Social Network Analysis for Computer Scientists

Frank Takes

LIACS, Leiden University

https://liacs.leidenuniv.nl/~takesfw/SNACS

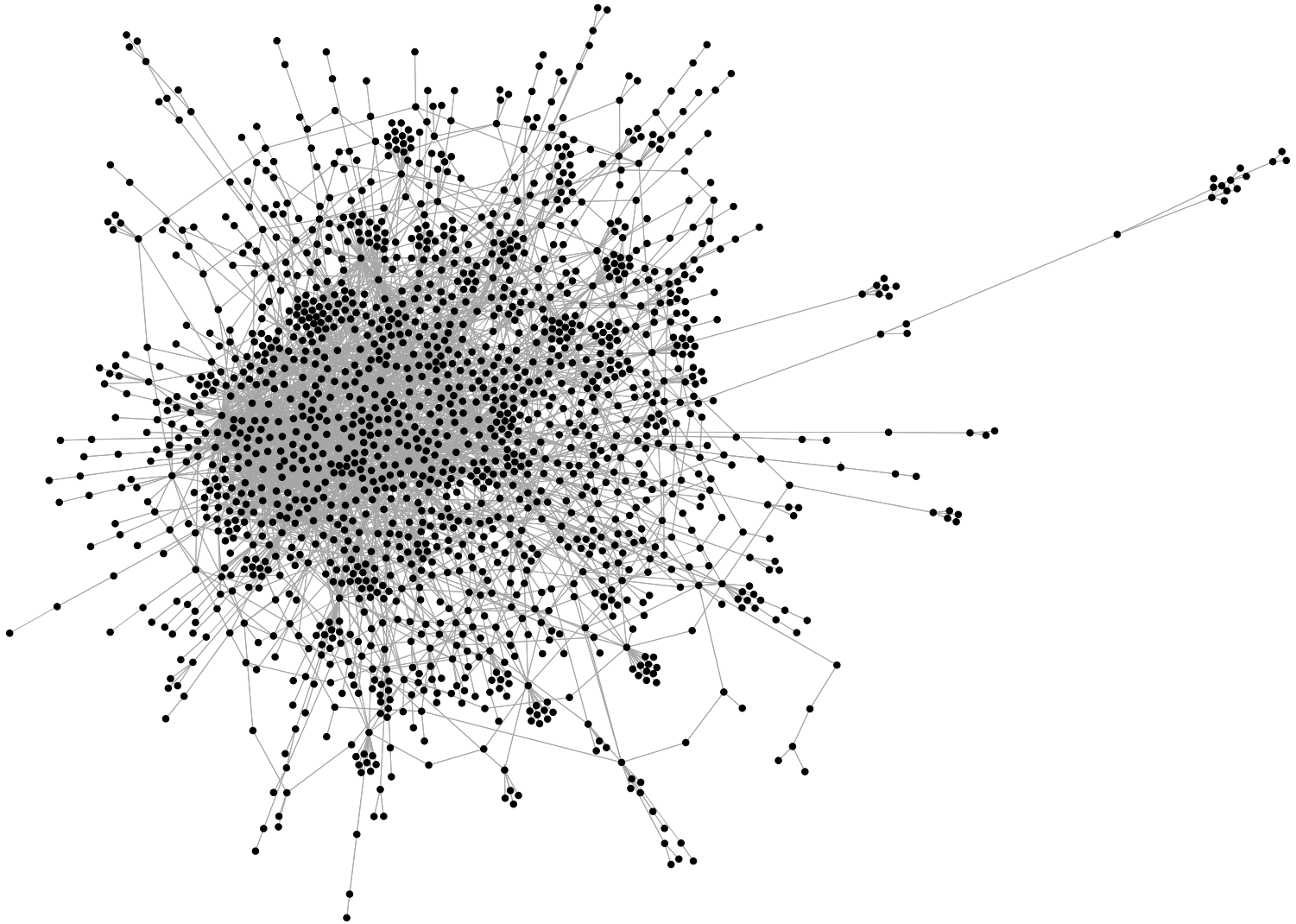
Lecture 9 — Link prediction and network motifs

Today

- Recap
- Link prediction
- Network motifs
- Motifs in corporate networks

Recap

Networks



Real-world networks

1. Sparse networks — density
2. Fat-tailed power-law degree distribution — degree
3. Giant component — components
4. Low pairwise node-to-node distances — distance
5. Many triangles — clustering coefficient

- Many examples: social networks, communication networks, citation networks, collaboration networks, protein interaction networks, information networks, webgraphs, financial networks, ...

Centrality measures

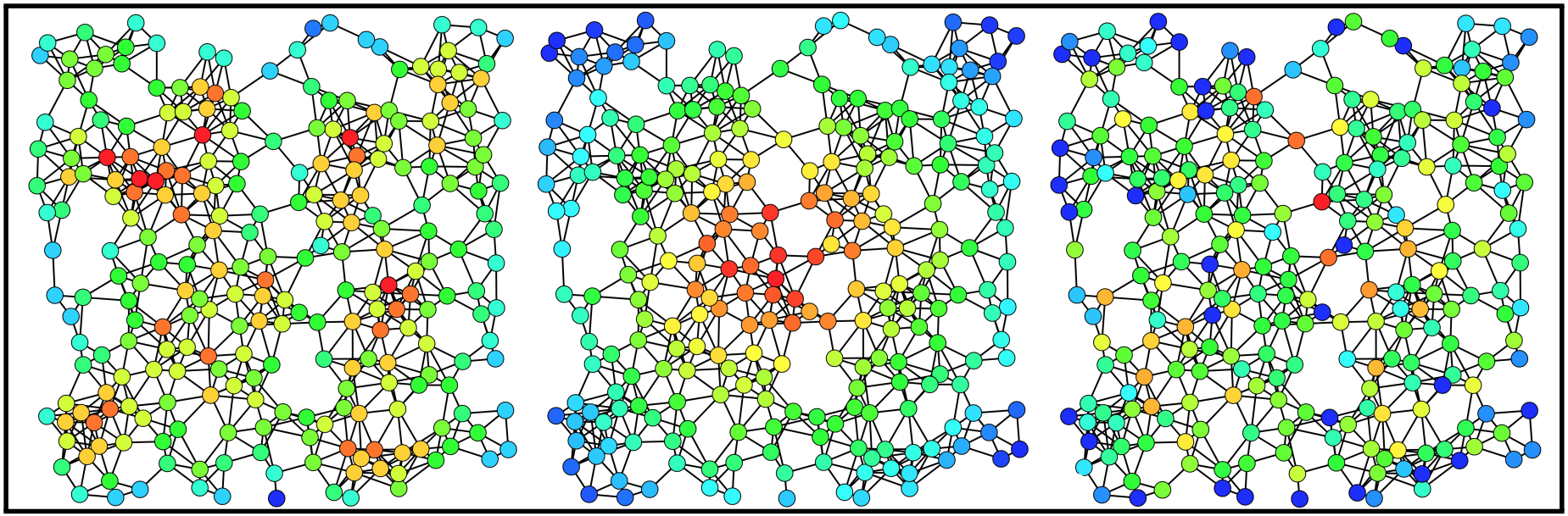
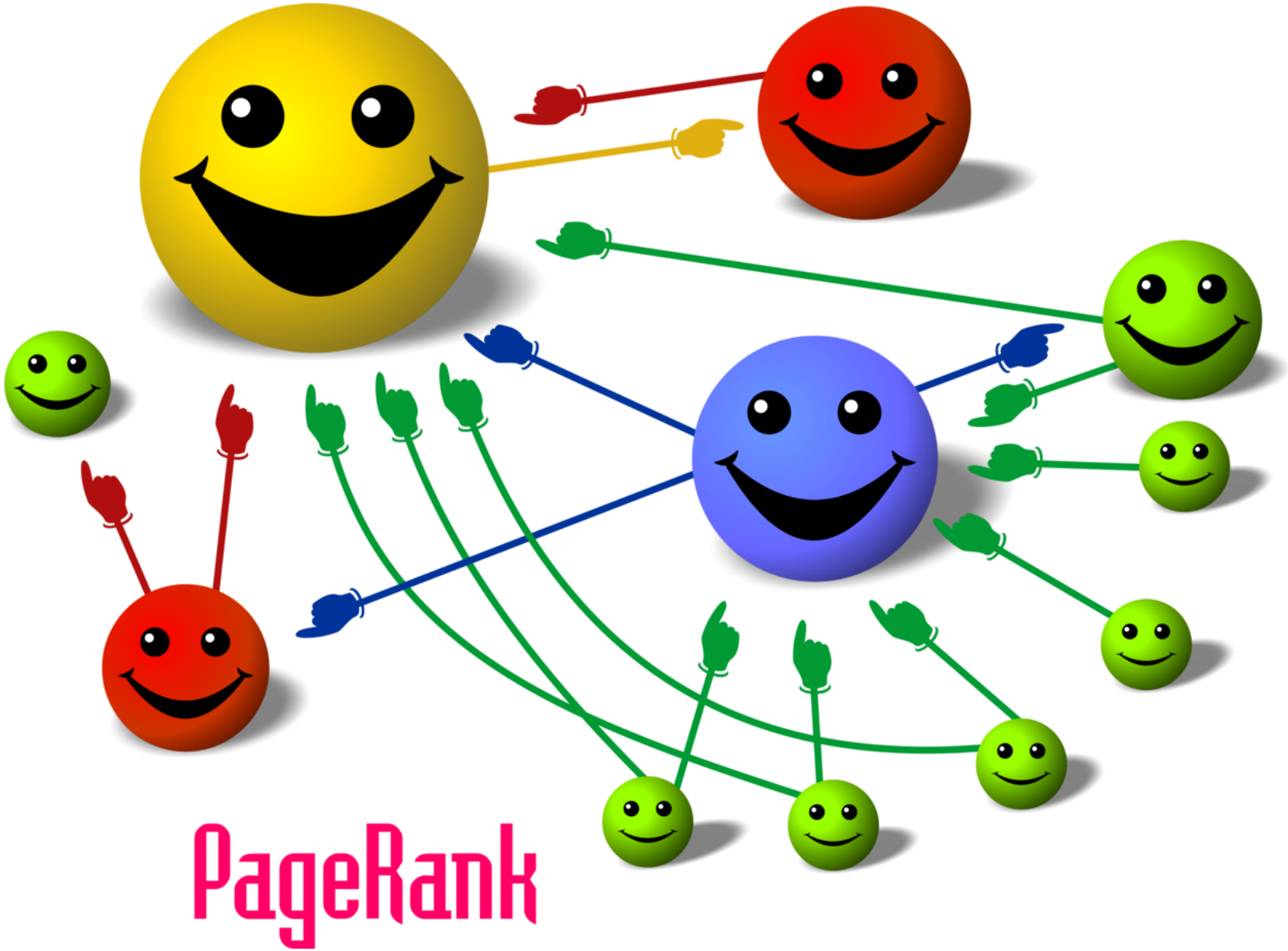
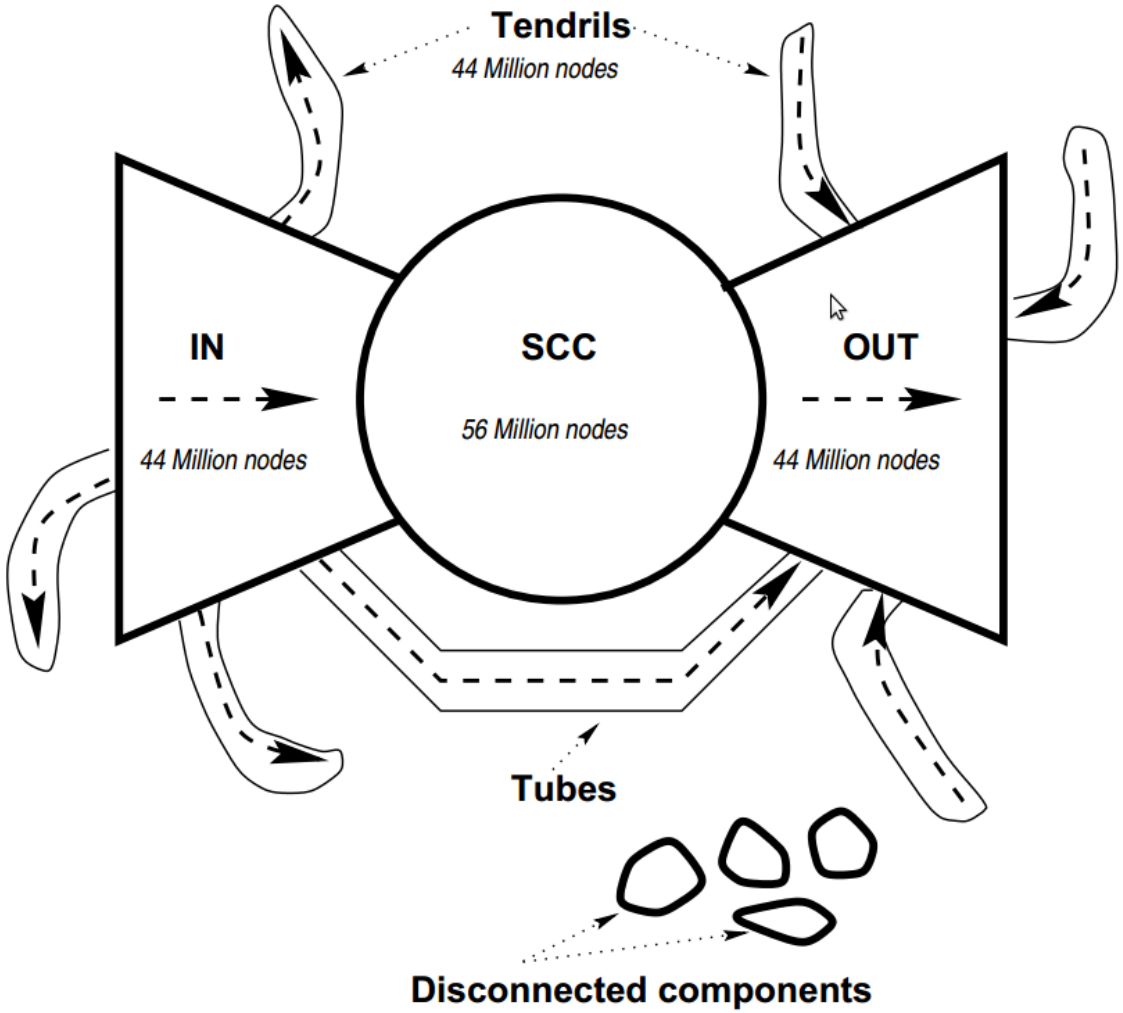


Figure: Degree, closeness and betweenness centrality.

Source: "Centrality" by Claudio Rocchini, Wikipedia `File:Centrality.svg`


PageRank

Webgraph



Broder et al., Graph structure in the web, Computer Networks 33(1): 309–320, 2000.

Attribute assortativity

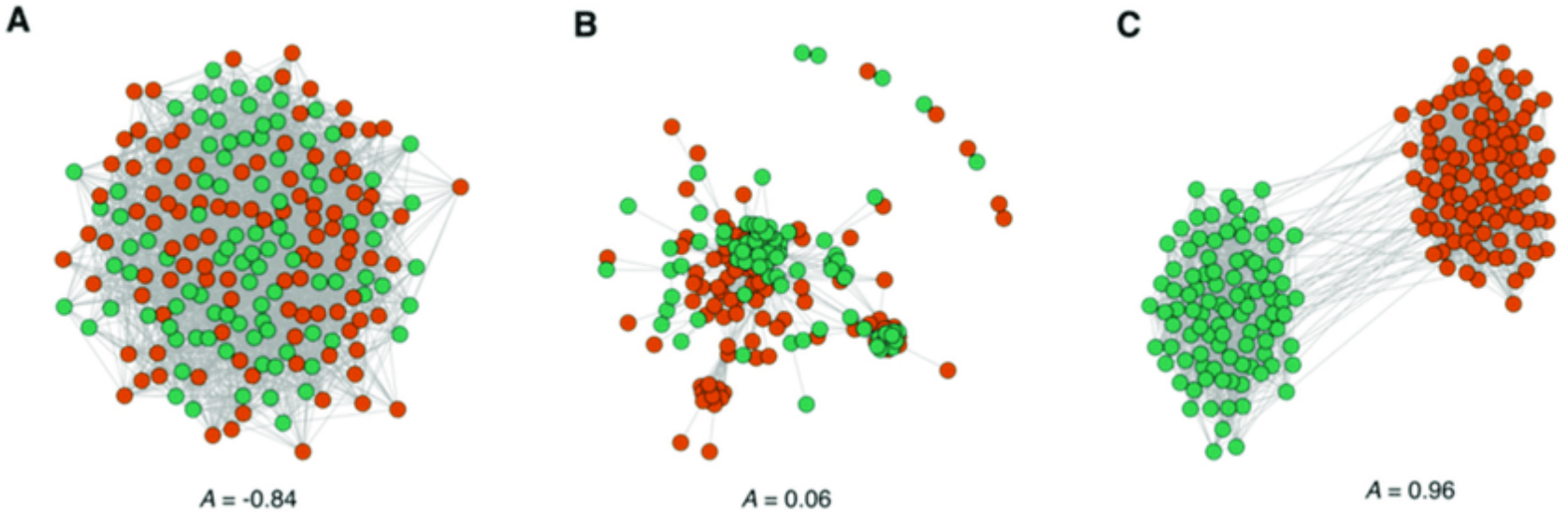


Figure: Attribute assortativity.

