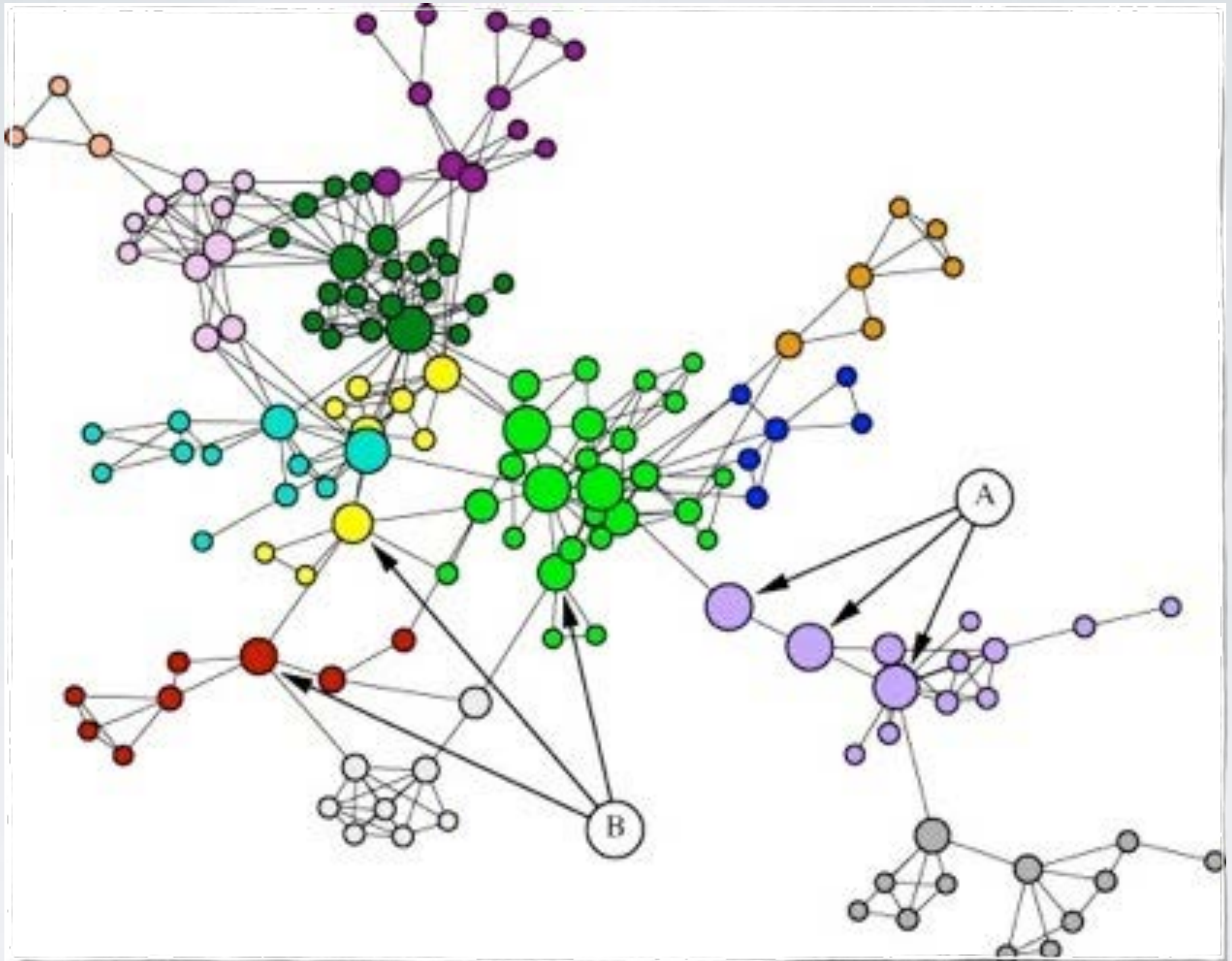
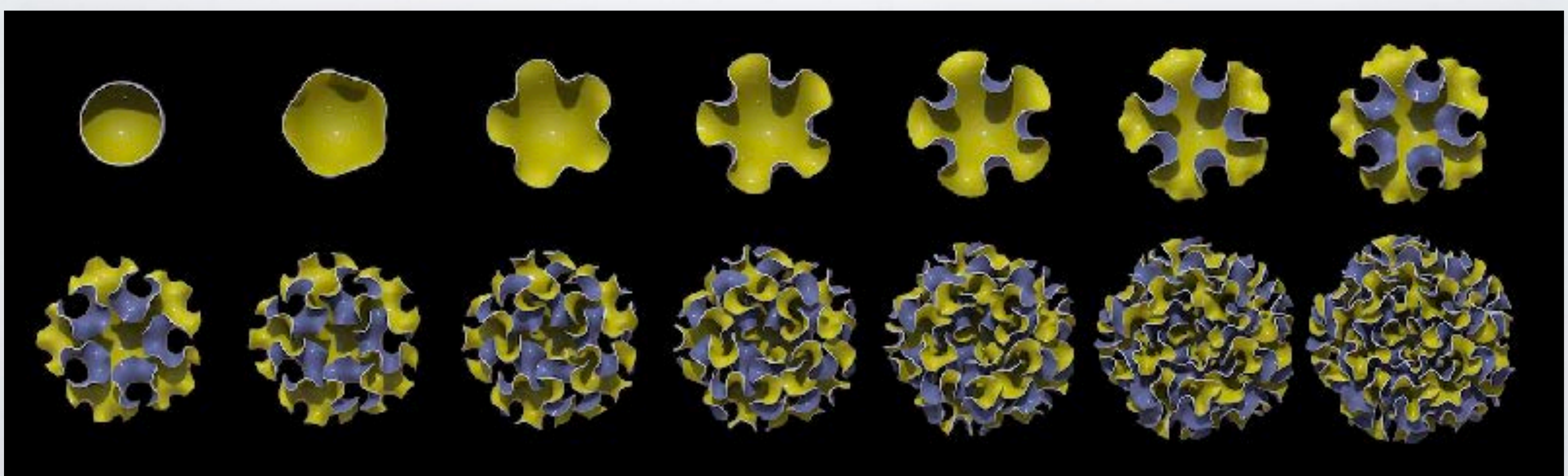
INTRODUCTION TO NETWORK MORPHOGENESIS

Camille Roth
CNRS

Centre Marc Bloch Berlin e.V.
(BMBF / CNRS / Humboldt Universität / MAE)


A
B



A BRIEF TAXONOMY...

reconstructing / using	processes	structure
processes	Preferential attachment Link prediction, classifiers Scoring methods	PA-based models Rewiring models Cost optimization Agent-based models
structure	ERGMs, p1, p* Markov graphs SOAMs	Prescribed structure, edge swaps Subgraph-based Kronecker graphs

A BRIEF TAXONOMY...

using \ reconstructing	processes	structure
processes	Preferential attachment Link prediction, classifiers Scoring methods	PA-based models Rewiring models Cost optimization Agent-based models
structure	ERGMs, p1, p* Markov graphs SOAMs	Prescribed structure, edge swaps Subgraph-based Kronecker graphs

PREFERENTIAL ATTACHMENT

de Solla Price, 1976

"A general theory of bibliometric and other cumulative advantage processes"

