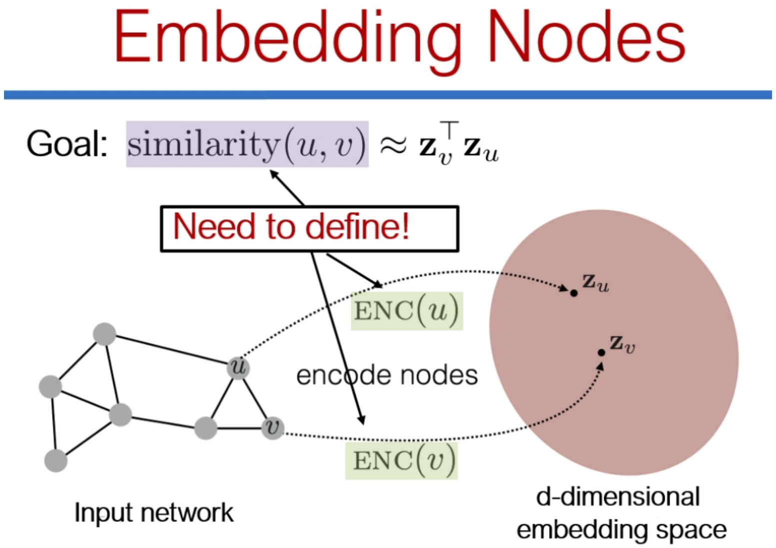
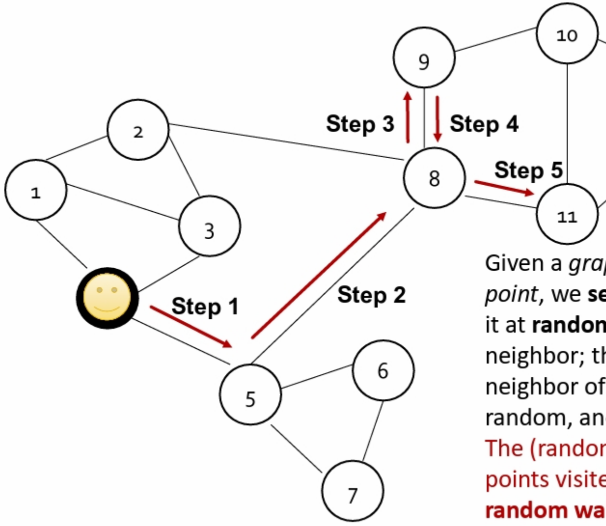


Illustration graph from: https://cs.stanford.edu/people/jure/teaching.html

Example: similarity based on random walks

- Given a random node u, predict its neighbor $N_R(u)$, equivalently minimizing L.
- Intuition: Optimize embedding z_u to max the likelihood of random walk co-occurrences.



Given a *graph* and a *starting point*, we **select a neighbor** of it at **random**, and move to this neighbor; then we select a neighbor of this point at random, and move to it, etc. The (random) sequence of points visited this way is a **random walk on the graph**.

1. Simulate many short random walks starting from each node using a strategy R
2. For each node u, get $N_R(u)$ as a sequence of nodes visited by random walks starting at u
3. For each node u, learn its embedding by predicting which nodes are in $N_R(u)$:

$$\mathcal{L} = \sum_{u \in V} \sum_{v \in N_R(u)} -\log(P(v|\mathbf{z}_u))$$

Can efficiently approximate using negative sampling

Example from: https://cs.stanford.edu/people/jure/teaching.html

Example: similarity based on random walks

- Given a random node u, predict its neighbor $N_R(u)$, equivalently minimizing L.
- Intuition: Optimize embedding z_u to max the likelihood of random walk co-occurrences.
- Use **softmax to parameterize $P(v|z_u)$** (make v to be most similar to u).