Image: Moya-García, A. et al. Identification of New Toxicity Mechanisms ... *Genes*, 13(7), 1292, 2022.

Projection

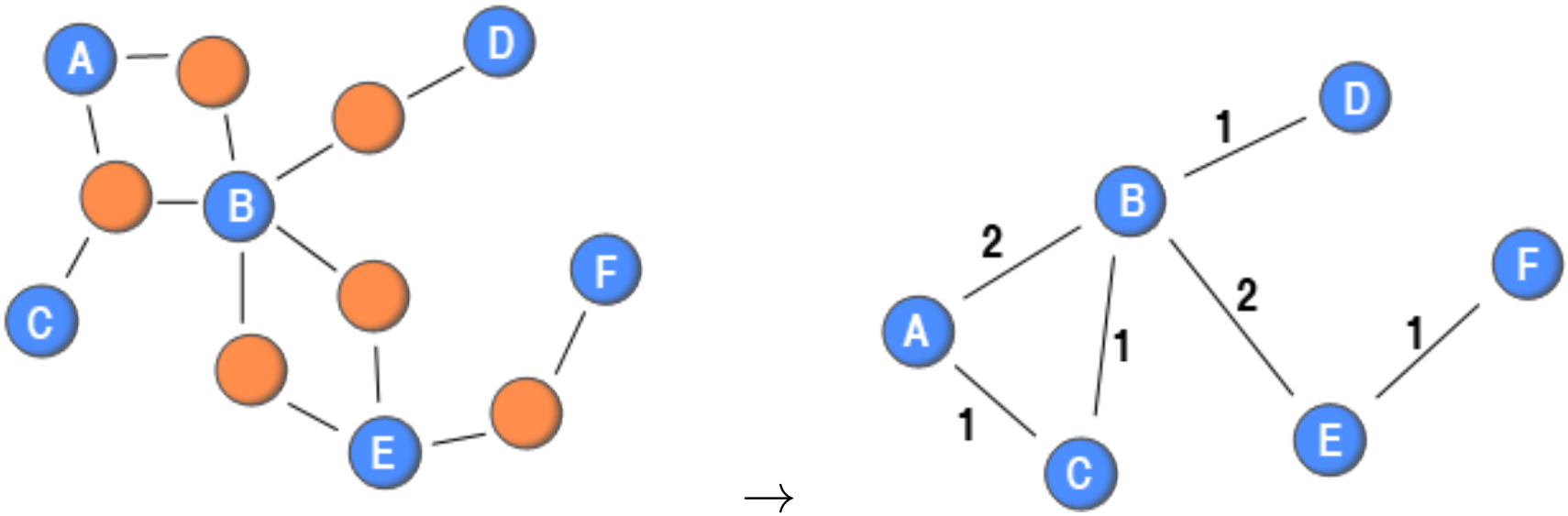


Image: `http://toreopsahl.com`

Community detection

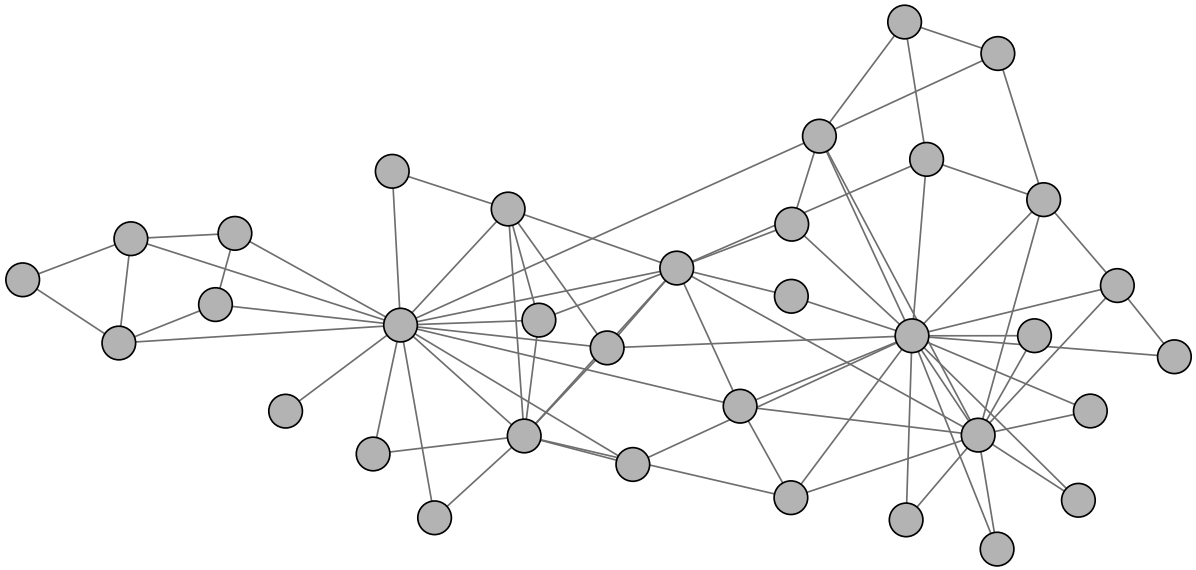


Figure: Communities: node subsets connected more strongly with each other.

- Partitioning a given graph into densely connected groups of nodes

Community detection

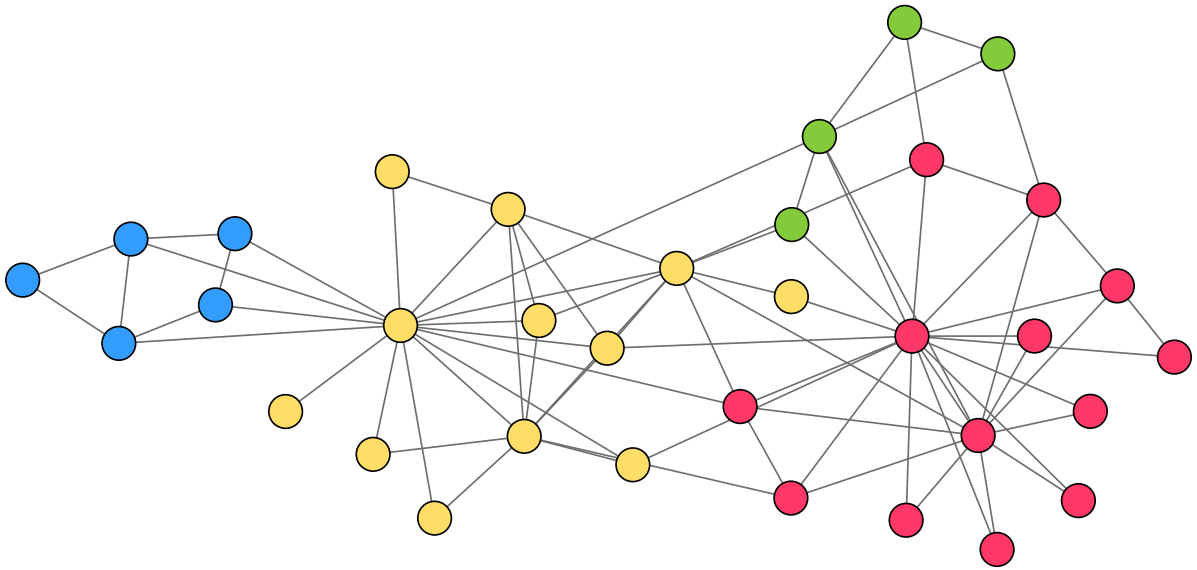


Figure: Communities: node subsets connected more strongly with each other.

- Divisive algorithms (Girvan–Newman), local search algorithms (Louvain) optimizing a quality function (e.g., modularity), label propagation and random walk based approaches.

Core/periphery structure

- Many large networks have a densely connected **core** and a **periphery** of nodes attaching to only a few neighbors
- **Core/periphery structure**: there is **one** dense area
- **Community structure**: there are **several**, with little periphery

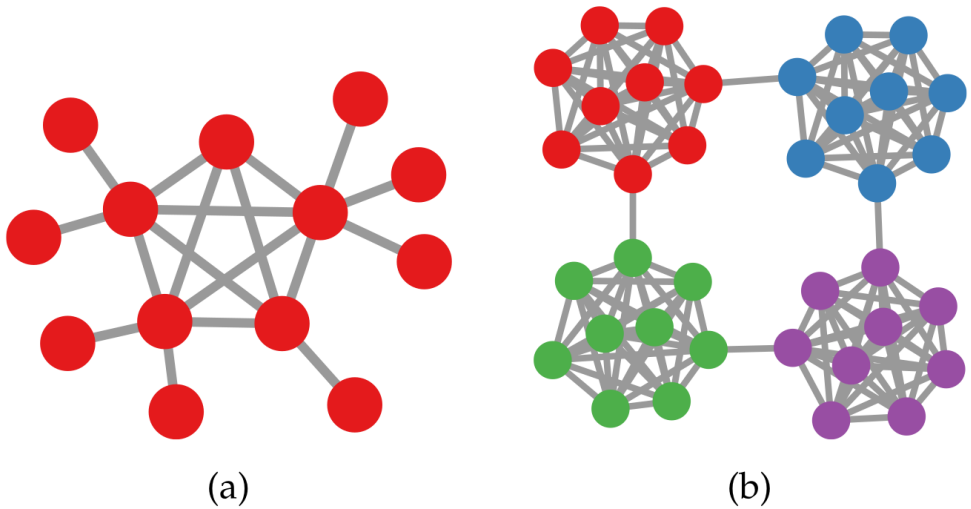
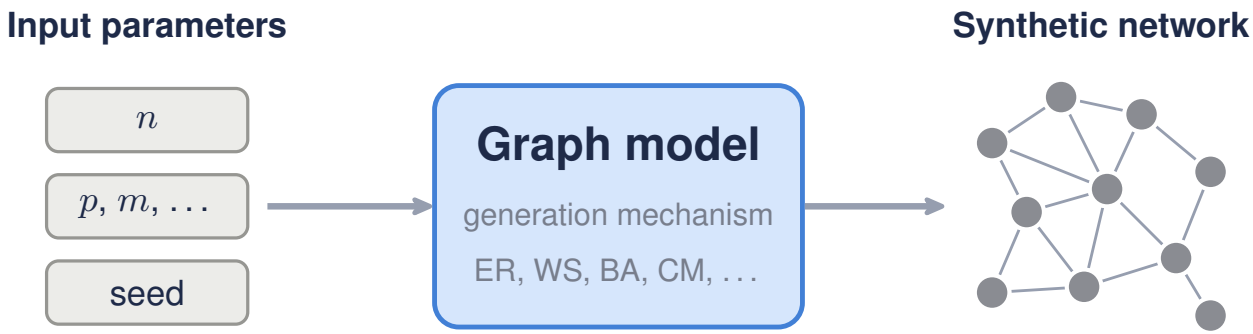
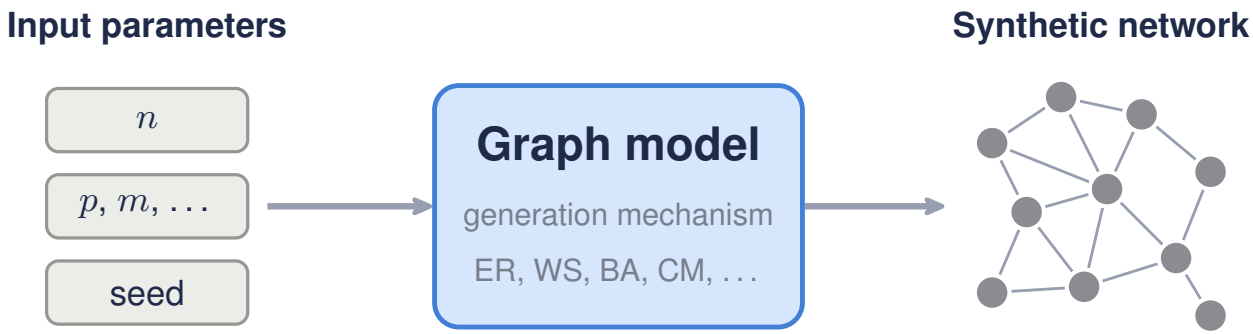


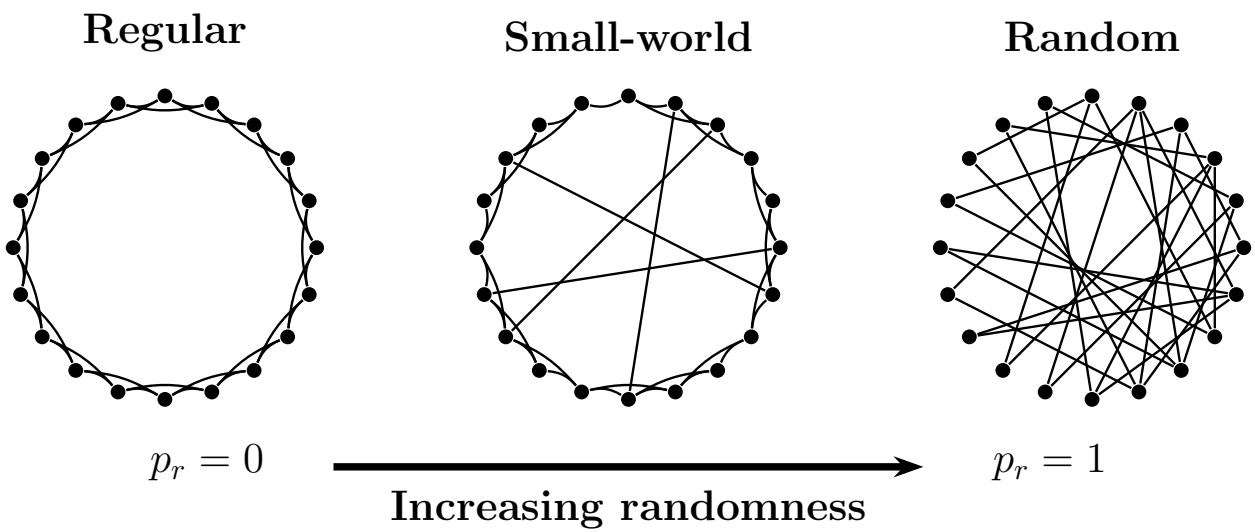
Figure: (a) Core/periphery structure and (b) Community structure.

Network models



Network models





Based on: Watts, D., Strogatz, S. Collective dynamics of 'small-world' networks. Nature 393, 440–442 (1998).