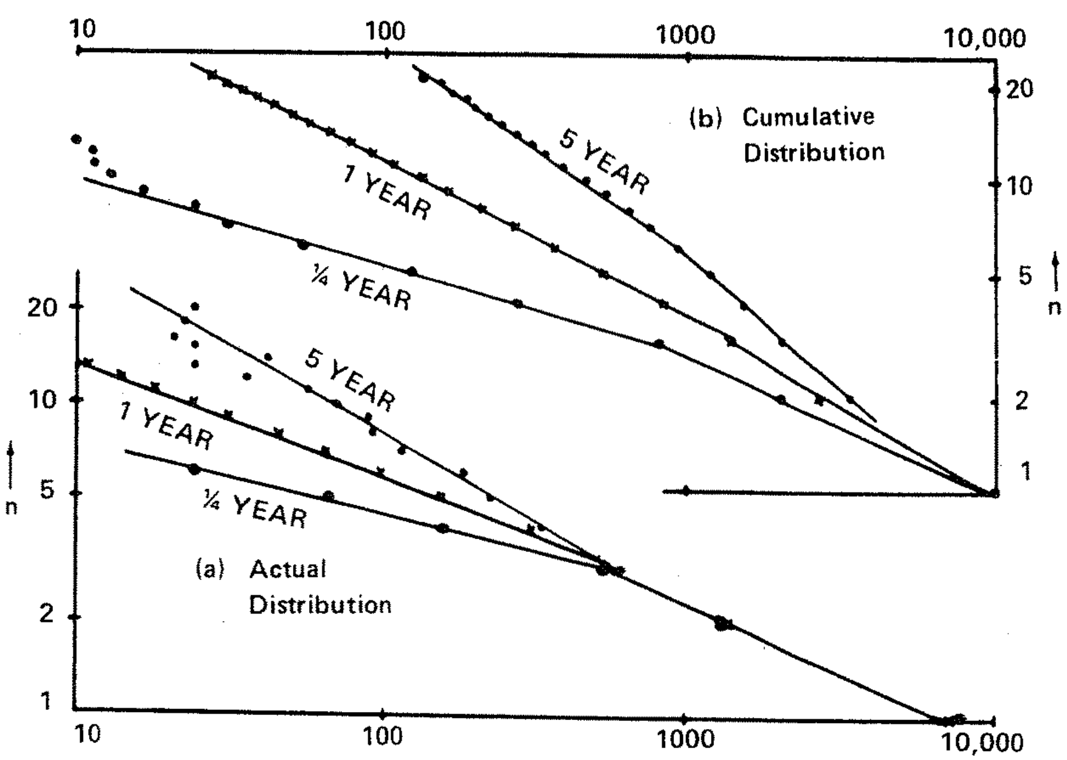
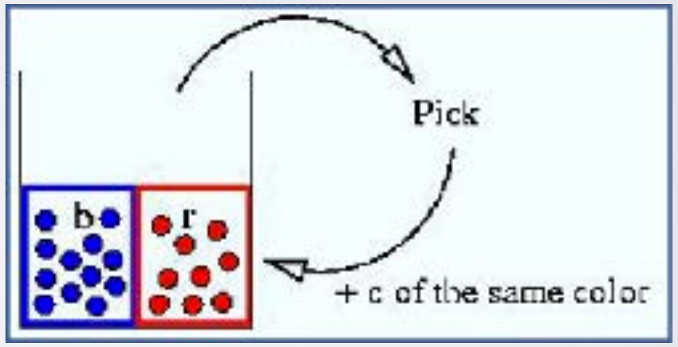


Fig. 1. Number of papers with (a) exactly and (b) at least n citations in ¼, 1, and 5-year indexes.

"cumulative advantage theory"

Polya Urn model



over *n* urns with uniform initial conditions

(converges to a power law with exponent 2)

...then relaxing uniformity

PREFERENTIAL ATTACHMENT

- “Classical” preferential attachment:

Barabasi, Jeong, Neda, Ravasz, Schubert, Vicsek, 2002

  - assuming that links do not attach uniformly with respect to degree k, with a bias function **∏(k)** depicting the degree increment of degree-k nodes

$$\boxed{\kappa(k) = \int_1^k \Pi(k')\,\mathrm{d}k'}$$

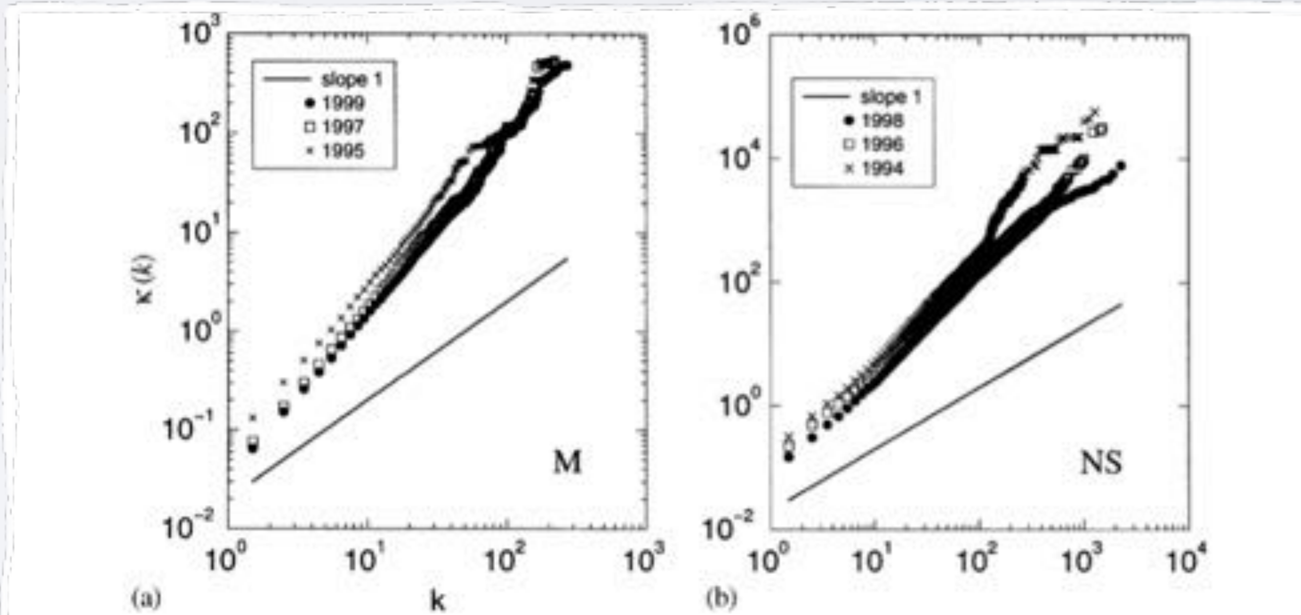


Fig. 7. Cumulated preferential attachment ($\kappa(k)$) of incoming new nodes for the M and NS database. Results computed by considering the new nodes coming in the specified year, and the network formed by nodes already present up to this year. In the absence of preferential attachment $\kappa(k) \sim k$, shown as continuous line on the figures.

PREFERENTIAL ATTACHMENT

"Classical" preferential attachment:

Barabasi, Jeong, Neda, Ravasz, Schubert, Vicsek, 2002

- assuming that links do not attach uniformly with respect to degree k, with a bias function $\prod(\mathbf{k})$ depicting the degree increment of degree-k nodes
- ...or that links attach with respect to degrees of link extremities $\prod(\mathbf{k_1,k_2})$