$$\mathcal{L} = \sum_{u \in V} \sum_{v \in N_R(u)} -\log\left(\frac{\exp(\mathbf{z}_u^\top \mathbf{z}_v)}{\sum_{n \in V} \exp(\mathbf{z}_u^\top \mathbf{z}_n)}\right)$$

sum over all nodes u

sum over nodes v seen on random walks starting from u

predicted probability of u and v co-occuring on random walk, i.e., use softmax to parameterize $P(v|\mathbf{z}_u)$

Random walk embeddings = z_u minimizing L

Example from: https://cs.stanford.edu/people/jure/teaching.html

Recall negative sampling in word2vec

- Calculating L is **expensive**: pick random negative samples to normalize
- **Negative sampling** (Jan 31)

$$\mathcal{L} = \sum_{u \in V} \sum_{v \in N_R(u)} -\log\left(\frac{\exp(\mathbf{z}_u^\top \mathbf{z}_v)}{\sum_{n \in V} \exp(\mathbf{z}_u^\top \mathbf{z}_n)}\right)$$

sum over all nodes u

sum over nodes v seen on random walks starting from u

predicted probability of u and v co-occuring on random walk, i.e., use softmax to parameterize $P(v|\mathbf{z}_u)$

Skip-gram: Negative sampling

$$\frac{\exp(u_o^T v_c)}{\sum_{i=1}^{V} \exp(u_i^T v_c)}$$

Encourage center word and context word to have similar vectors

Encourage center word and all other words to have different vectors

Random walk embeddings = $\boldsymbol{z}_u$ minimizing **L**

Example from: https://cs.stanford.edu/people/jure/teaching.html

Recall negative sampling in word2vec

- Calculating L is expensive: pick random negative samples to normalize
- Negative sampling (Jan 31): **Sample k negative nodes** each with prob. proportional to its **degree** (k=5~20)
- **Gradient Descent** to minimize L

Skip-gram: Negative sampling

$$\frac{\exp(u_o^T v_c)}{\sum_{i=1}^{V} \exp(u_i^T v_c)}$$

Encourage center word and context word to have similar vectors

Encourage center word and all other words to have different vectors

Solution: Negative sampling (Mikolov et al., 2013)

$$\log\left(\frac{\exp(\mathbf{z}_u^\top \mathbf{z}_v)}{\sum_{n\in V}\exp(\mathbf{z}_u^\top \mathbf{z}_n)}\right)$$

$$\approx \log(\sigma(\mathbf{z}_u^\top \mathbf{z}_v)) - \sum_{i=1}^{k} \log(\sigma(\mathbf{z}_u^\top \mathbf{z}_{n_i})), n_i \sim P_V$$

sigmoid function

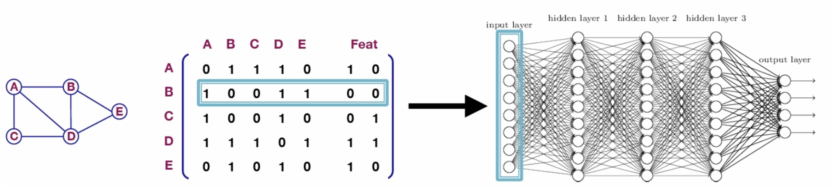
random distribution over all nodes

i.e., instead of normalizing w.r.t. all nodes, just normalize against k random negative samples

Example from: https://cs.stanford.edu/people/jure/teaching.html

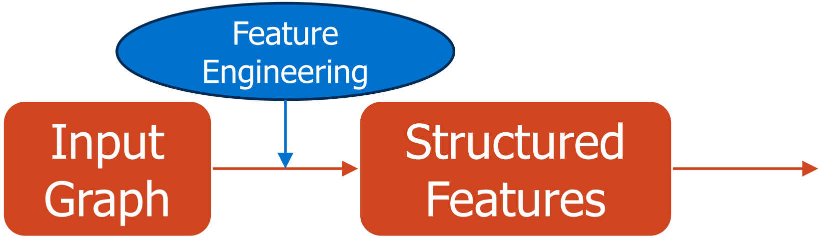
Call back Neural Network (Feb 14)

- Can we directly apply neural network to graph, taking adjacency matrix and network metrics as input?