Lancichinetti–Fortunato–Radicchi (LFR) graphs

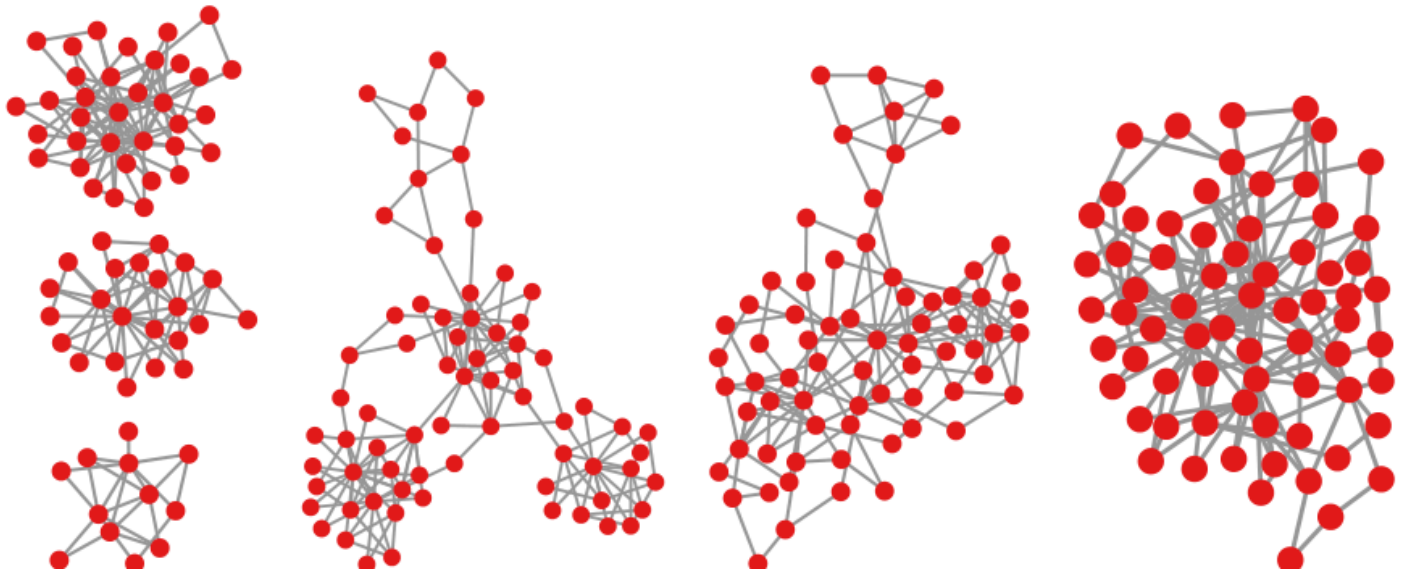


Figure: LFR models with community mixing parameter $\mu = 0, 0.05, 0.2, 1$.

Micro, meso and macro scale

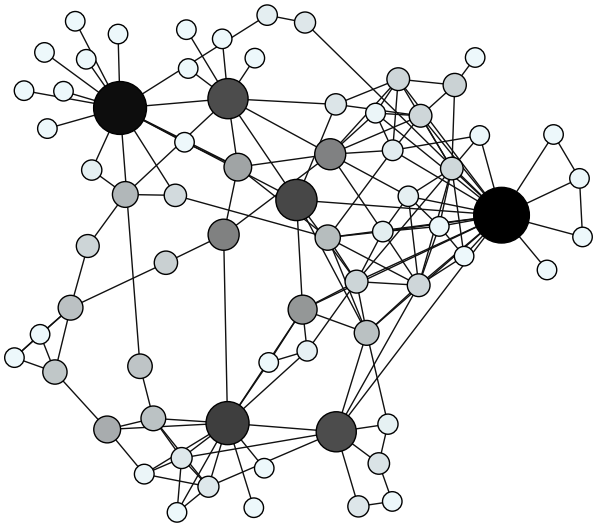


Figure: Micro: node centrality.

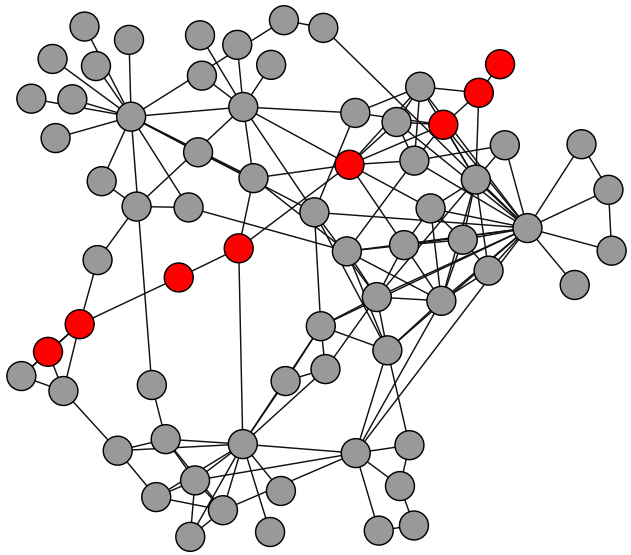


Figure: Macro: average distance.

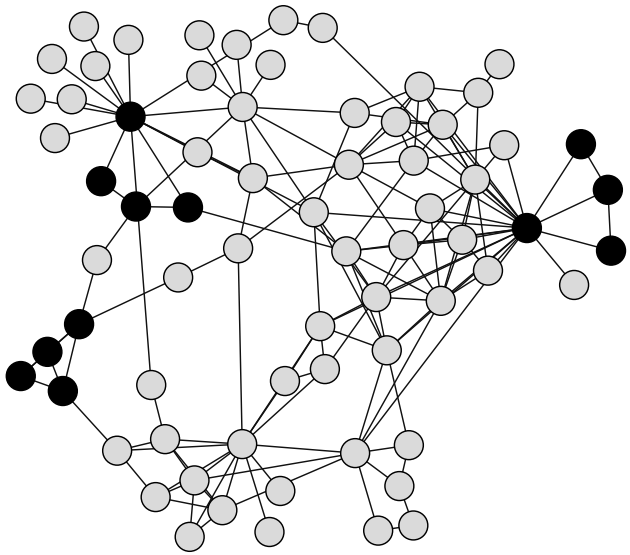


Figure: Meso: network motifs.

Graph evolution

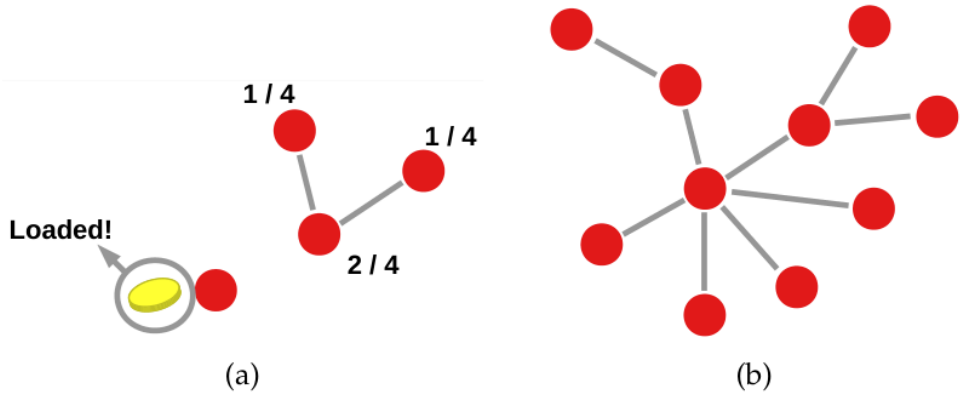


Figure: PA: target node based on degree (a), graph formed through PA (b).

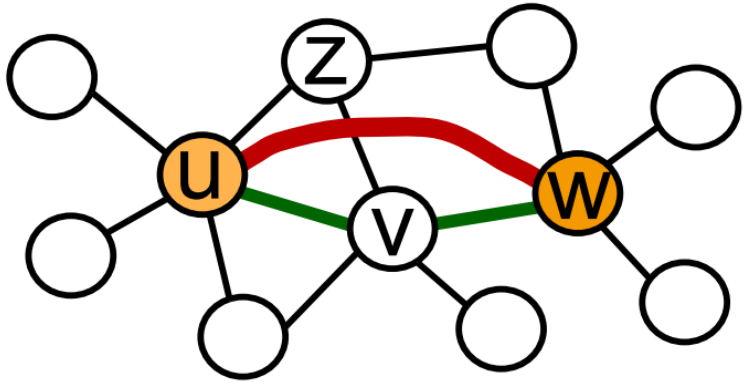
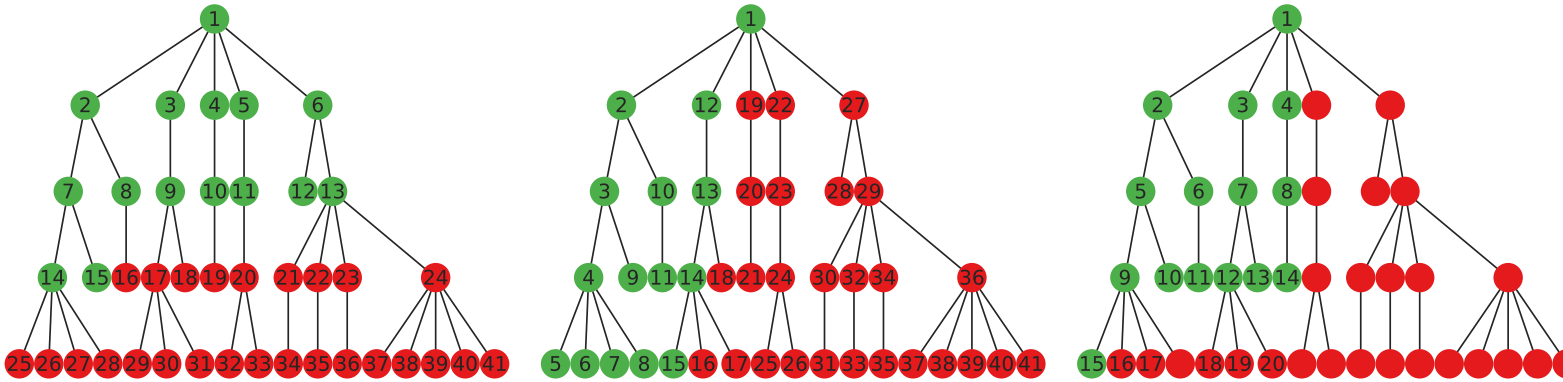


Figure: Triadic closure.

Network sampling



(a) BFS (b) DFS (c) Snowball

(d) Random Walk (e) MHRW (f) Forest Fire

Figure: Six sampling strategies on the same network with a budget of 15 nodes. Labels give the exploration order; nodes are sampled (green) or not (red).

Machine learning and networks

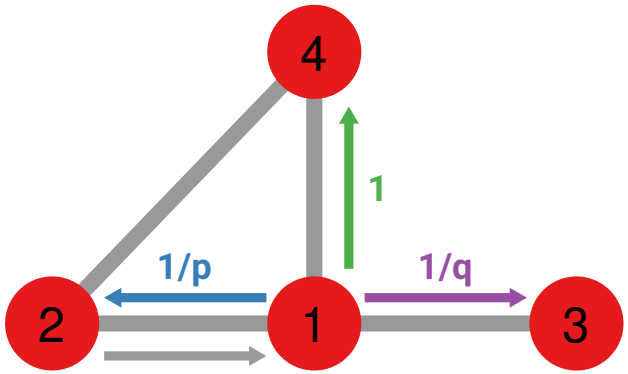


Figure: Node embeddings based on random walks.

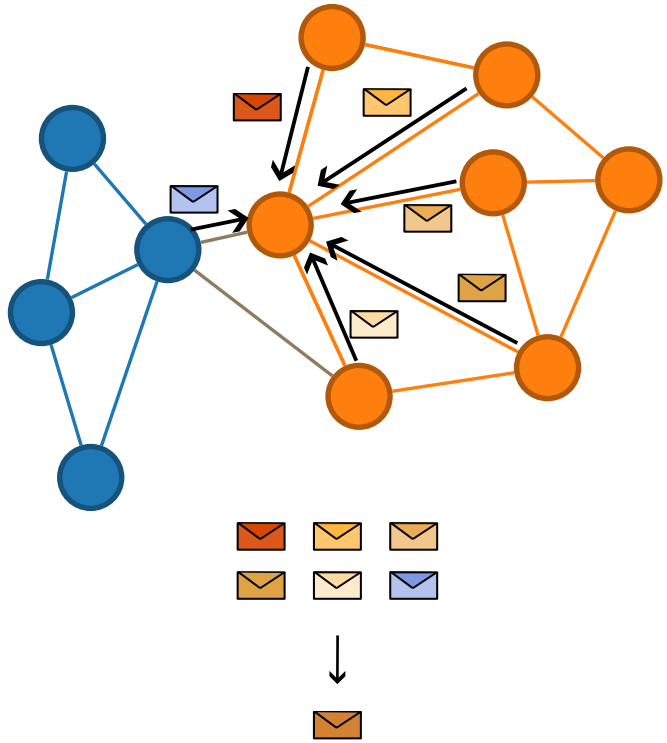


Figure: Message passing in GNNs.

Link prediction

Link prediction

- **Link prediction**: given a network $G = (V, E)$, predict which of the node pairs that are not connected are most likely to be connected

Link prediction

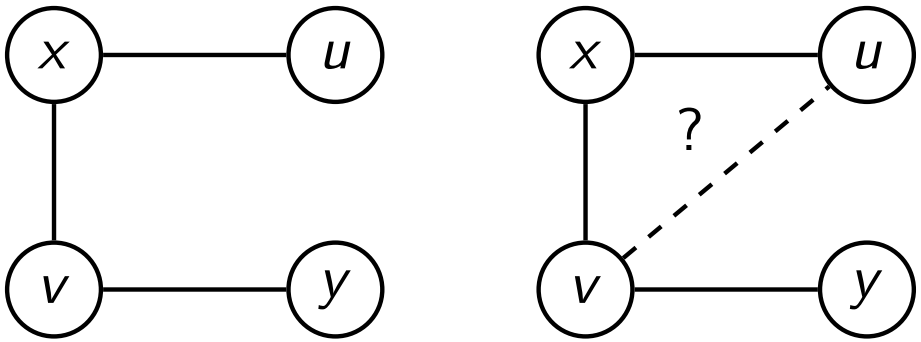
- **Link prediction**: given a network $G = (V, E)$, predict which of the node pairs that are not connected are most likely to be connected
  - **Dynamic**: given snapshot G^t, predict new edges of G^{t+1}

Link prediction

- **Link prediction**: given a network $G = (V, E)$, predict which of the node pairs that are not connected are most likely to be connected
  - **Dynamic**: given snapshot G^t, predict new edges of G^{t+1}
  - **Static**: given an incomplete observation of G, predict **missing** edges

Link prediction

- **Link prediction**: given a network $G = (V, E)$, predict which of the node pairs that are not connected are most likely to be connected
  - **Dynamic**: given snapshot G^t, predict new edges of G^{t+1}
  - **Static**: given an incomplete observation of G, predict **missing** edges
- **Approach**: Formulate a hypothesis on how nodes link, turn it into a **link prediction score** $score(u, v)$ for each unconnected pair, and rank the pairs by score



Link prediction approaches

- **Neighborhood-based**: look at u, v and their neighbors
  - Preferential attachment: the rich get richer
  - Common neighbors: triangles close
  - Jaccard: common neighbors, relative to all neighbors
  - Adamic-Adar: low-degree common neighbors count more
- **Path-based**: look at the network around u and v
  - Katz: many short paths between u and v
- Many more: hierarchical random graphs, stochastic block models, graph evolution rules, node embeddings (last week), ...

Preferential attachment

- Hypothesis: the rich get richer, so high-degree nodes are more likely to attract new edges
- $score(u, v) = deg(u) \cdot deg(v)$

Preferential attachment

- Hypothesis: the rich get richer, so high-degree nodes are more likely to attract new edges
- $score(u, v) = deg(u) \cdot deg(v)$

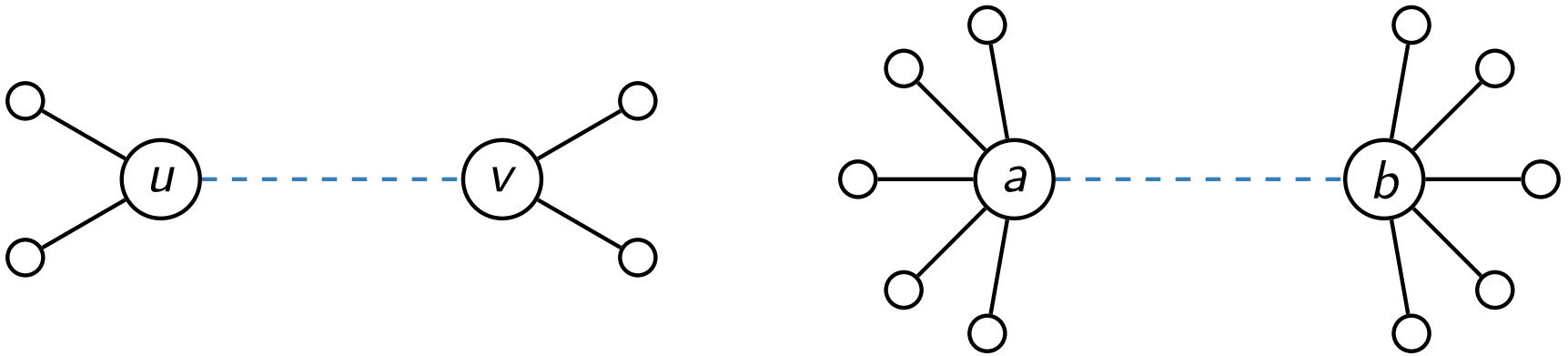


Figure: Candidate links (dashed): $score(u, v) = 2 \cdot 2 = 4$ and $score(a, b) = 5 \cdot 5 = 25$, so the link between the hubs a, b is predicted first.