$$\kappa(k_1 k_2) = \int_1^{k_1 k_2} \Pi(k_1' k_2')\,\mathrm{d}(k_1' k_2')$$

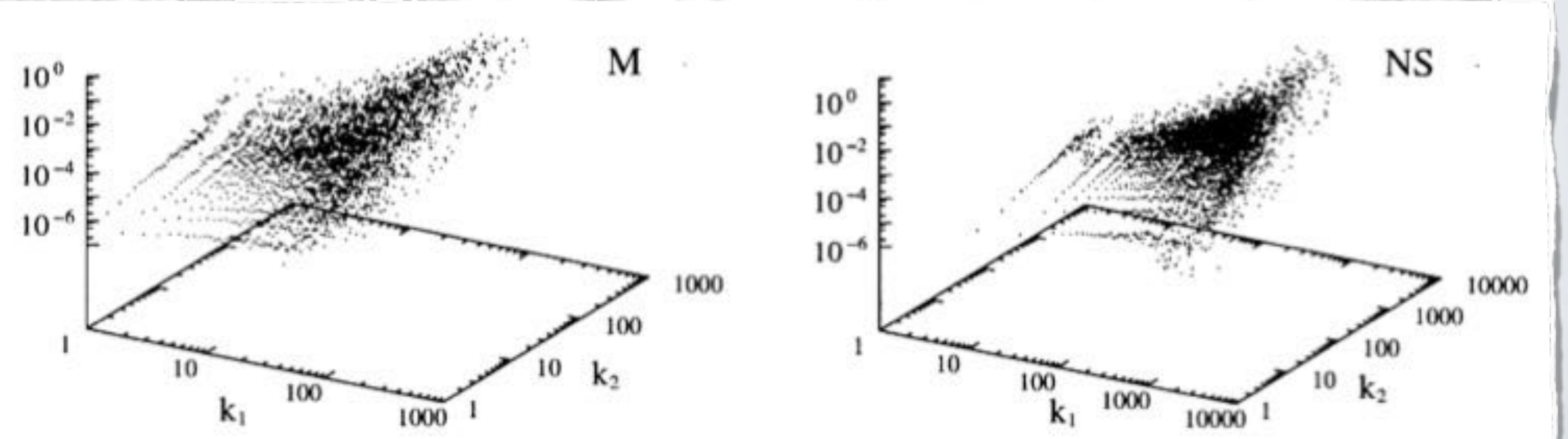


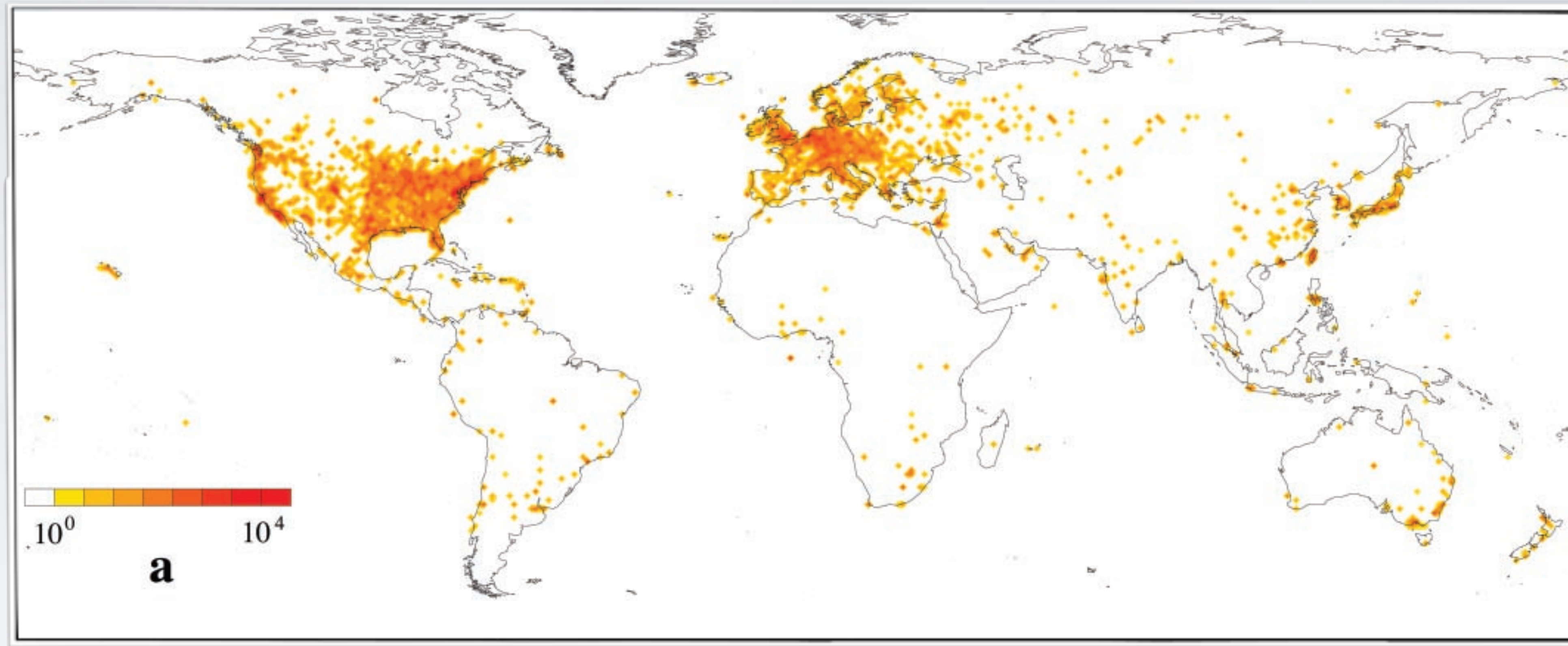
Fig. 8. Internal preferential attachment for the M and NS database, 3D plots: $\pi(k_1, k_2)$ as a function of k_1 and k_2. Results computed on the cumulative data in the last considered year.

PREFERENTIAL ATTACHMENT

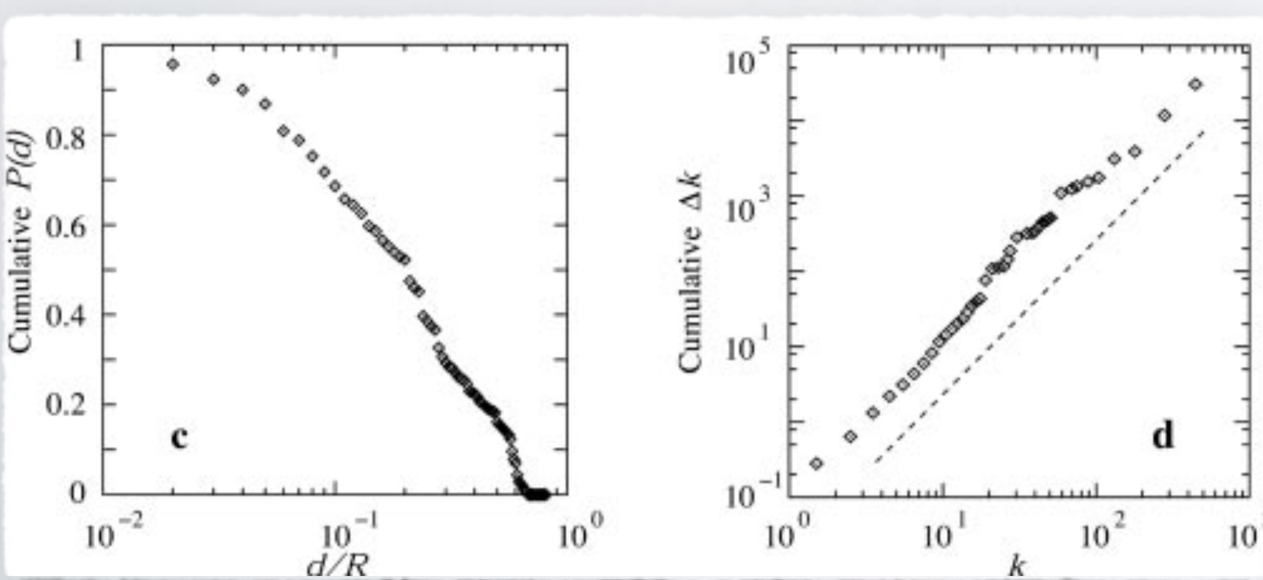
Yook, Jeong, Barabasi, 2002

"Modeling the internet's large-scale topology"

spatial distance



Worldwide router density map (2002)



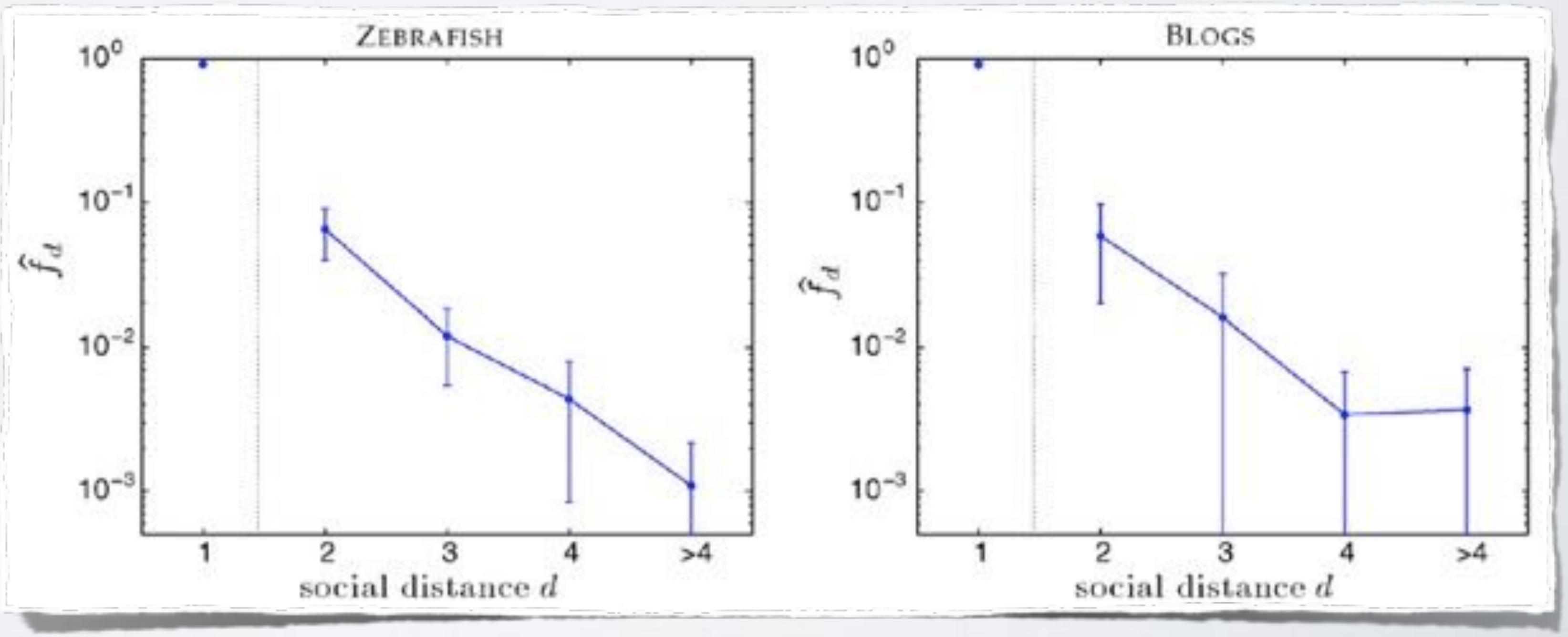
$$\Pi(k_j, d_{ij}) \sim k_j^{\alpha} / d_{ij}^{\sigma},$$