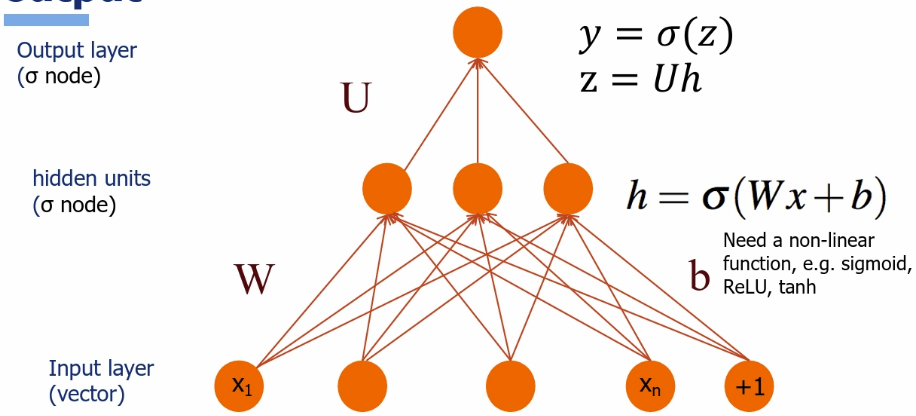


- Issues with naïve neural network

Node order; Graph size change…



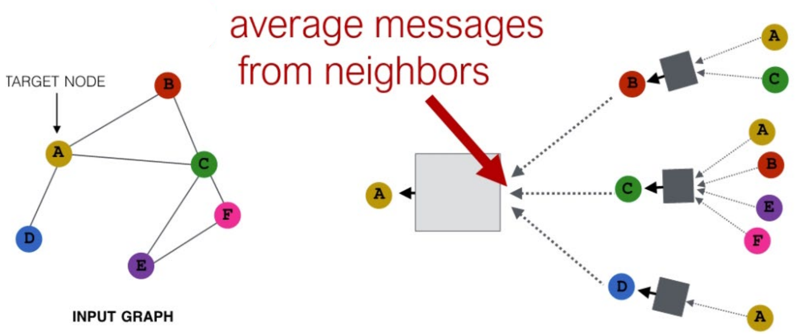
Two-layer Neural Network with scalar output



Graph Neural Network

- Logic of GNN

1) Network neighborhood defines a computation **graph**
2) Generate **node embeddings/link messages** based on local network neighborhoods
3) **Aggregate** information across layers
4) **Train** the neural network



- Basic approach: Average neighbor messages and apply a neural network

Initial 0-th layer embeddings are equal to node features

$$\boxed{\mathbf{h}_v^0 = \mathbf{x}_v}$$

Previous layer embedding of v

$$\mathbf{h}_v^k = \boxed{\sigma}\left(\mathbf{W}_k \boxed{\sum_{u \in N(v)} \frac{\mathbf{h}_u^{k-1}}{|N(v)|}} + \mathbf{B}_k \boxed{\mathbf{h}_v^{k-1}}\right), \quad \forall k \in \{1, \dots, K\}$$

$$\boxed{\mathbf{z}_v} = \mathbf{h}_v^K$$

Average of neighbor's previous layer embeddings

Embedding after K layers of neighborhood aggregation

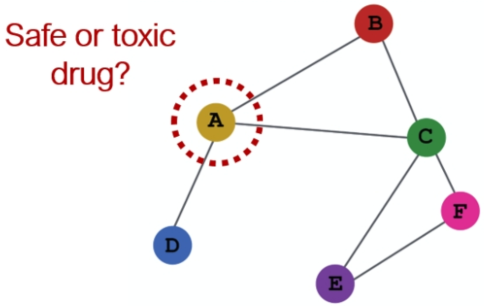
Non-linearity (e.g., ReLU)

Example from: https://cs.stanford.edu/people/jure/teaching.html

Graph Neural Network Training

Supervised Training

Directly train the model for a supervised task (e.g., node classification)