Common neighbors

- Hypothesis: triangles frequently close, so nodes with many shared neighbors connect
- $score(u, v) = |N(u) \cap N(v)|$

Common neighbors

- Hypothesis: triangles frequently close, so nodes with many shared neighbors connect
- $score(u, v) = |N(u) \cap N(v)|$

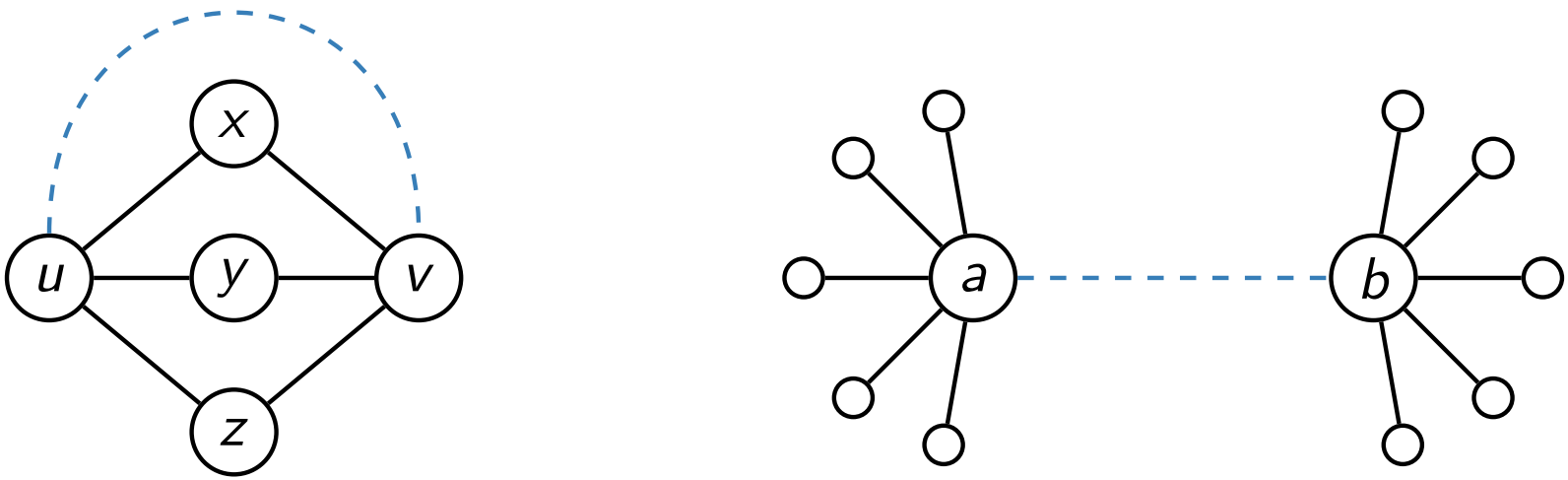


Figure: Candidate links (dashed): $score(u, v) = |\{x, y, z\}| = 3$ and $score(a, b) = 0$, although a and b have a much higher degree than u and v.

Jaccard

- Hubs share many neighbors simply because they have many neighbors
- **Jaccard coefficient**: the common neighbors as a fraction of all neighbors of u and v:

$$score(u, v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$$

Jaccard

- Hubs share many neighbors simply because they have many neighbors
- **Jaccard coefficient**: the common neighbors as a fraction of all neighbors of u and v:

$$score(u, v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$$

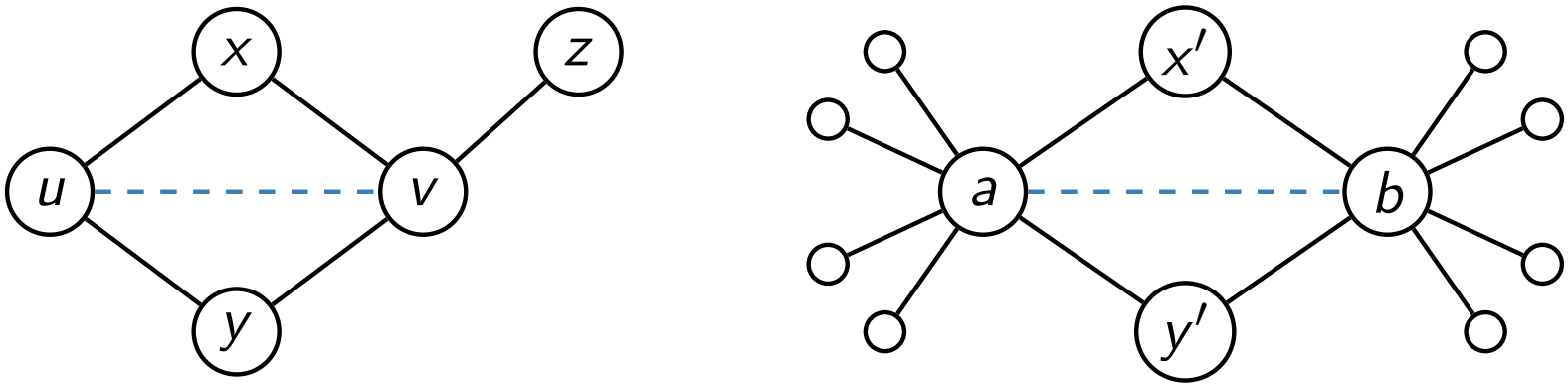


Figure: Both candidate links (dashed) have two common neighbors, but the Jaccard coefficient is $2/3$ for u, v and only $2/10$ for the hubs a, b.

Adamic-Adar

- Hypothesis: a low-degree common neighbor has time to introduce its neighbors to each other, a hub does not
- **Adamic-Adar**: each common neighbor z counts $1/\log deg(z)$ instead of 1:

$$score(u, v) = \sum_{z \in N(u) \cap N(v)} \frac{1}{\log deg(z)}$$

Adamic-Adar

- Hypothesis: a low-degree common neighbor has time to introduce its neighbors to each other, a hub does not
- **Adamic-Adar**: each common neighbor z counts $1/\log deg(z)$ instead of 1:

$$score(u, v) = \sum_{z \in N(u) \cap N(v)} \frac{1}{\log deg(z)}$$



Figure: Both dashed links have one common neighbor, but $score(u, v) = 1/\ln 2 \approx 1.44$ and $score(a, b) = 1/\ln 8 \approx 0.48$, because hub h has degree 8.

Katz

- High for u and v if there are **many short paths** between them
- **Katz score**; $P^{\ell}_{u,v}$ the paths of length ℓ, and $0 < \alpha < 1$:

$$score(u, v) = \sum_{\ell=1}^{\infty} \alpha^{\ell} \, |P^{\ell}_{u,v}|$$

Katz

- High for u and v if there are **many short paths** between them
- **Katz score**; $P^{\ell}_{u,v}$ the paths of length ℓ, and $0 < \alpha < 1$:

$$score(u, v) = \sum_{\ell=1}^{\infty} \alpha^{\ell} \, |P^{\ell}_{u,v}|$$

- Lower α penalizes long paths more; for very small α it is practically common neighbors

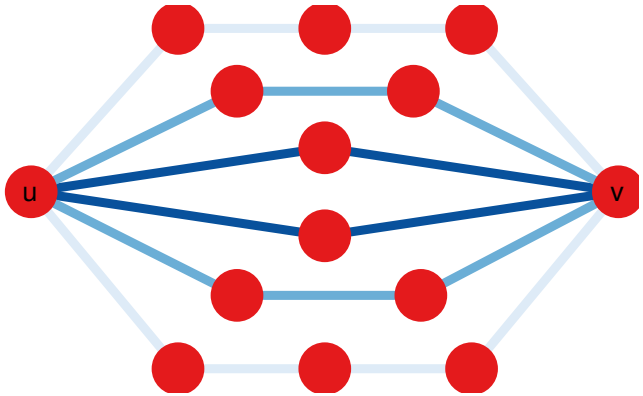


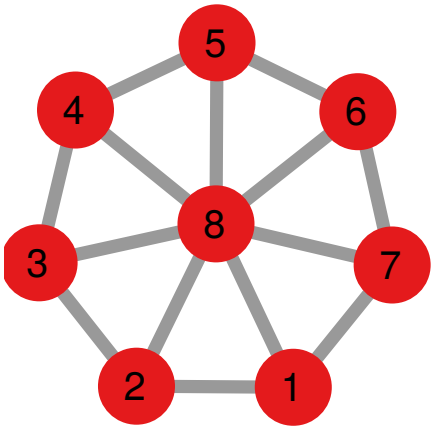
Figure: Paths between u and v: the shorter and darker, the more they contribute.

Link prediction as a machine learning task

- Which predictor works best on a given network? We need to **evaluate** predictions
- Link prediction is a **binary classification** task: for each unconnected pair, does the edge appear, yes or no?
- Predict "yes" when $score(u, v)$ is above a **threshold**
- Better: use many scores as features in a machine learning model (e.g., logistic regression, decision tree, random forest, etc.)

Training and test set

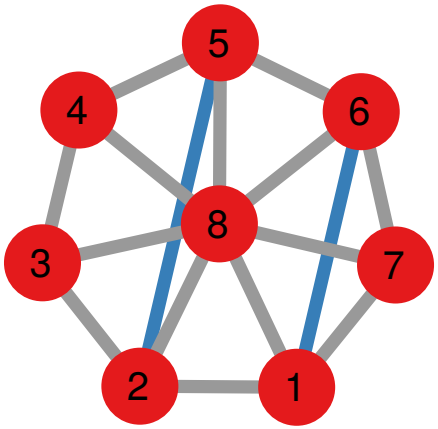
- **Training set**: the edges used to compute the scores
- **Test set**: the edges used to check the predictions
- Training and test set must be **disjoint**



(a) Train

Edge	Score
1, 2	1
1, 3	2
1, 4	1
1, 5	1
1, 6	2
2, 4	2
...	...

(b) Scores



(c) Test

Figure: (a) Training set. (b) Common neighbor scores from (a); pair $1, 2$ is already connected and is discarded. (c) Test set: the new edges (blue).

Splitting the edges

- With **timestamps**: training set has the edges up to time t, test set those after t
- Without: **k-fold cross validation**: split the edges into k parts, use each part once as test set and the other $k-1$ parts as training set (e.g., $k = 10$)

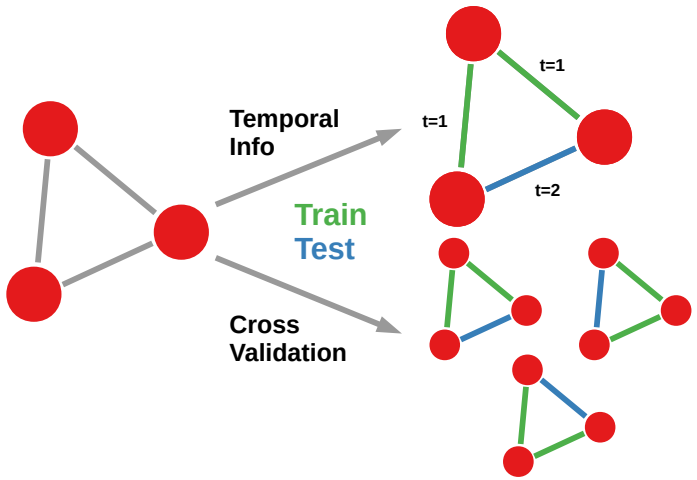
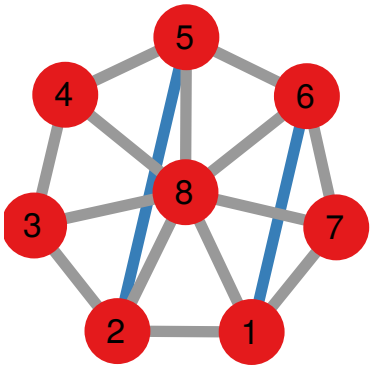


Figure: Splitting the edges of a network (left) into training set (green) and test set (blue), by time (top) or by cross validation (bottom).

Sparsity

- Real-world networks are sparse: almost all pairs are **negatives** (no edge appears), very few are **positives**
- Solution: a **balanced** test set with all positives and an equally large random sample of negatives (negative sampling, lecture 8)



(a) Test

Edge	Prediction	Correct?
1, 3	0	1
1, 4	0	1
1, 5	0	1
1, 6	0	0
2, 4	0	1
2, 5	0	0
2, 6	0	1
...	...	

(b) Scores

Figure: The "always negative" predictor on the test set of the previous slides: correct for about 85% of the pairs, without ever predicting a link.

Confusion matrix

- **True positive** (TP): predicted, and appeared
- **False positive** (FP): predicted, but did not appear
- **True negative** (TN): not predicted, and did not appear
- **False negative** (FN): not predicted, but appeared
- **Confusion matrix**: the four counts in a grid
- **Accuracy** $\frac{TP+TN}{TP+TN+FP+FN}$ hides difference between FP and FN

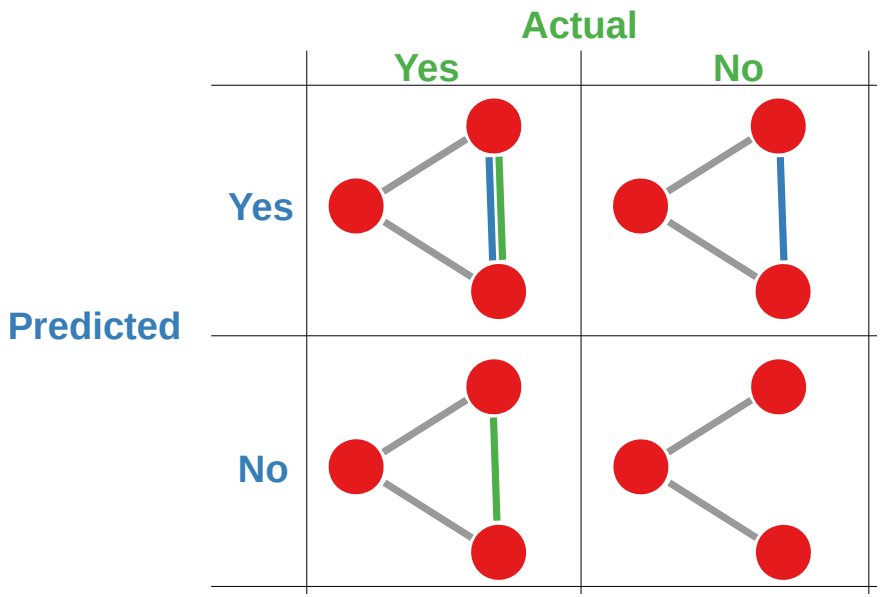


Figure: Predicted links in blue, actual new links in green. Clockwise from the top left: TP, FP, TN, FN.