PREFERENTIAL ATTACHMENT

Preferential attachment may work for any type of variable

- propension to create/receive links with respect to dyad properties, i.e. comparing the values of **P(L|d)** for various values of **d**
- or computing the relative propension of appearance of a link between a dyad **d** relatively to the baseline

$$\boxed{\frac{P(L|d)}{P(L)} = \frac{\nu(d)}{\nu} \cdot \frac{N}{N(d)}}$$



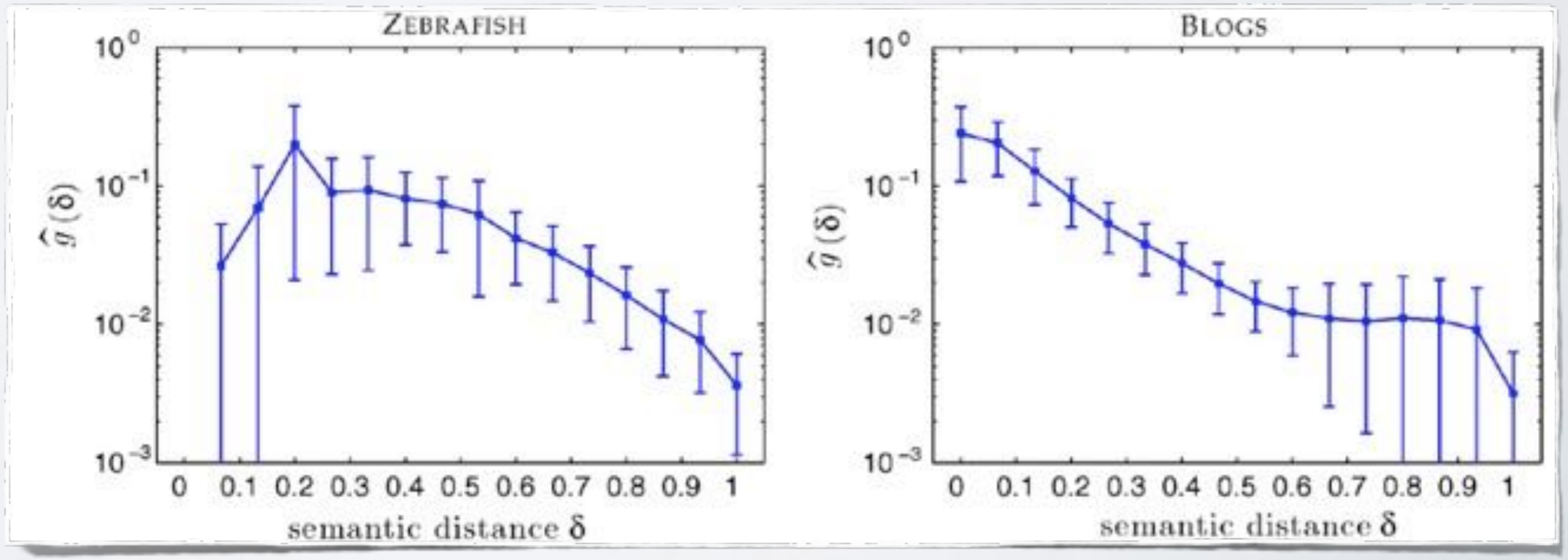
(Roth, 2005; Cointet, Roth, 2011; Roth, Cointet, 2010)

PREFERENTIAL ATTACHMENT

Preferential attachment may work for any type of variable

- propension to create/receive links with respect to dyad properties, i.e. comparing the values of **P(L|d)** for various values of **d**
- or computing the relative propension of appearance of a link between a dyad **d** relatively to the baseline

$$\boxed{\frac{P(L|d)}{P(L)} = \frac{\nu(d)}{\nu} \cdot \frac{N}{N(d)}}$$



(Roth, 2005; Cointet, Roth, 2011; Roth, Cointet, 2010)

PREFERENTIAL ATTACHMENT

Preferential attachment may work for any type of variable

- propension to create/receive links with respect to dyad properties, i.e. comparing the values of **P(L|d)** for various values of **d**
- or computing the relative propension of appearance of a link between a dyad **d** relatively to the baseline

$$\boxed{\frac{P(L|d)}{P(L)} = \frac{\nu(d)}{\nu} \cdot \frac{N}{N(d)}}$$

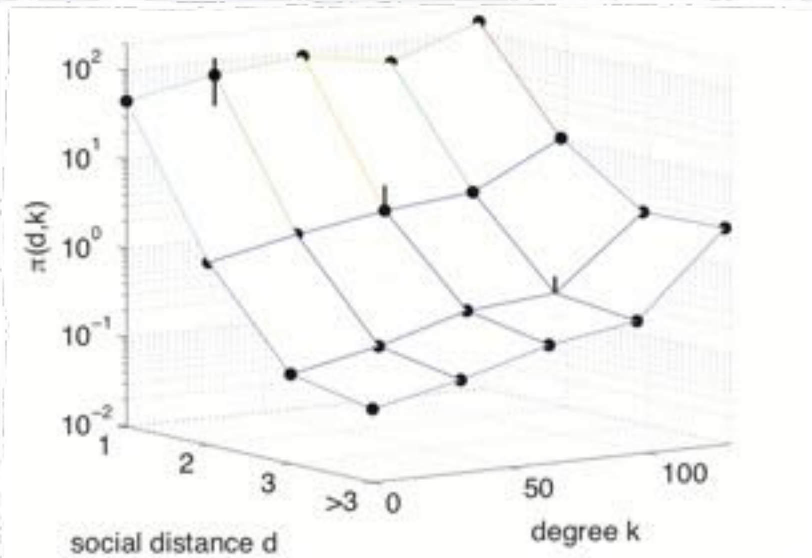


Figure 2: Interaction propensity $\Pi(d, k)$ with respect to social distance and degree (averaged over the 8 last periods).

(Roth, 2005; Cointet, Roth, 2011; Roth, Cointet, 2010)

PREFERENTIAL ATTACHMENT

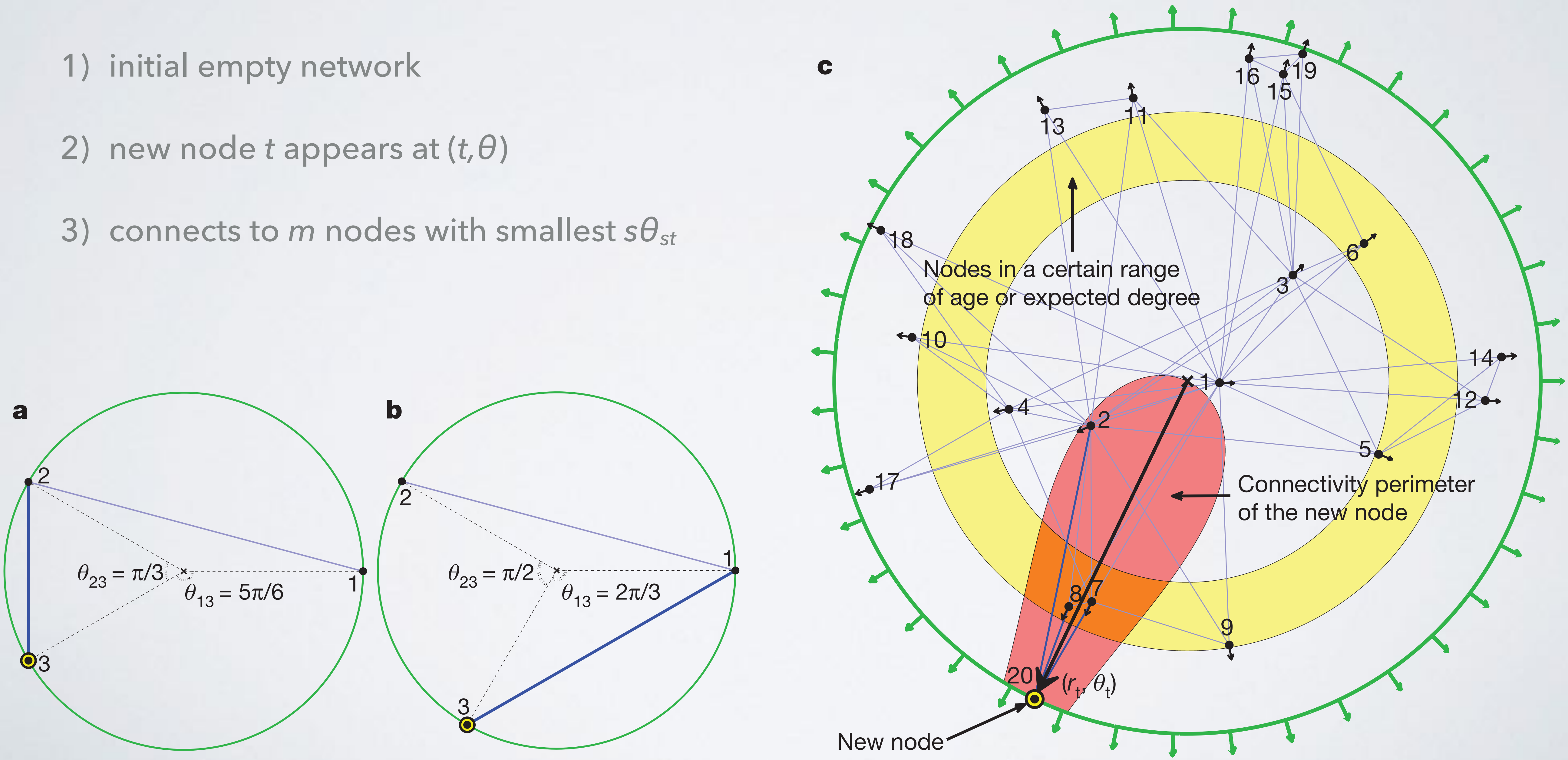
Papadopoulos, Kitsak, Ángeles-Serrano, Boguná, Krioukov, 2012