E.g., a drug-drug interaction network

Unsupervised Training

- Train in an unsupervised manner:
  - Use only the graph structure
  - "Similar" nodes have similar embeddings
- Unsupervised loss function can be anything from the last section, e.g., a loss based on
  - Random walks (node2vec, DeepWalk, struc2vec)
  - Graph factorization
  - Node proximity in the graph

Example from: https://cs.stanford.edu/people/jure/teaching.html

Example 1: Predict Twitter (X) Interaction

- Dynamic GNN

TEMPORAL GRAPH NETWORKS FOR DEEP LEARNING ON DYNAMIC GRAPHS

Emanuele Rossi* Twitter
Ben Chamberlain Twitter
Fabrizio Frasca Twitter
Davide Eynard Twitter
Federico Monti Twitter
Michael Bronstein Twitter

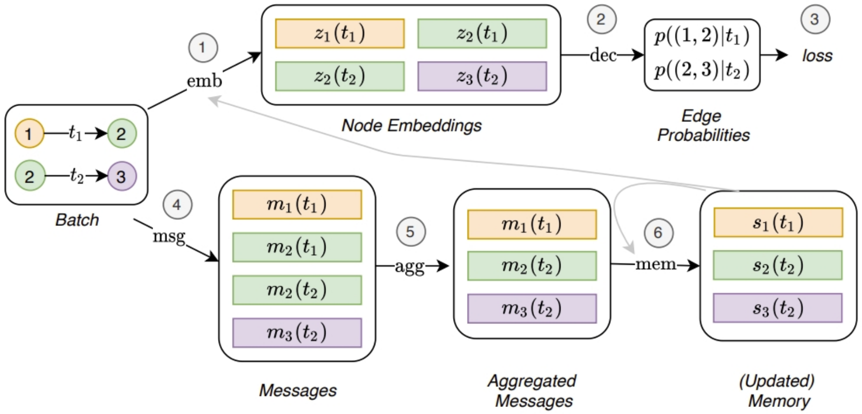


Figure 1: Computations performed by TGN on a batch of time-stamped interactions. *Top:* embeddings are produced by the embedding module using the temporal graph and the node's memory (1). The embeddings are then used to predict the batch interactions and compute the loss (2, 3). *Bottom:* these same interactions are used to update the memory (4, 5, 6). This is a simplified flow of operations which would prevent the training of all the modules in the bottom as they would not receiving a gradient. Section 3.2 explains how to change the flow of operations to solve this problem and figure 2 shows the complete diagram.

Rossi, E., Chamberlain, B., Frasca, F., Eynard, D., Monti, F., & Bronstein, M. (2020). Temporal graph networks for deep learning on dynamic graphs. *arXiv preprint arXiv:2006.10637*.

Example 2: GraphSAGE

- Heterogeneous Nodes and Edges

Inductive Representation Learning on Large Graphs

William L. Hamilton* wleif@stanford.edu
Rex Ying* rexying@stanford.edu
Jure Leskovec jure@cs.stanford.edu

Department of Computer Science
Stanford University
Stanford, CA, 94305

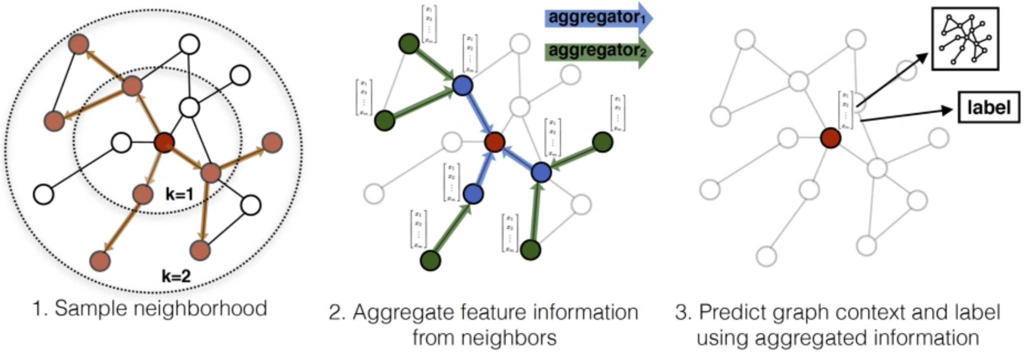


Figure 1: Visual illustration of the GraphSAGE sample and aggregate approach.

Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*, *30*..

Issues with GNN

- Lost global information (Complex system studies are good at dealing with global info)
- Interpretability (Ongoing research)

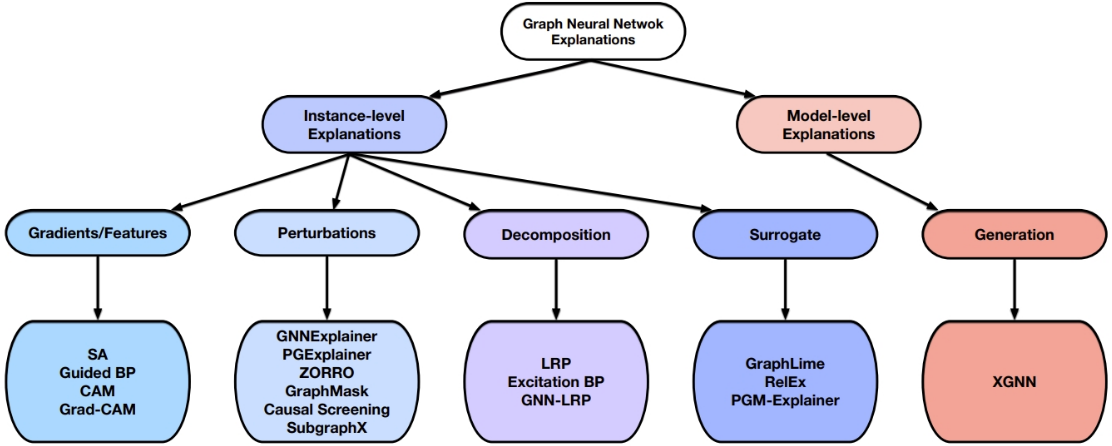


Fig. 1. An overview of our proposed taxonomy. We categorize existing GNN explanation approaches into two branches: instance-level methods and model-level methods. For the instance-level methods, the gradients/features-based methods include SA [54], Guided BP [54], CAM [55], and Grad-CAM [55]; the perturbation-based methods are GNNExplainer [46], PGExplainer [47], ZORRO [56], GraphMask [57], Causal Screening [58], and SubgraphX [48]; the decomposition methods contains LRP [54], [59], Excitation BP [55] and GNN-LRP [60]; the surrogate methods include GraphLime [61], RelEx [62], and PGM-Explainer [63]. For the model-level methods, the only existing approach is XGNN [45].

Yuan, H., Yu, H., Gui, S., & Ji, S. (2022). Explainability in graph neural networks: A taxonomic survey. IEEE transactions on pattern analysis and machine intelligence, 45(5), 5782-5799.

Examples of complex system network studies

- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’networks. nature, 393(6684), 440-442.
- Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. science, 286(5439), 509-512.
- Muscoloni, A., Thomas, J. M., Ciucci, S., Bianconi, G., & Cannistraci, C. V. (2017). Machine learning meets complex networks via coalescent embedding in the hyperbolic space. Nature communications, 8(1), 1615.
- Wang, D., & Barabási, A. L. (2021). The science of science. Cambridge University Press.

Recommended readings

- Grover, A., & Leskovec, J. (2016, August). node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 855-864).
- Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2018). How powerful are graph neural networks?. arXiv preprint arXiv:1810.00826.
- Yuan, H., Yu, H., Gui, S., & Ji, S. (2022). Explainability in graph neural networks: A taxonomic survey. IEEE transactions on pattern analysis and machine intelligence, 45(5), 5782-5799.