ROC curve

- **ROC curve** (receiver operating characteristic): FPR on the horizontal axis, TPR on the vertical axis, one point per threshold
- Random guessing gives the diagonal: TPR and FPR grow equally fast

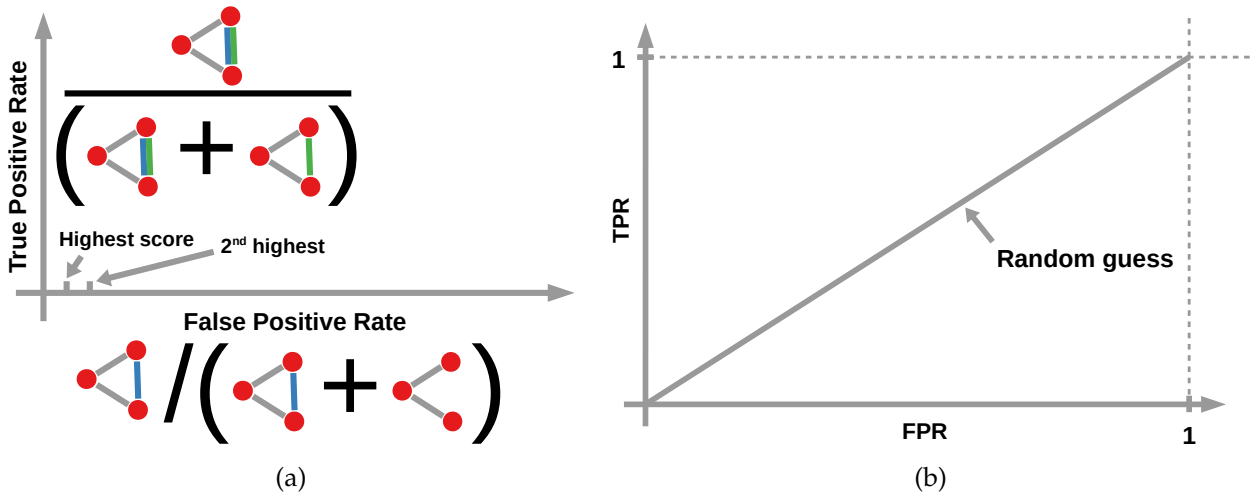


Figure: (a) Going down the ranking, each pair moves the curve up (TP) or to the right (FP). (b) Random guessing.

ROC curves and AUC

- The higher the curve, the more true positives for the same number of false positives
- **AUC**: area under ROC curve; 0.5 for random guessing, 1 for perfect predictor

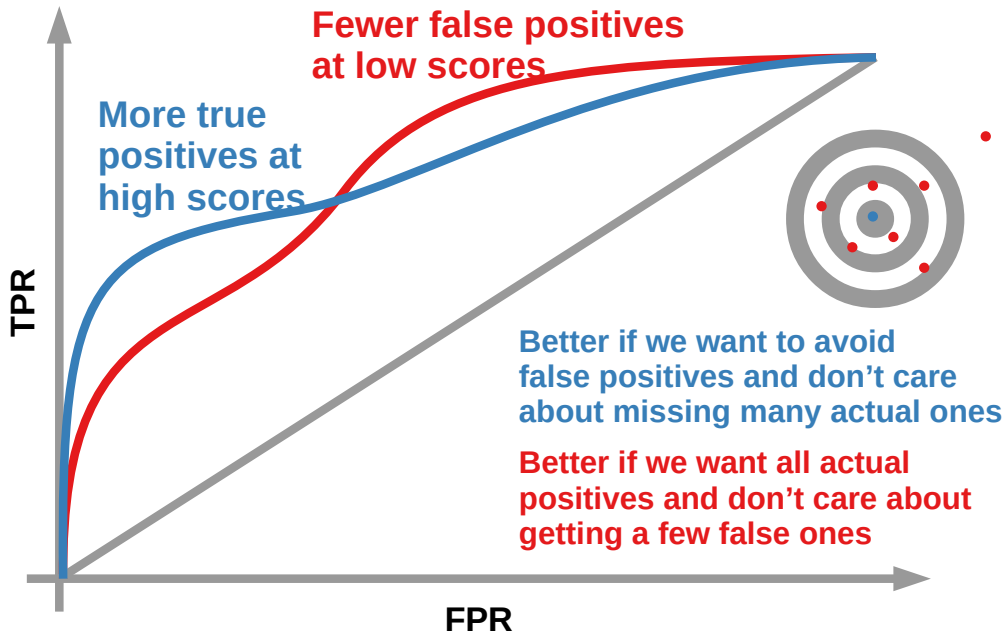


Figure: ROC curves of two predictors and of random guessing (gray).

Explainable and efficient link prediction in real-world networks

J.E. van Engelen, H.D. Boekhout and F.W. Takes, Explainable and Efficient Link Prediction in Real-World Network Data, in *Proceedings of the 15th International Symposium on Intelligent Data Analysis (IDA 2016)*, Lecture Notes in Computer Science 9897: 295-307, Springer, 2016.

Link prediction

- **Link prediction problem**: given a network $G_t = (V_t, E_t)$, denoting the network at time t, predict the newly formed links in the evolved network $G_{t'} = (V_{t'}, E_{t'})$ at time $t' > t$, i.e., predict the contents of $E_{t'} \setminus E_t$.

Link prediction

- **Link prediction problem**: given a network $G_t = (V_t, E_t)$, denoting the network at time t, predict the newly formed links in the evolved network $G_{t'} = (V_{t'}, E_{t'})$ at time $t' > t$, i.e., predict the contents of $E_{t'} \setminus E_t$.
- Applicable to weighted and unweighted, directed and undirected networks
- Supervised learning problem
- Features based on the structure of the network
- Train on first 95%, test on last 5% (randomized)
- Validate result using AUROC

Feature set goals

- **efficient** in terms of time complexity;
- **accurate** in its future link predictions;
- **explainable** in its performance based on simple features;
- **consistent** in its accuracy relative to larger feature sets across networks;
- **generic**, yielding reliable results across a broad range of networks.

Link prediction features

- Compute features for each possible future edge $(i, j) \notin E_t$
- **Node features**: degree, volume (total weight)
- **Neighborhood features**: neighbor count, common neighbor count, transitive common neighborhood, Jaccard coefficient, preferential attachment, and others
- **Path features**: shortest path length, number of shortest paths, restricted Katz measure, and others

Efficient Feature Set (EFS)

- Large number of features
- Black box type of approach
- Cover individual, local and global properties
- Explainable result
- **Efficient Feature Set**

Feature	Variant	Complexity	EFS
Node features			
Degree (source)	-	$O(1)$	✓
Degree (source)	d_{in}	$O(1)$	
Degree (source)	d_{out}	$O(1)$	
Degree (target)	-	$O(1)$	✓
Degree (target)	d_{in}	$O(1)$	
Degree (target)	d_{out}	$O(1)$	
Volume (source)	-	$O(m/n)$	
Volume (source)	d_{in}	$O(m/n)$	✓
Volume (source)	d_{out}	$O(m/n)$	✓
Volume (target)	-	$O(m/n)$	
Volume (target)	d_{in}	$O(m/n)$	✓
Volume (target)	d_{out}	$O(m/n)$	✓
Neighbourhood features			
Total neighbours	-	$O(m/n)$	
Total neighbours	Γ_{in}	$O(m/n)$	
Total neighbours	Γ_{out}	$O(m/n)$	
Common neighbours	-	$O(m/n)$	✓
Common neighbours	Γ_{in}	$O(m/n)$	
Common neighbours	Γ_{out}	$O(m/n)$	
Transitive comm. neigh.	-	$O(m/n)$	
Jaccard Coeff.	-	$O(m/n)$	✓
Jaccard Coeff.	Γ_{in}	$O(m/n)$	
Jaccard Coeff.	Γ_{out}	$O(m/n)$	
Transitive Jacc. Coeff.	-	$O(m/n)$	✓
Adamic/Adar	-	$O(m/n)$	
Preferential attachment	-	$O(1)$	
Preferential attachment	Γ_{in}	$O(1)$	
Preferential attachment	Γ_{out}	$O(1)$	
Opposite direction link	-	$O(1)$	✓
Path features			
Shortest path length	-	$O(m+n)$	✓
Num. shortest paths	$\ell_{\text{MAX}} = 3$	$O(m+n)$	
Restricted Katz measure	$\ell_{\text{MAX}} = 3$, $\beta = 0.05$	$O(m+n)$	
PropFlow	$\ell_{\text{MAX}} = 3$	$O(m+n)$	✓

Datasets

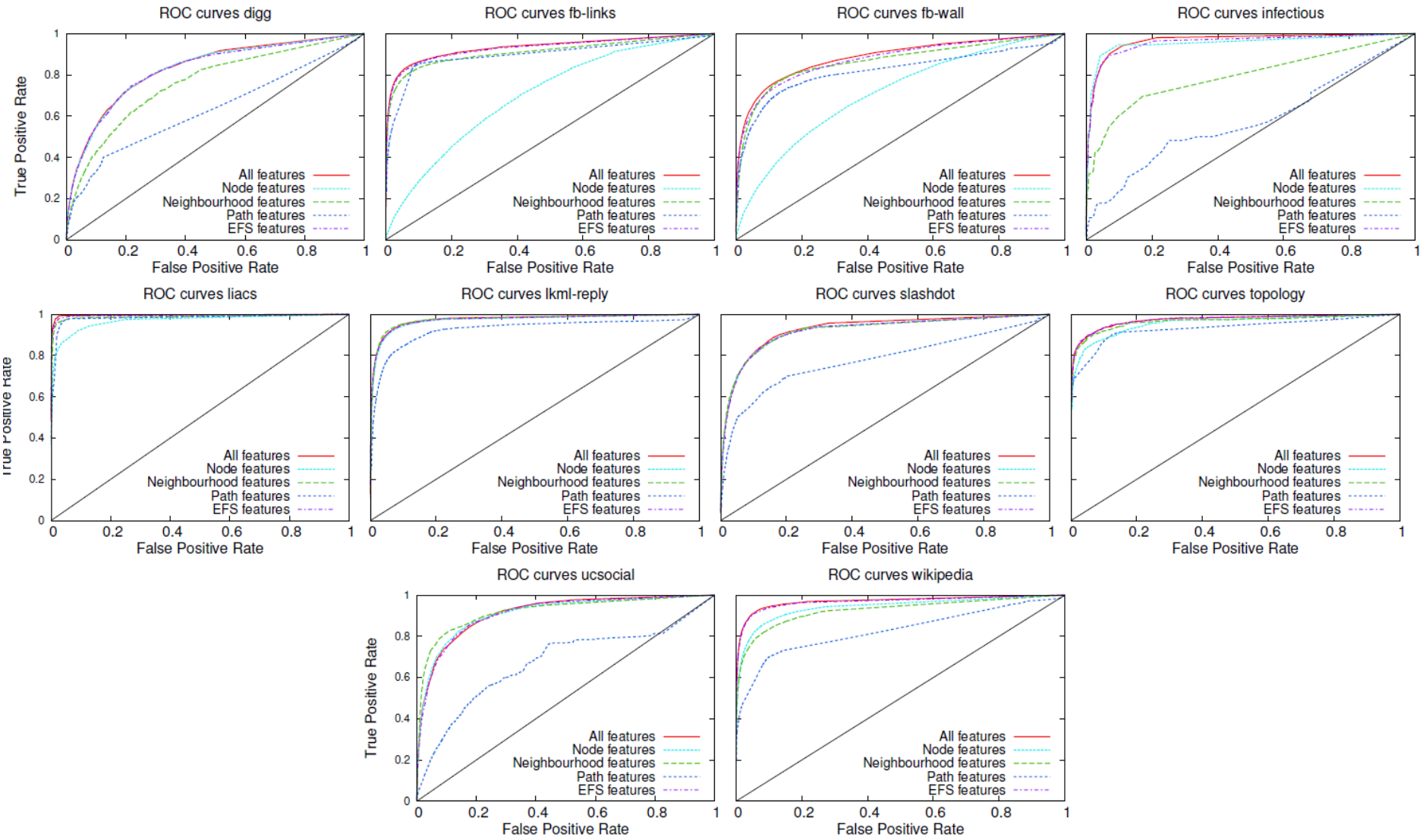
Table: Characteristics of network data sets used for testing.

Data set	Nodes	Links	CC	Type	Dist	3N
digg	30,398	86,404	0.01	+ D	4.68	45%
fb-links	63,731	817,035	0.22	- U	4.31	88%
fb-wall	46,952	274,086	0.11	+ D	5.71	61%
infectious	410	2,765	0.46	+ U	3.57	83%
liacs	1,036	4,650	0.84	+ U	3.86	100%
lkml-reply	27,927	242,976	0.30	+ D	5.19	99%
slashdot	51,083	131,175	0.02	+ D	4.59	75%
topology	34,761	107,720	0.29	+ U	3.78	97%
ucsocial	1,899	20,296	0.11	+ D	3.07	99%
wikipedia	100,312	746, 114	0.21	- D	3.83	89%

Experiments

- Large candidate set of size $(|V| \times |V - 1|) - |E|$
- Restrict based on maximum distance a new edge bridges
- Class imbalance
- Randomly leave out edges in training to get to $9:1$ ratio
- Measure result using AUROC
- Determine difference between All features, Node features, Neighborhood Features and EFS

Results


ROC curves digg
ROC curves fb-links
ROC curves fb-wall
ROC curves infectious
ROC curves liacs
ROC curves lkml-reply
ROC curves slashdot
ROC curves topology
ROC curves ucsocial
ROC curves wikipedia
True Positive Rate
False Positive Rate
All features
Node features
Neighbourhood features
Path features
EFS features

Results

Features	*digg*	*fb-links*	*fb-wall*	*infectious*	*liacs*	*lkml*	*slashdot*	*topology*	*ucsocial*	*wikipedia*
All	0.830	0.933	0.887	0.967	0.997	0.975	0.928	0.967	0.913	0.970
Node	**0.827**	0.700	0.710	**0.955**	0.969	0.971	**0.922**	0.949	0.911	**0.941**
Neighbourhood	0.761	**0.911**	**0.866**	0.794	**0.986**	**0.974**	0.920	**0.961**	**0.920**	0.926
Path	0.632	0.897	0.819	0.579	0.979	0.925	0.777	0.940	0.673	0.827
EFS	0.825	0.930	0.876	0.958	0.995	0.973	0.921	0.965	0.910	0.967
EFS Performance	**99.4%**	**99.6%**	**98.8%**	**99.1%**	**99.8%**	**99.8%**	**99.2%**	**99.8%**	**99.7%**	**99.7%**

Table: AUROC for each network and each set of features. EFS Performance lists performance of EFS relative to All features.

Conclusion

- A **small set** of 12 cheap features (EFS) reaches 98.8–99.8% of the AUROC of all features

Conclusion

- A **small set** of 12 cheap features (EFS) reaches 98.8–99.8% of the AUROC of all features
- **No single feature class** is best everywhere: node features win on some networks, neighborhood features on others

Conclusion

- A **small set** of 12 cheap features (EFS) reaches 98.8–99.8% of the AUROC of all features
- **No single feature class** is best everywhere: node features win on some networks, neighborhood features on others
- Combining node, neighborhood and path features gives **consistent** results across networks

Conclusion

- A **small set** of 12 cheap features (EFS) reaches 98.8–99.8% of the AUROC of all features
- **No single feature class** is best everywhere: node features win on some networks, neighborhood features on others
- Combining node, neighborhood and path features gives **consistent** results across networks
- Simple features keep the predictions **explainable** and efficient to compute

Graph Mining

Graph mining

- **Graph mining**: applying machine learning and data mining techniques to networks
- Today: **frequent subgraph mining**, finding the subgraphs that occur often
  - Network motifs
  - Graph isomorphism
  - Frequent itemset mining
  - Transactional graph mining: many small graphs
  - Single graph mining: one large graph
- Lecture 8: node embeddings and graph neural networks

Machine learning on graphs

- Classical approach: describe each node by features such as degree and betweenness, and feed these to a classifier
- Same for entire networks: degree distribution, clustering coefficient, or the **counts of small subgraphs**

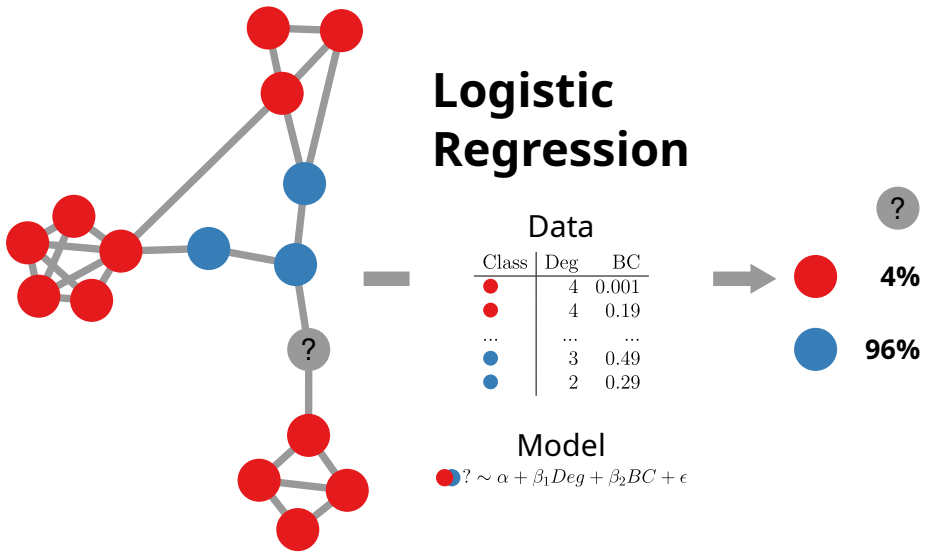


Figure: A logistic regression model trained on the degree and betweenness of the labeled nodes predicts that the gray node is blue.