"Popularity versus similarity in growing networks"

1) initial empty network

2) new node t appears at (t,θ)

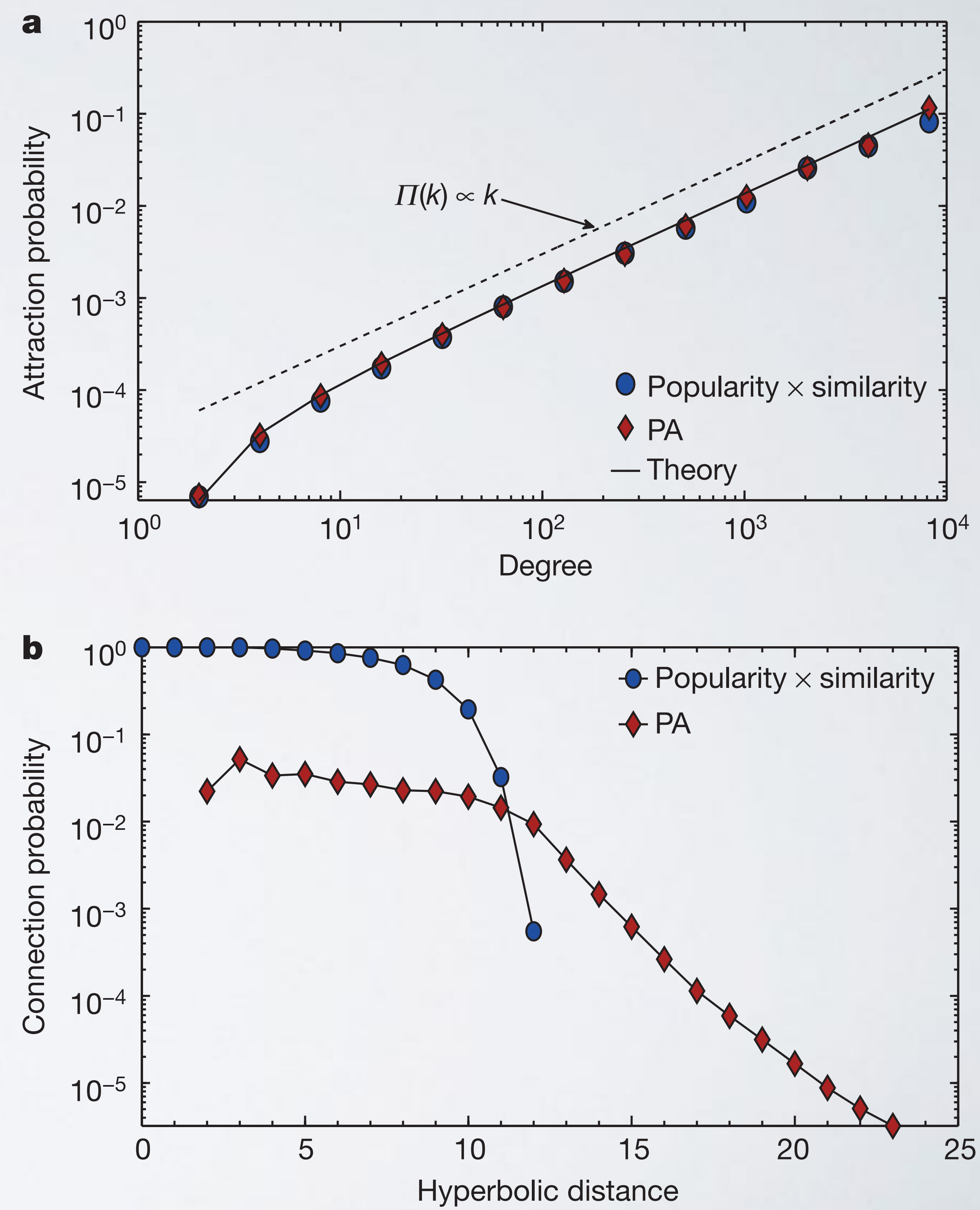
3) connects to m nodes with smallest $s\theta_{st}$



PREFERENTIAL ATTACHMENT

Papadopoulos, Kitsak, Ángeles-Serrano, Boguná, Krioukov, 2012

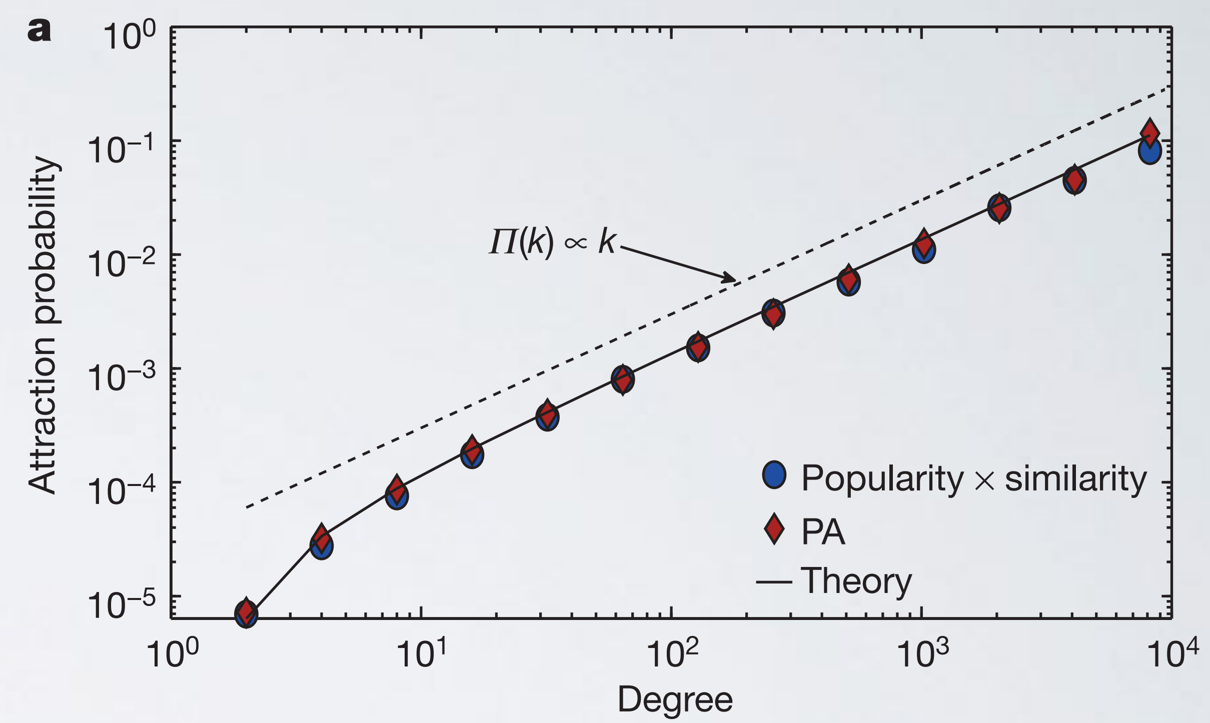
"Popularity versus similarity in growing networks"