Network motifs

- **Network motif**: small subgraph with a given topology

Network motifs

- **Network motif**: small subgraph with a given topology
- Real-world networks have **overexpressed** motifs, such as triangles
- Is a motif significant? Compare its count with the count in a null model, e.g., the configuration model
- Counting motifs requires deciding whether two graphs are the same

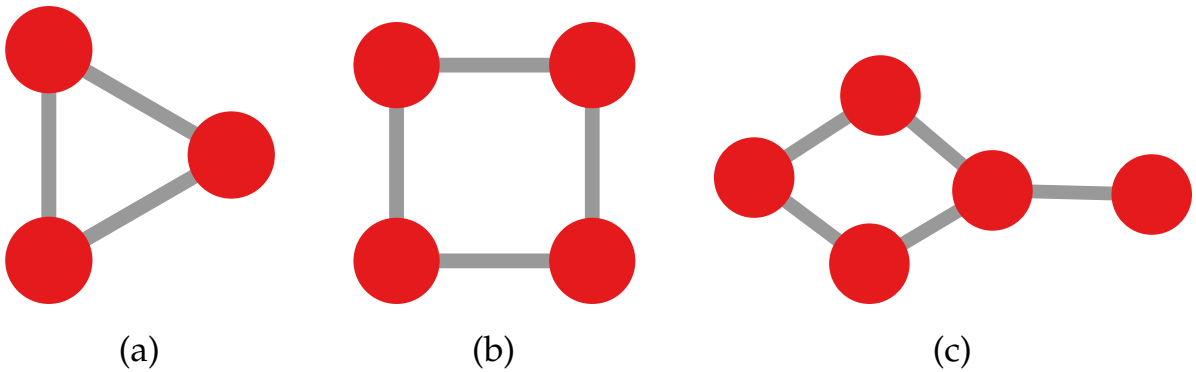
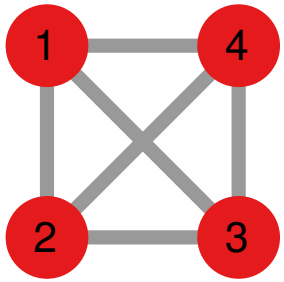


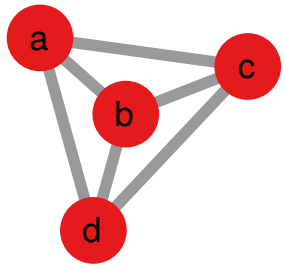
Figure: Three motifs: (a) a triangle, (b) a square and (c) a square with a dangling edge.

Graph isomorphism

- **Graph isomorphism**: G_1 and G_2 are isomorphic if there is a one-to-one mapping between their nodes such that two nodes are connected in G_1 exactly when their images are connected in G_2
- **Subgraph isomorphism**, does G_1 contain a copy of G_2, is NP-complete
- Either fast but sometimes wrong (approximate), or always right but slow (exact)



(a)



(b)

$G1$	$G2$
1	c
2	a
3	d
4	b

(c)

Figure: (a, b) Two 4-cliques. (c) An edge-preserving mapping from (a) to (b): the graphs are isomorphic.

Frequent itemset mining

- Input: a collection of sets of items, the **itemsets**
- **Support** of a set of items: the number of itemsets containing it
- **Frequent itemset mining**: find all sets of items with support at least a given **support threshold**

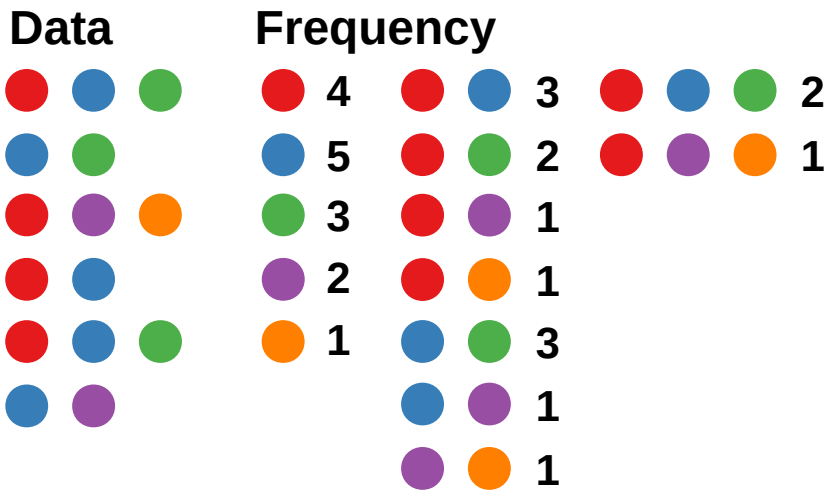


Figure: A collection of itemsets (left) and the support of each set of items (right).

Apriori

- Support is **anti-monotone**: set is at most as frequent as its subsets
- **Apriori** algorithm: once a set is infrequent, skip all sets that contain it

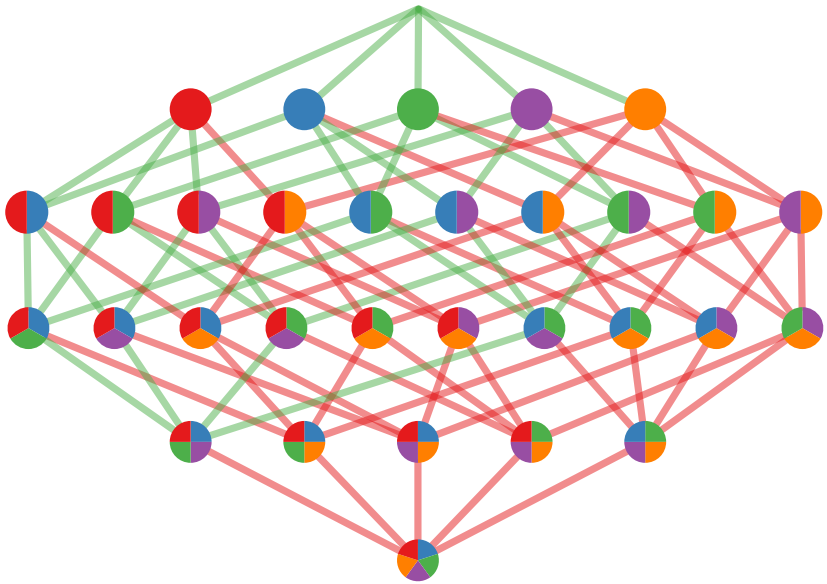


Figure: All subsets of five items. Orange is infrequent, so Apriori does not follow the red links, which lead to sets containing orange.

Transactional graph mining

- **Transactional graph mining**: itemsets become graphs, items become motifs
- Input: a database of many graphs
- Support of a motif: the number of graphs that contain it
- Again anti-monotone, so Apriori-style pruning works

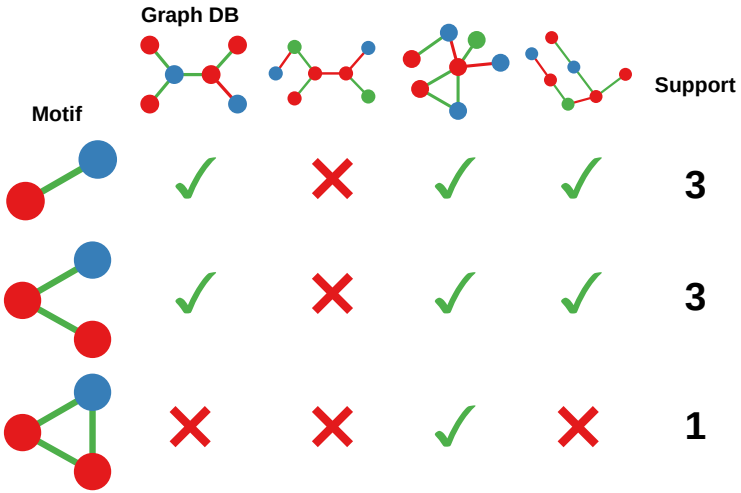


Figure: Support of three motifs (left) in a database of four graphs (top).

Network motifs

in multiplex networks of corporations

F.W. Takes, W.A. Kosters, B. Witte and E.M. Heemskerk, Multiplex Network Motifs as Building Blocks of Corporate Networks, *Applied Network Science* 3: 39, Springer, 2018.

Overview

- Introduction to corporate networks
- Multiplex networks
- Network motif detection algorithms
- Multiplex adjustments
- Experiments and results
- Conclusion and future work

Multiplex corporate networks

- **Nodes**: firms/corporations

Multiplex corporate networks

- **Nodes**: firms/corporations
- **Links**:
  - Ownership (directed)



  - Board interlocks (undirected)



Multiplex corporate networks

- **Nodes**: firms/corporations
- **Links**:
  - Ownership (directed)



  - Board interlocks (undirected)



- Typical small-world networks
- Insight in power and control of corporations and institutions

Multiplex corporate networks

- **Nodes**: firms/corporations
- **Links**:
  - Ownership (directed)



  - Board interlocks (undirected)

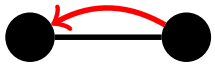


- Typical small-world networks
- Insight in power and control of corporations and institutions
- Interlayer assortativity: different types of edges are related to each other.

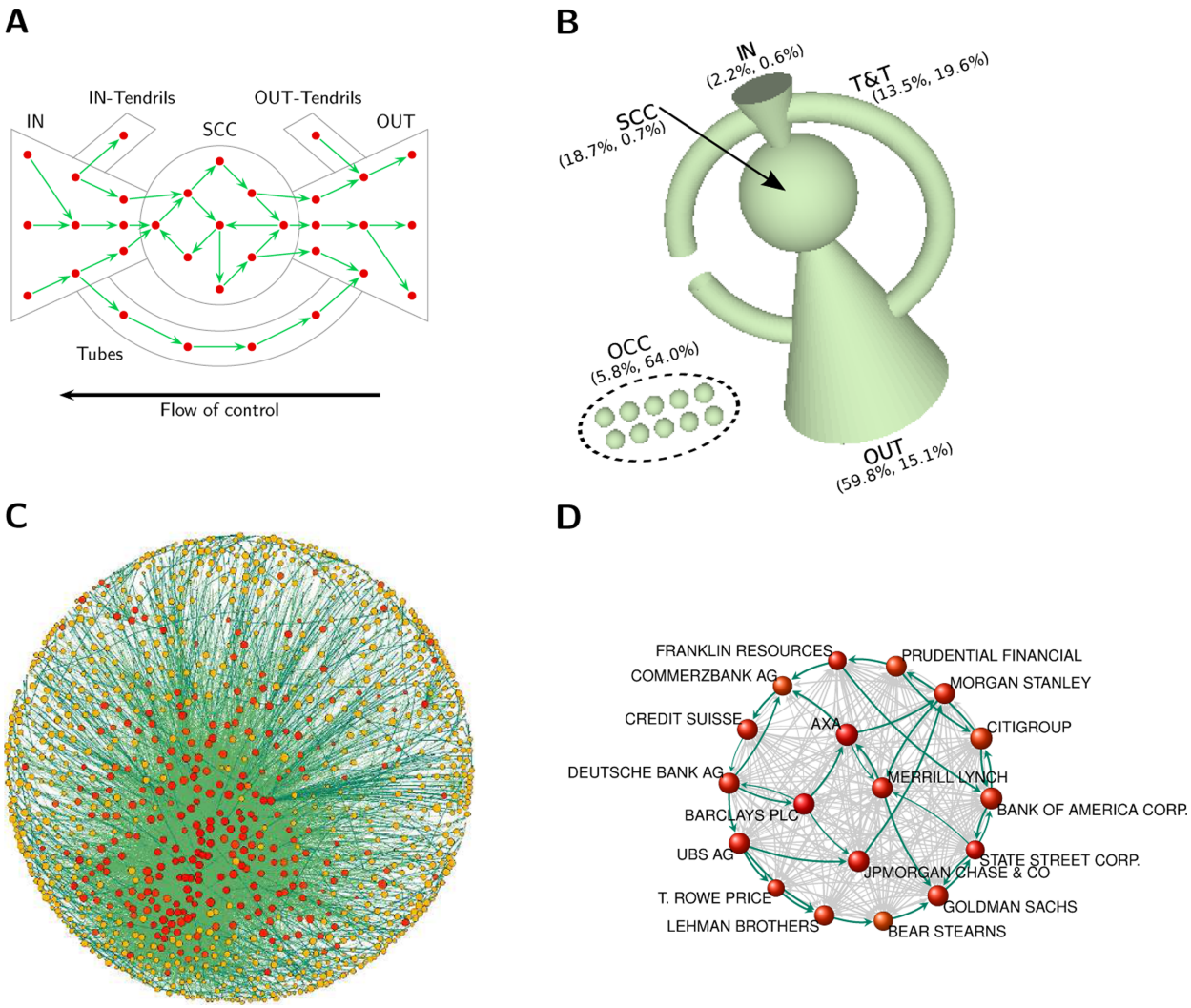
Multiplex corporate networks

- **Nodes**: firms/corporations
- **Links**:
  - Ownership (directed)

    
  - Board interlocks (undirected)

    
- Typical small-world networks
- Insight in power and control of corporations and institutions
- Interlayer assortativity: different types of edges are related to each other. Frequently: 

Ownership



S. Vitali, J.B. Glattfelder and S. Battiston, S, The network of global corporate control, *PloS one* 6(10), e25995, 2011.