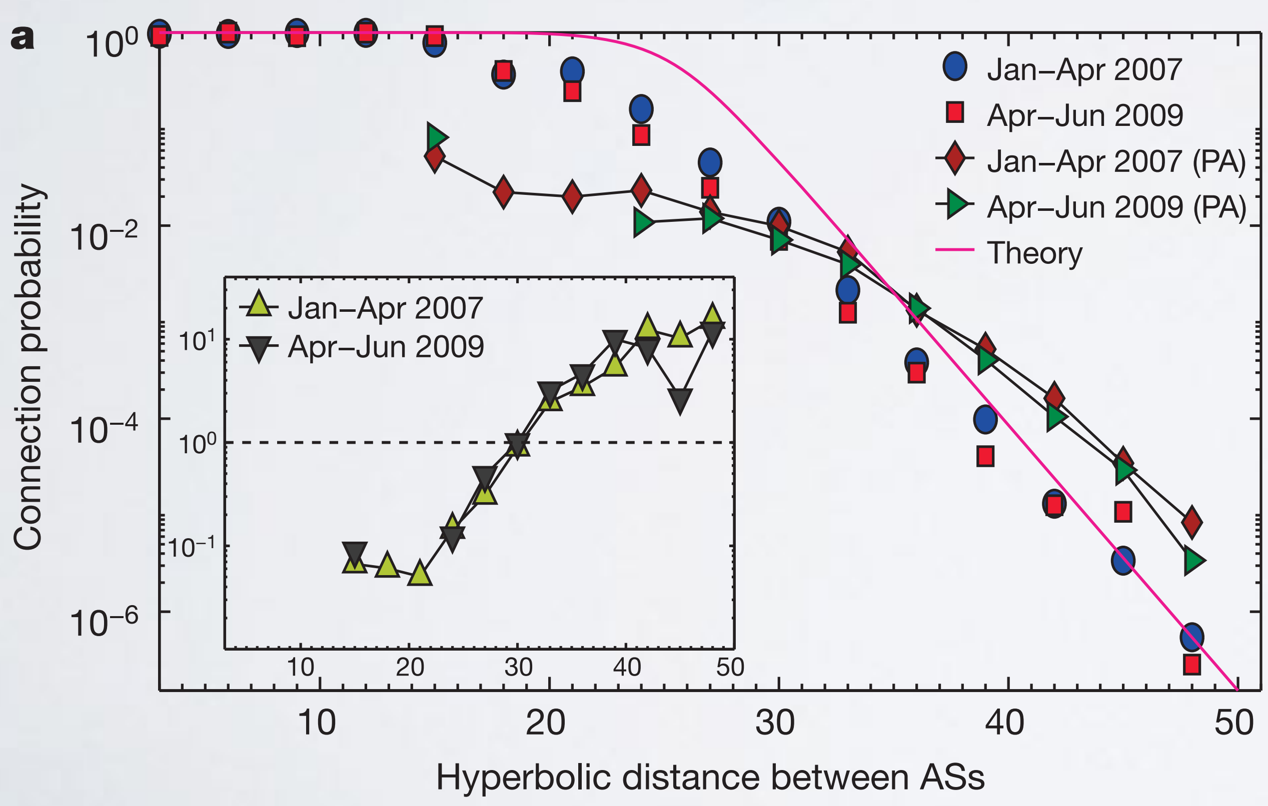


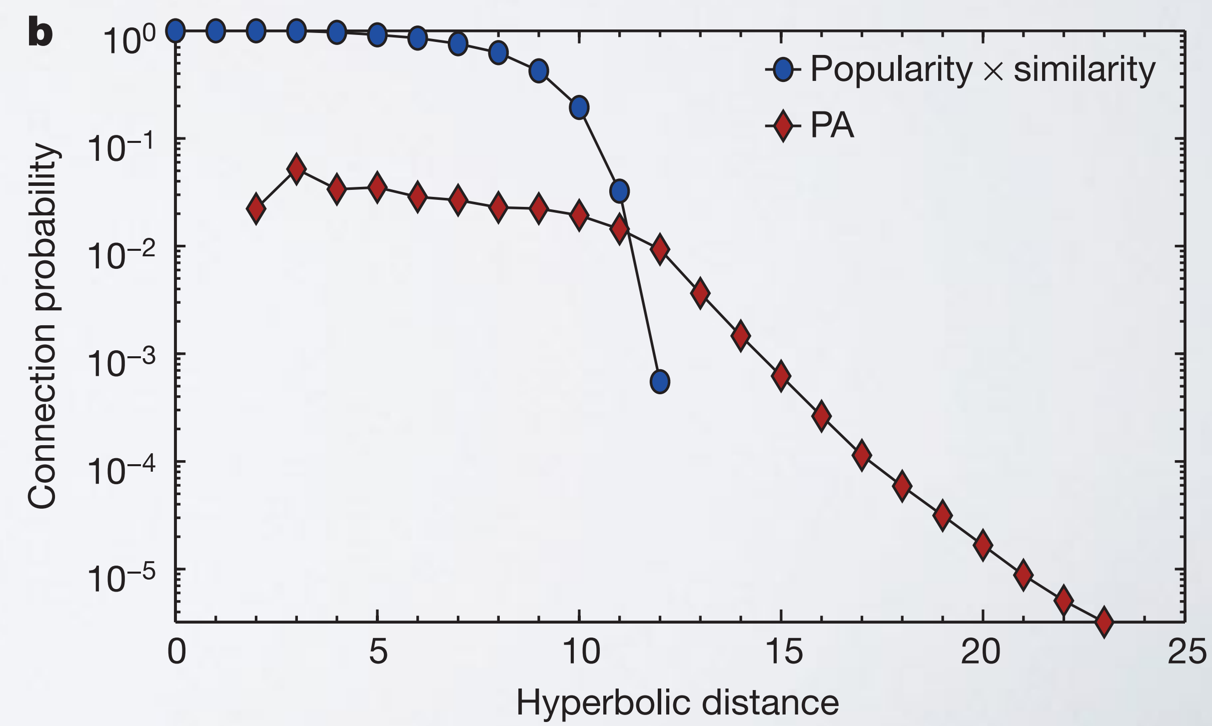
PREFERENTIAL ATTACHMENT

Papadopoulos, Kitsak, Ángeles-Serrano, Boguná, Krioukov, 2012

"Popularity versus similarity in growing networks"



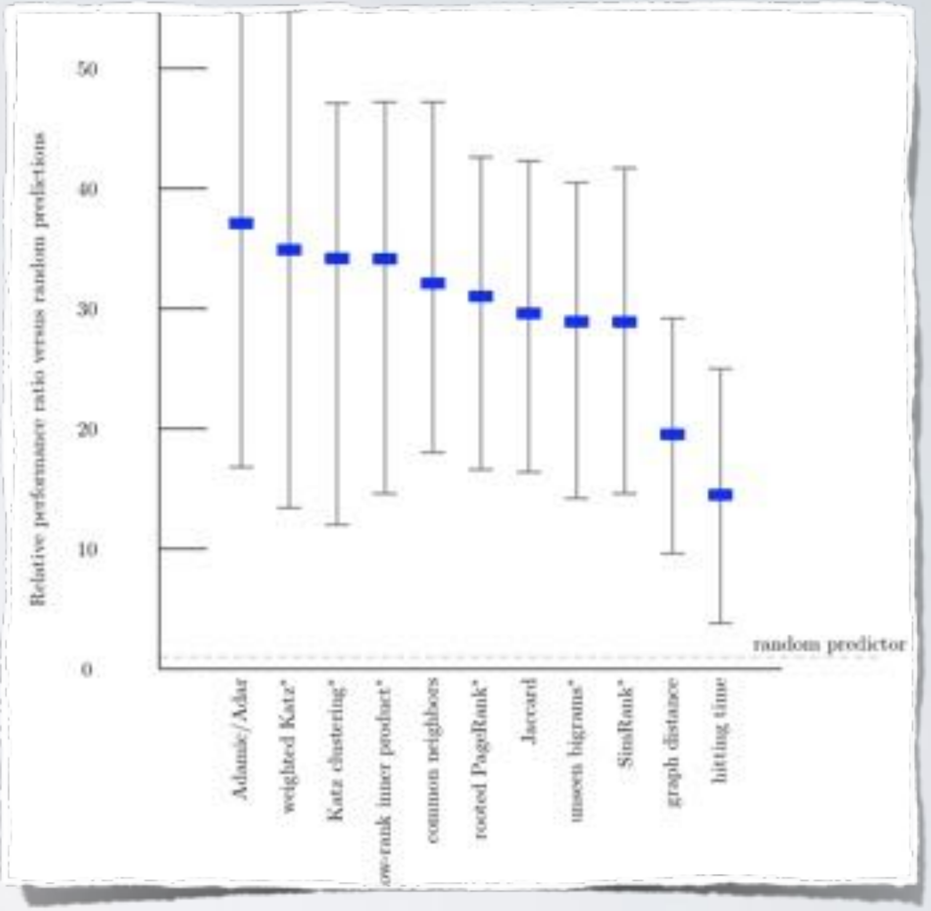




SCORING METHODS - I

Derived from AI / machine learning

- targeted at link prediction
  rather than behavior estimation



Scoring methods

(Liben-Nowell, Kleinberg, 2003)

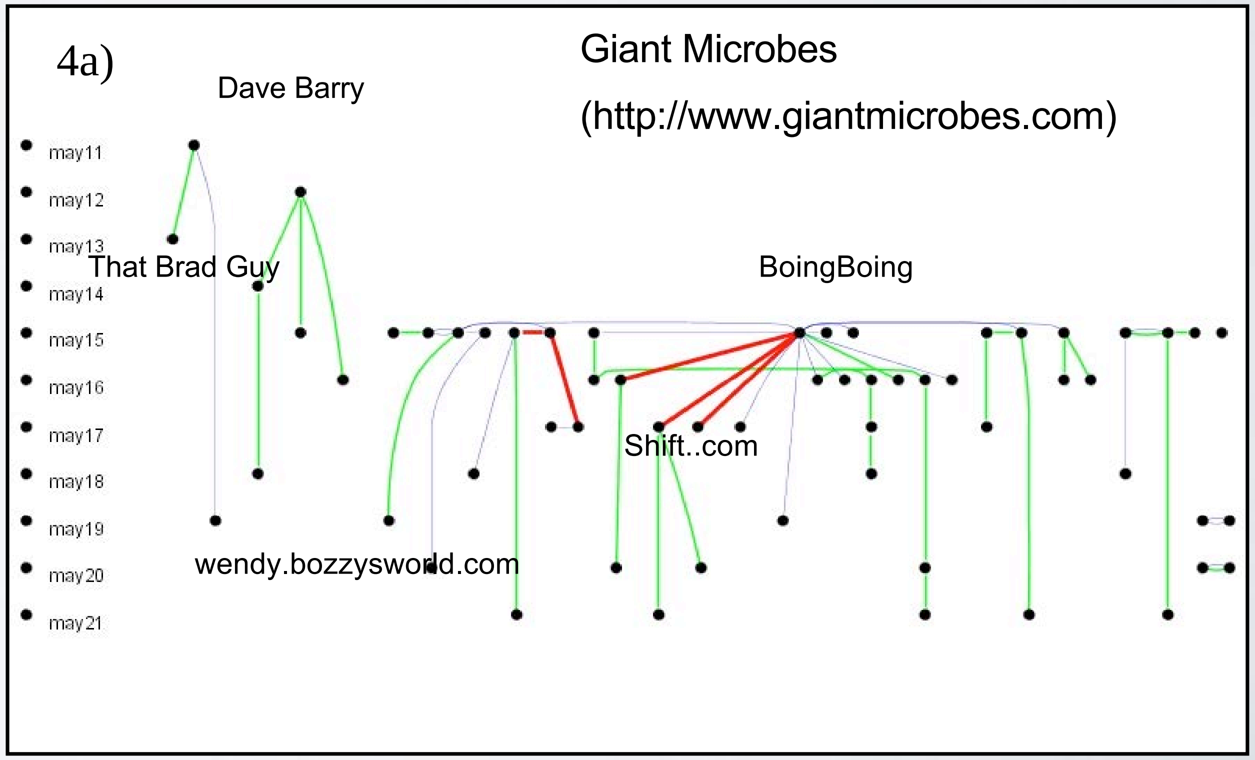
- based on a predictor function score(x,y)
  using measures such as number of common neighbors,
  Jaccard coefficients, Katz' distance $\boxed{\sum_{\ell=1}^{\infty} \beta^{\ell} \cdot |\mathsf{paths}_{x,y}^{\langle \ell \rangle}|}$
- computes the list of scores of pairs (x,y) of a network observed over $[t_0,t_1]$
- predicts new links for $t>t_1$ according to decreasing values of score, among the non-connected pairs during $[t_0,t_1]$

SCORING METHODS - II

Classifier-based methods

(Adar, Adamic, 2004)

- using a variety of features altogether:
  - number of common linked blogs
  - common URLs
  - textual cosine similarity
  - degree similarity



- and SVM-classifiers or classical logistic regressions in order to predict the existence (or not) of:
  - one-way / two-way links,
  - explicit infection links

SCORING METHODS - III

Clauset, Moore, Newman, 2008

"Hierarchical structure and the prediction of missing links in networks"

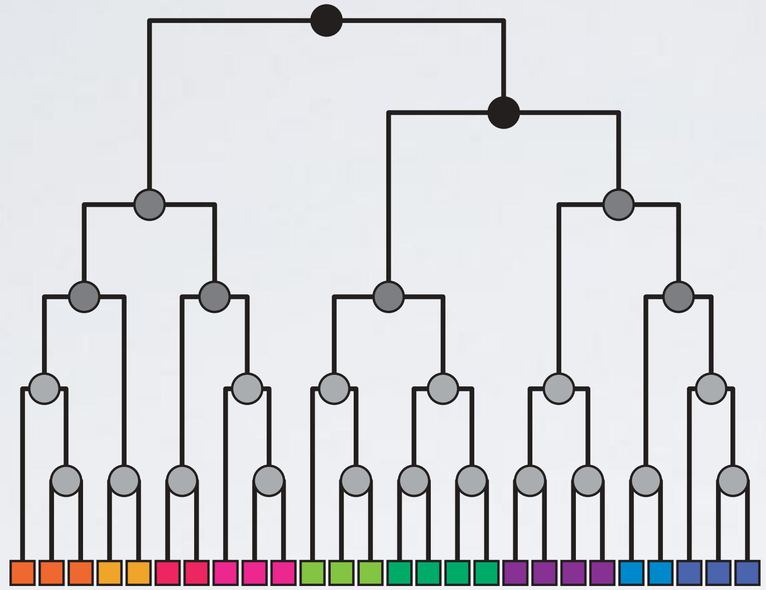
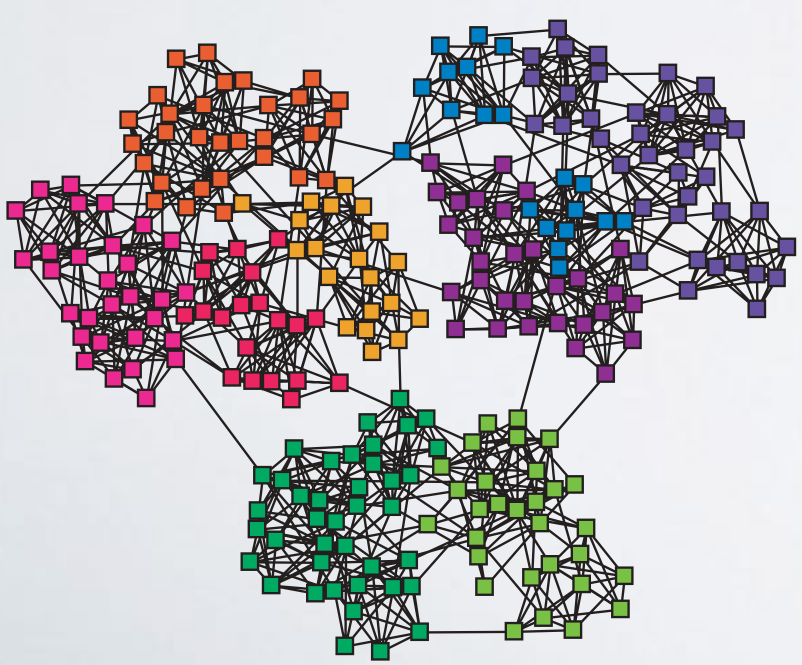


Figure 1 | A hierarchical network with structure on many scales, and the corresponding hierarchical random graph. Each internal node r of the dendrogram is associated with a probability p_r that a pair of vertices in the left and right subtrees of that node are connected. (The shades of the internal nodes in the figure represent the probabilities.)