Board interlocks

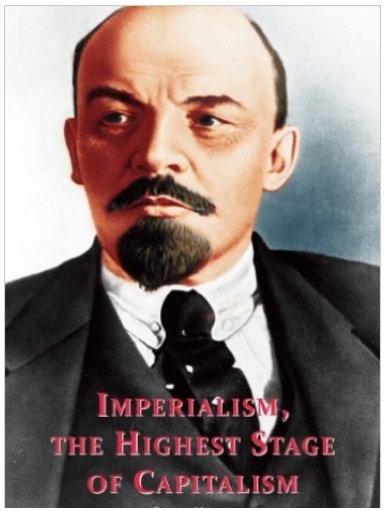
- **Board interlock**: there is a relationship between firms because they share a board member or director

Board interlocks

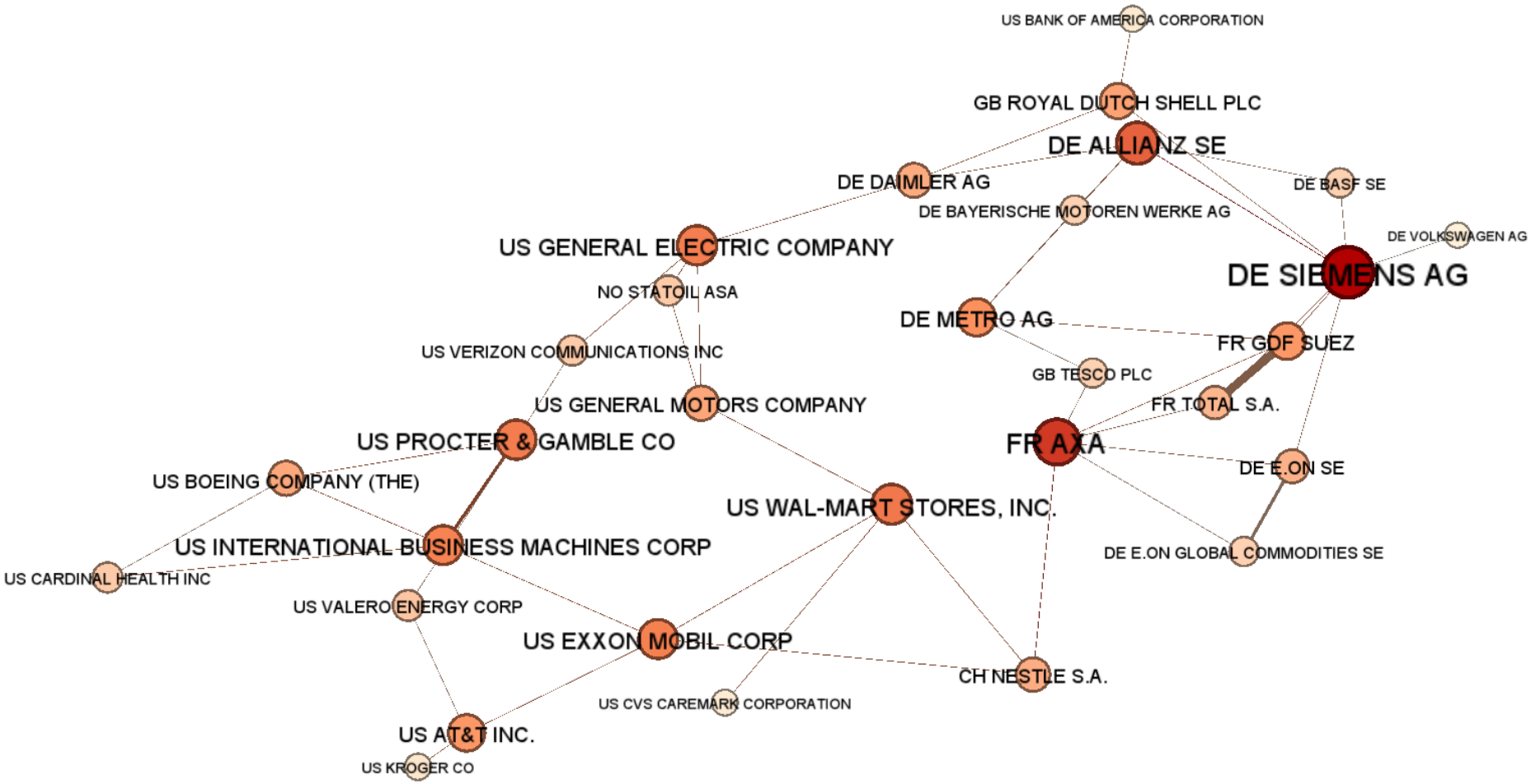
- **Board interlock**: there is a relationship between firms because they share a board member or director
- Reasons: collusion, monitoring, legitimacy, career advancement, social cohesion, access to resources, corporate control

Board interlocks

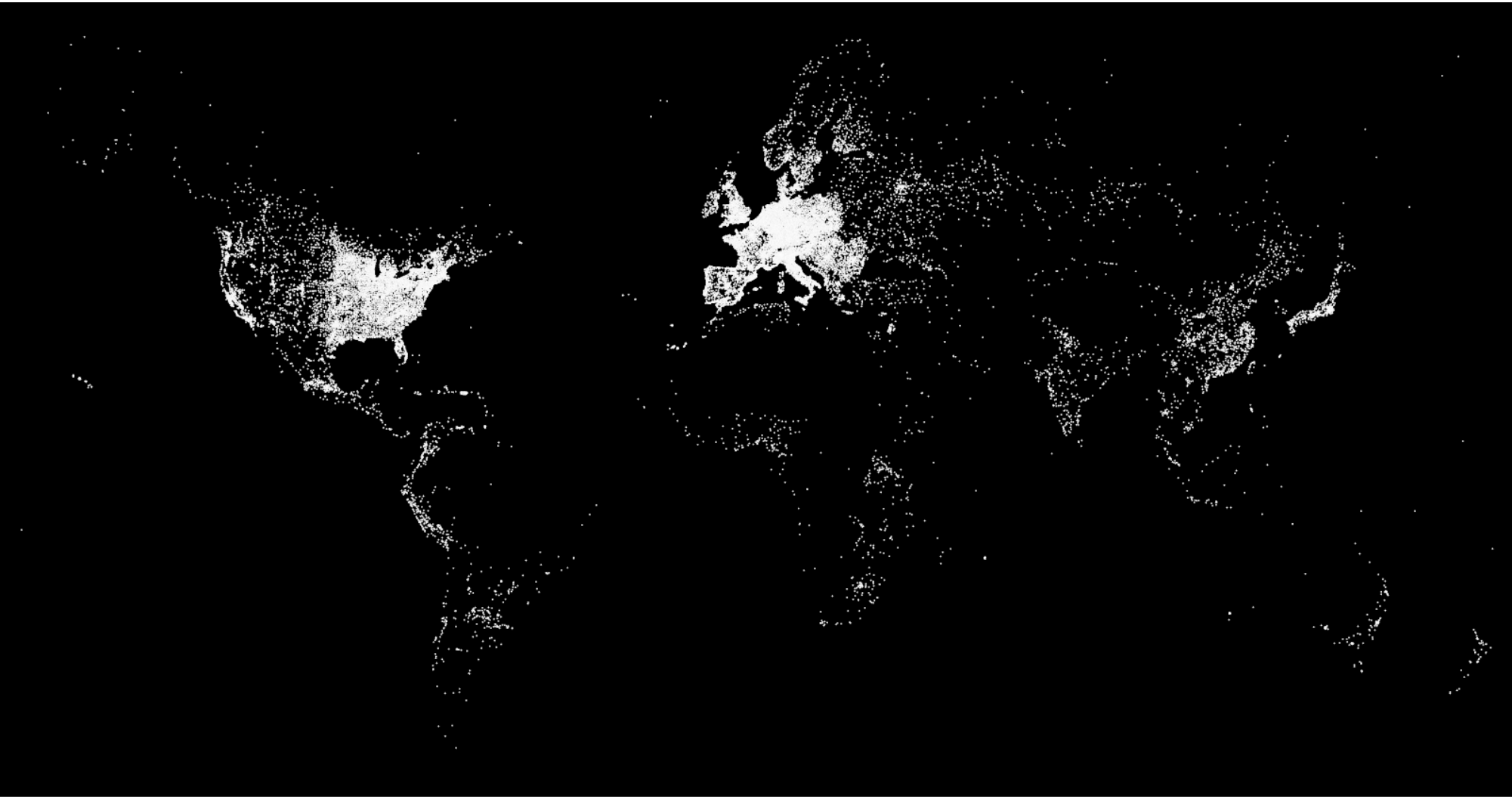
- **Board interlock**: there is a relationship between firms because they share a board member or director
- Reasons: collusion, monitoring, legitimacy, career advancement, social cohesion, access to resources, corporate control
- Vladimir I. Lenin, *Imperialism, The Highest Stage of Capitalism*, 1916.
- "... a personal union, so to speak, is established between the banks and the biggest industrial and commercial enterprises, the merging of one with another through the acquisition of shares, through the appointment of bank directors to the Supervisory Boards (or Boards of Directors) of industrial and commercial enterprises, and vice versa."



Board interlock networks



Corporations

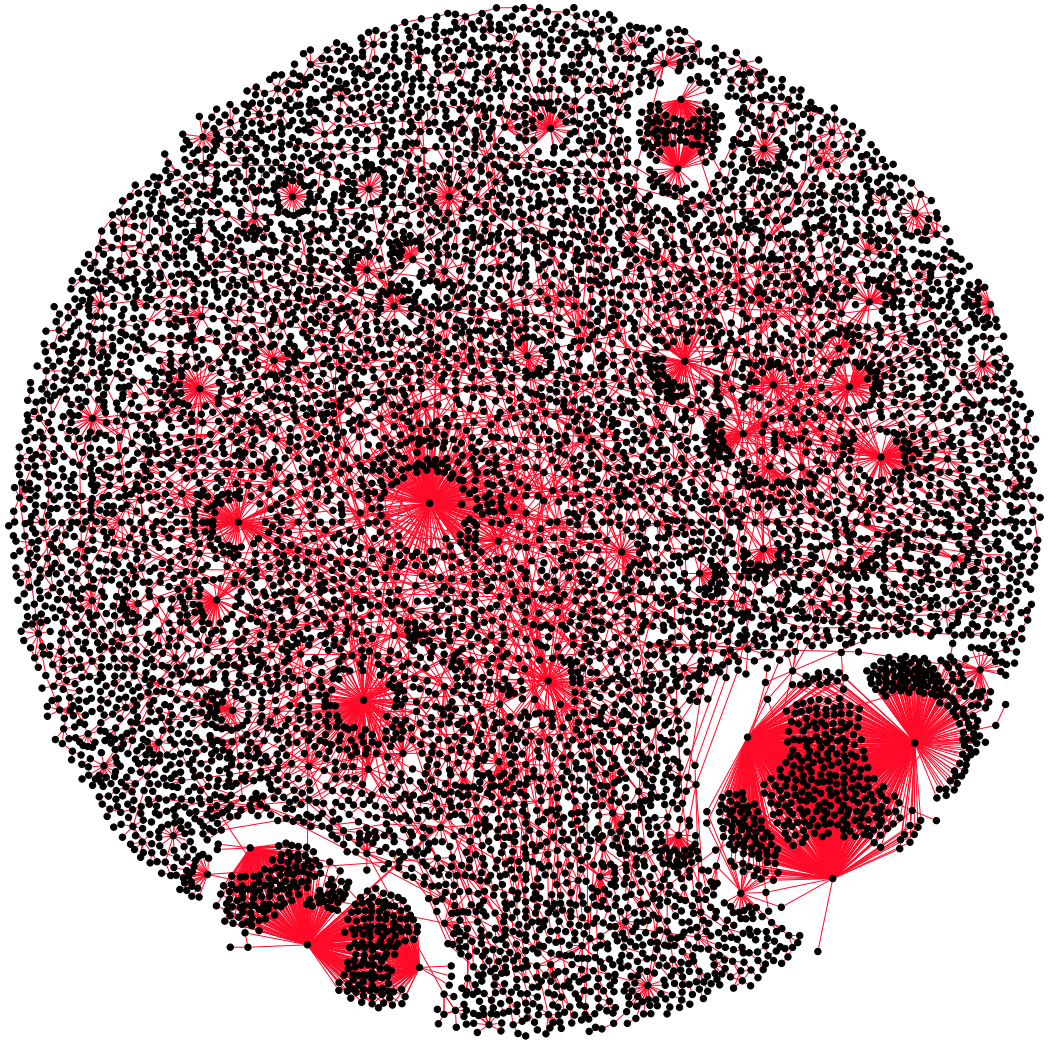


Board interlock network

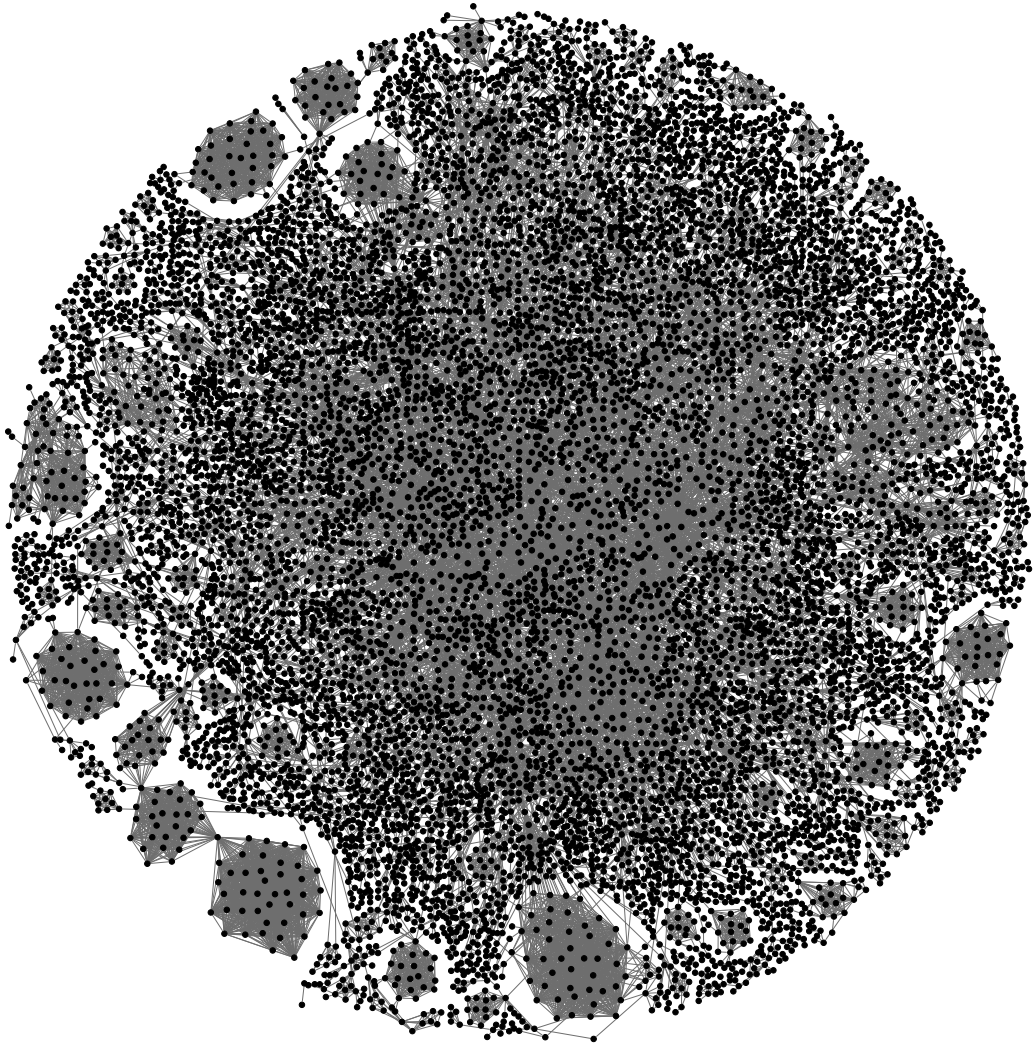


F.W. Takes and E.M. Heemskerk, Centrality in the Global Network of Corporate Control, *Social Network Analysis and Mining* 6(1): 1-18, 2016.

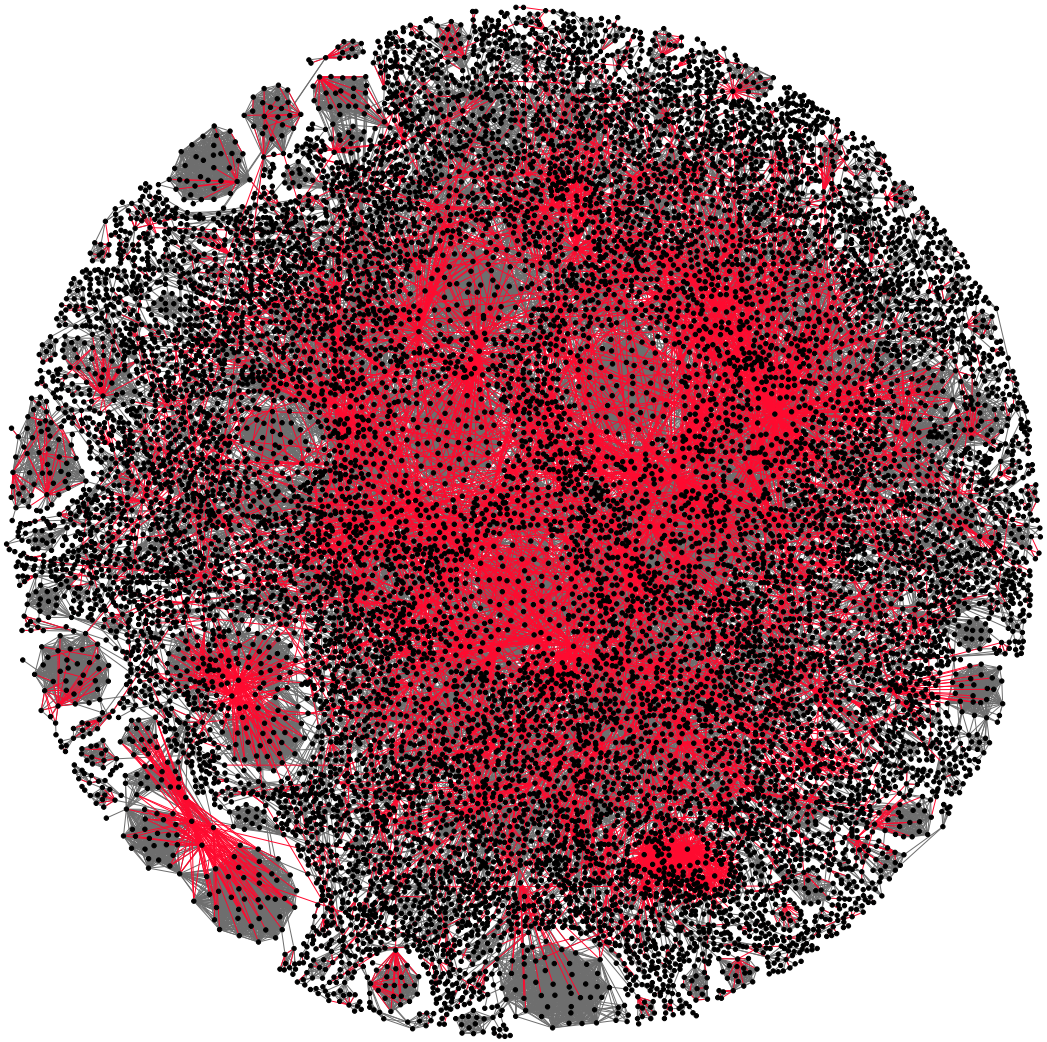
Ownership network



Board interlock network



Multiplex network



Data

Table: Network statistics.

Network	**Nodes**	**Edges**	**Density**	**Clustering**
Ownership	37 724	31 506	$2.25 \cdot 10^{-5}$	0.033
Board interlock	61 209	175 108	$4.67 \cdot 10^{-5}$	0.384
Multiplex	75 224	195 073	$1.72 \cdot 10^{-5}$	0.277

Data

Table: Network statistics.

Network	Nodes	Edges	Density	Clustering
Ownership	37 724	31 506	$2.25 \cdot 10^{-5}$	0.033
Board interlock	61 209	175 108	$4.67 \cdot 10^{-5}$	0.384
Multiplex	75 224	195 073	$1.72 \cdot 10^{-5}$	0.277

- Corporate network of Germany
- $23,709$ nodes are involved in both edge types
- $11,541$ edges are multiplex edges, which is
  - 37% of the ownership links
  - 6.7% of the board interlock links
  - 5.9% of the multiplex links

Data

Table: Division of firms over economic sectors.

Sector	**Ownership**		**Board interlock**		**Multiplex**	
Bank	474	1.25%	865	1.41%	972	1.29%
Financial	4 648	12.32%	6 250	10.21%	8 338	11.08%
Foundation/Research	55	0.14%	51	0.08%	88	0.12%
Industrial	32 350	85.75%	53 767	87.84%	65 484	87.05%
Insurance	19	0.05%	26	0.04%	34	0.05%
Mutual/Pension Fund	112	0.30%	175	0.29%	213	0.28%
Private Equity	29	0.08%	30	0.05%	37	0.05%
Public Authority	22	0.06%	31	0.05%	41	0.05%
Venture Capital	15	0.04%	14	0.02%	17	0.02%

Motifs

Motifs



Motifs

- Meso-level of networks
- Induced subgraphs of size > 2

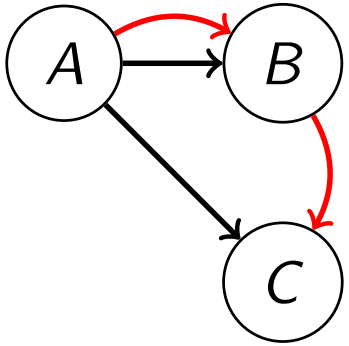
Motifs

- Meso-level of networks
- Induced subgraphs of size > 2
- Count the frequency of each subgraph (NP-hard)
- Compare frequency of subgraphs to a random network
- **Motif**: subgraph occurring at numbers significantly greater than in a random network

Motifs

- Meso-level of networks
- Induced subgraphs of size > 2
- Count the frequency of each subgraph (NP-hard)
- Compare frequency of subgraphs to a random network
- **Motif**: subgraph occurring at numbers significantly greater than in a random network
- Challenge: motifs in *multiplex* networks

Counting *multiplex* subgraphs



$$\begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 0 & 3 & 1 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix} \end{array} \rightarrow 031002000$$

Figure: Pattern labeling in directed multiplex networks.

Null model

- Stub matching model
- Modeling each *layer* will not realize 5.9% overlap between layers

Null model

- Stub matching model
- Modeling each *layer* will not realize 5.9% overlap between layers
- Instead, we model each *combination of layers* separately
  For L layers, we have $2^L - 1$ models

Null model

- Stub matching model
- Modeling each *layer* will not realize 5.9% overlap between layers
- Instead, we model each *combination of layers* separately
  For L layers, we have $2^L - 1$ models
- Combine modeled layers into multiplex network
- Sample a large number of models, and count its subgraphs

Significance testing

1. Subgraph **ratio**:

$$\frac{\text{frequency in data}}{\text{frequency in random model samples}}$$

2. Subgraph **concentration**:

$$\frac{\text{frequency in data}}{\text{frequency of all patterns of that size}}$$

- Cut-off value for ratio and concentration determines which subgraphs (patterns) are **motifs**

Experimental setup

- Ratio cut-off: 5
- Concentration cut-off: 0.01
- 1 000 random graph model samples
- Motif size $k = 3$, $k = 4$ and $k = 5$
- Assess composition of motifs based on industry sector

Results

Table: Patterns and motifs per network.

	Pattern size				**Motif size**			
	3	**4**	**5**	**All**	**3**	**4**	**5**	**All**
Ownership	11	63	391	465	3	4	6	13
Board interlock	2	6	21	29	0	2	10	12
Multiplex	58	1 132	21 858	23 048	14	48	73	135

Results

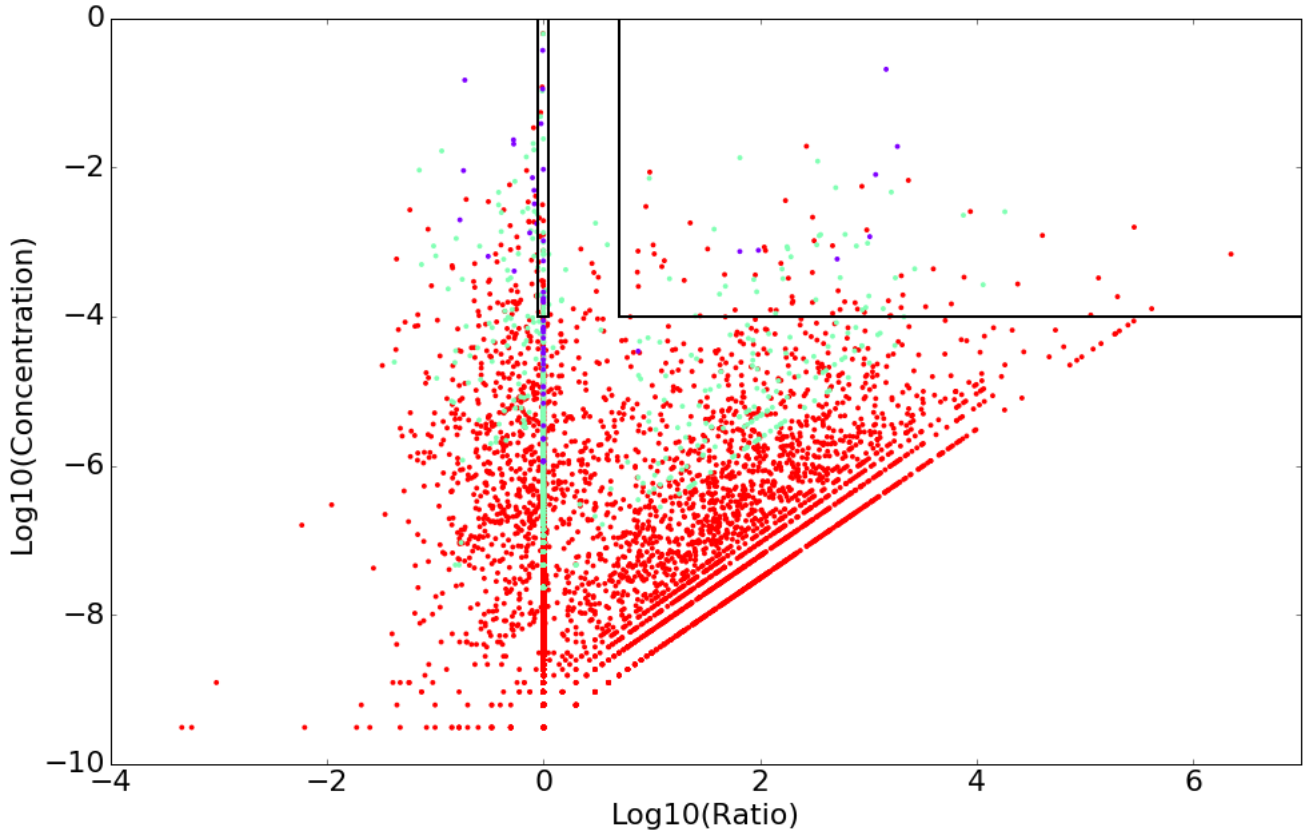
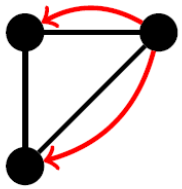
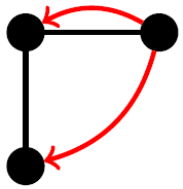


Figure: Ratio (horizontal axis) vs concentration (vertical axis) for all patterns. Top right box indicates cut-off values. Patterns of size 3 in blue, size 4 in green and size 5 in red.

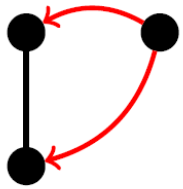
Highlighted motifs



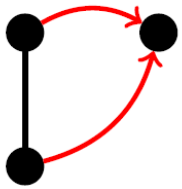
(a) Motif of size 3. $r = \infty$, $c = 0.945$.



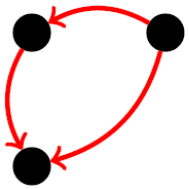
(b) Motif of size 3. $r = 1032$, $c = 0.118$.



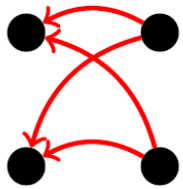
(c) Motif of size 3. $r = 1170$, $c = 0.805$.



(d) Motif of size 3. $r = 516$, $c = 0.059$.



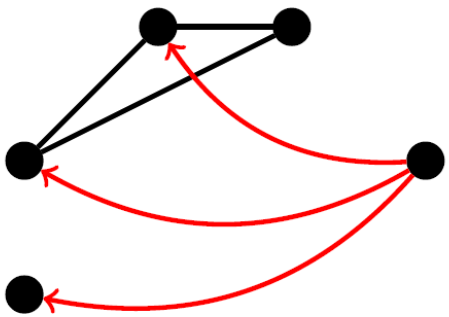
(e) Motif of size 3. $r = 278$, $c = 0.418$.



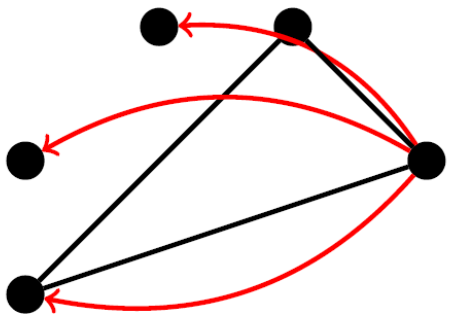
(f) Motif of size 4. $r = 2024$, $c = 0.351$.

Figure: Highlighted motifs of size 3 and 4.

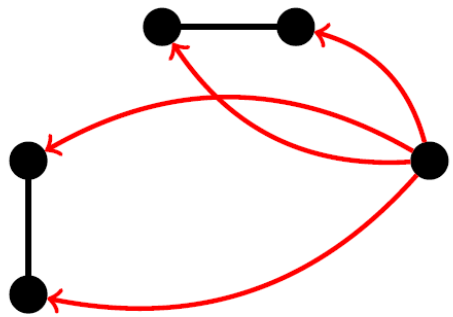
Highlighted motifs



(a) Motif of size 5. $r = 285\,266$, $c = 0.158$.



(b) Motif of size 5. $r = 2346$, $c = 0.673$.



(c) Motif of size 5. $r = 113\,400$, $c = 0.010$.

Figure: Highlighted motifs of size 5.

Results

- Investment and board interlocks often go together
- Well-connected boards attract investments
- Co-investment by interlocked firms happens frequently
- Shared investment strategies

Results

- Investment and board interlocks often go together
- Well-connected boards attract investments
- Co-investment by interlocked firms happens frequently
- Shared investment strategies
- Size 2 motif of reciprocated links: crossholdings

Results

- Investment and board interlocks often go together
- Well-connected boards attract investments
- Co-investment by interlocked firms happens frequently
- Shared investment strategies
- Size 2 motif of reciprocated links: crossholdings
- Over-representation of financial firms in certain motifs
- Substantial involvement of finance sector in complex motifs

Sectoral results

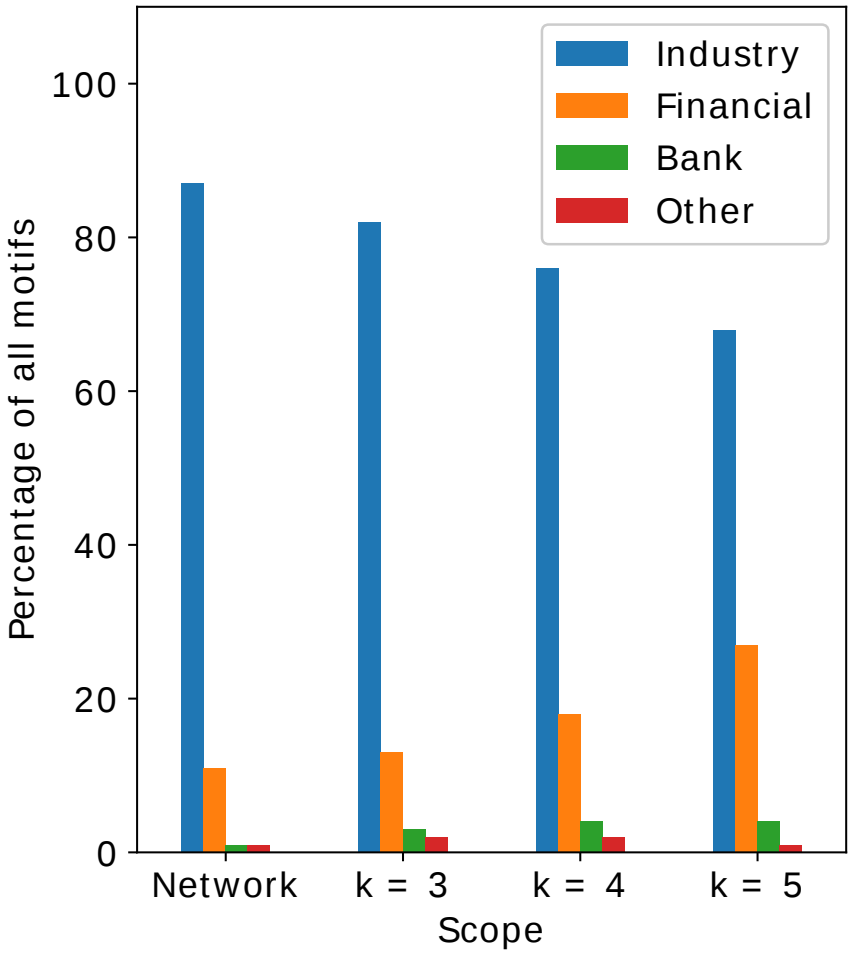


Figure: Division of firms over economic sectors for the full network and motifs of different sizes.

Conclusion

- Multiplex motif recognition is nontrivial
  - Layer encoding scheme for enumerating and counting
  - Null models to account for interlayer assortativity
- Motifs provide insight in the organization of corporations
- There is added value in the multiplex aspect of corporate networks
- The financial sector is overrepresented in complex network patterns
- Future work: cross-country comparison

Upcoming lab session and "Homework"

- Progress substantially with Assignment 2; ask your questions during the lab session!

Credits

- Several images are gratefully reused from
  Michele Coscia, The Atlas for the Aspiring Network Scientist, v3 of
  `https://arxiv.org/abs/2101.00863`