Given a dendrogram and a set of probabilities p_r, the hierarchical random graph model generates artificial networks with a specified hierarchical structure



SCORING METHODS - III

Clauset, Moore, Newman, 2008

"Hierarchical structure and the prediction of missing links in networks"

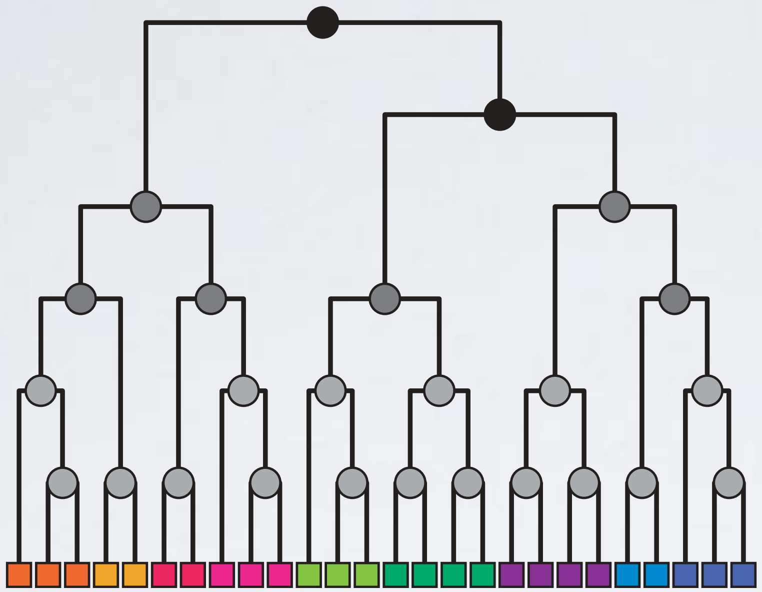


Figure 1 | A hierarchical network with structure on many scales, and the corresponding hierarchical random graph. Each internal node r of the

Given a dendrogram and a set of probabilities p_r, the hierarchical random graph model generates artificial networks with a specified hierarchical structure

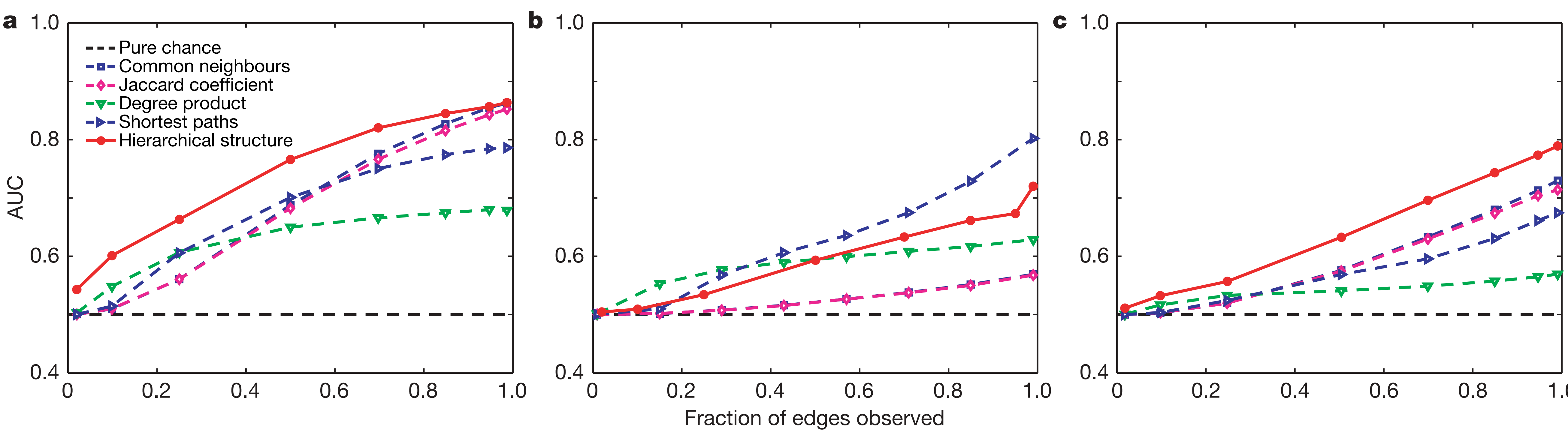


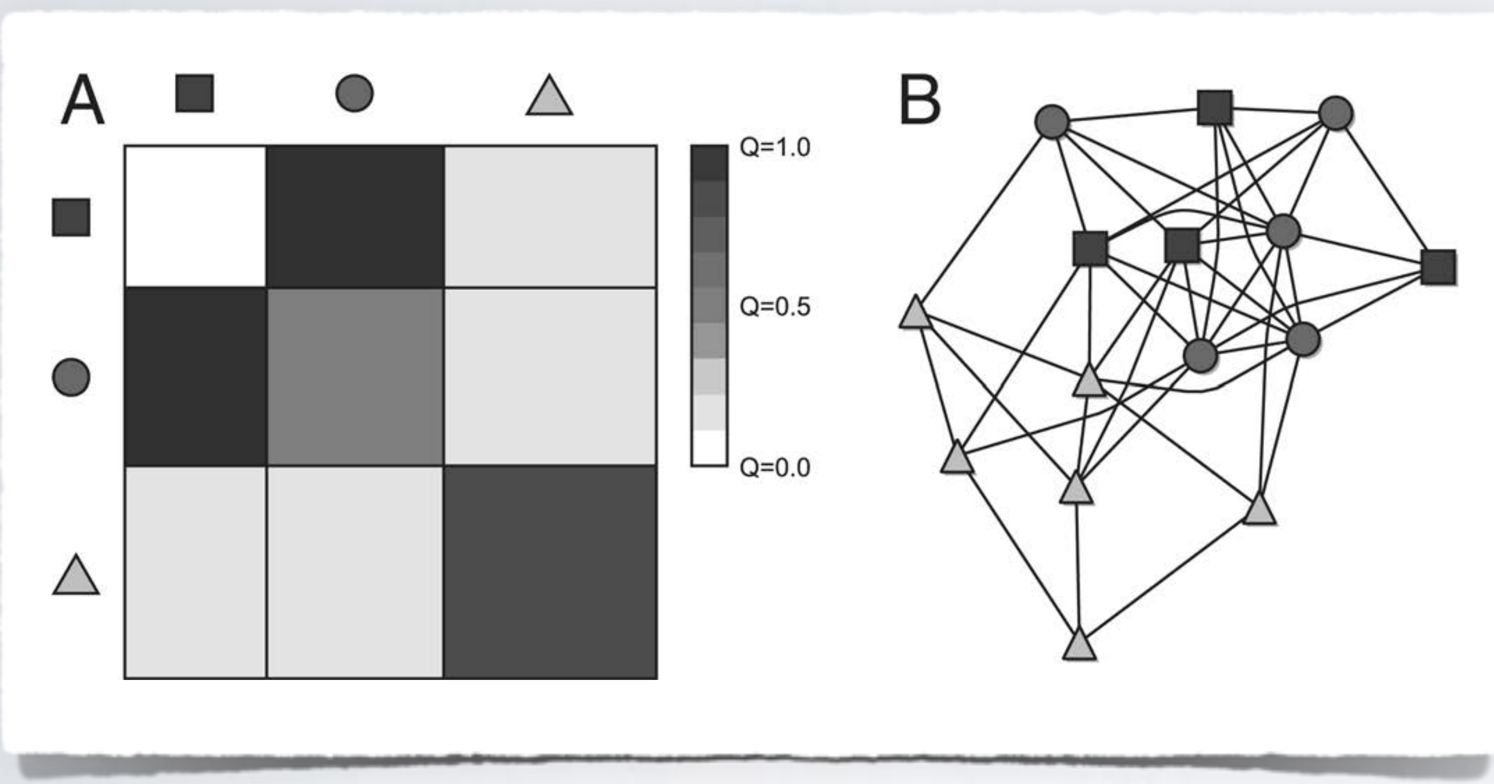
Figure 3 | Comparison of link prediction methods. Average AUC statistic—that is, the probability of ranking a true positive over a true negative—as a function of the fraction of connections known to the algorithm, for the link prediction method presented here and a variety of previously published methods. **a**, Terrorist association network; **b**, *T. pallidum* metabolic network; **c**, grassland species network.

SCORING METHODS - III

Guimerà, Sales-Pardo, 2009

"Missing and spurious interactions and the reconstruction of complex networks"

Fig. 1. Stochastic block models. A stochastic block model is fully specified by a partition of nodes into groups and a matrix $\boldsymbol{Q}$ in which each element $Q_{\alpha\beta}$ represents the probability that a node in group α connects to a node in group β. (*A*), A simple matrix of probabilities $\boldsymbol{Q}$. Nodes are divided in three groups (which contain 4, 5, and 6 nodes, respectively) and are represented as squares, circles, and triangles depending on their group. The value of each element $Q_{\alpha\beta}$ is indicated by the shade of gray; for example, squares do not connect to other squares, and connect to triangles with small probability, but squares connect to circles with high probability. (*B*) A realization of the model in *A*. In this realization, the number of links between the square and the triangle group is $l_{\square\triangle} = 4$, whereas the maximum possible number of links between these groups is $r_{\square\triangle} = 24$.

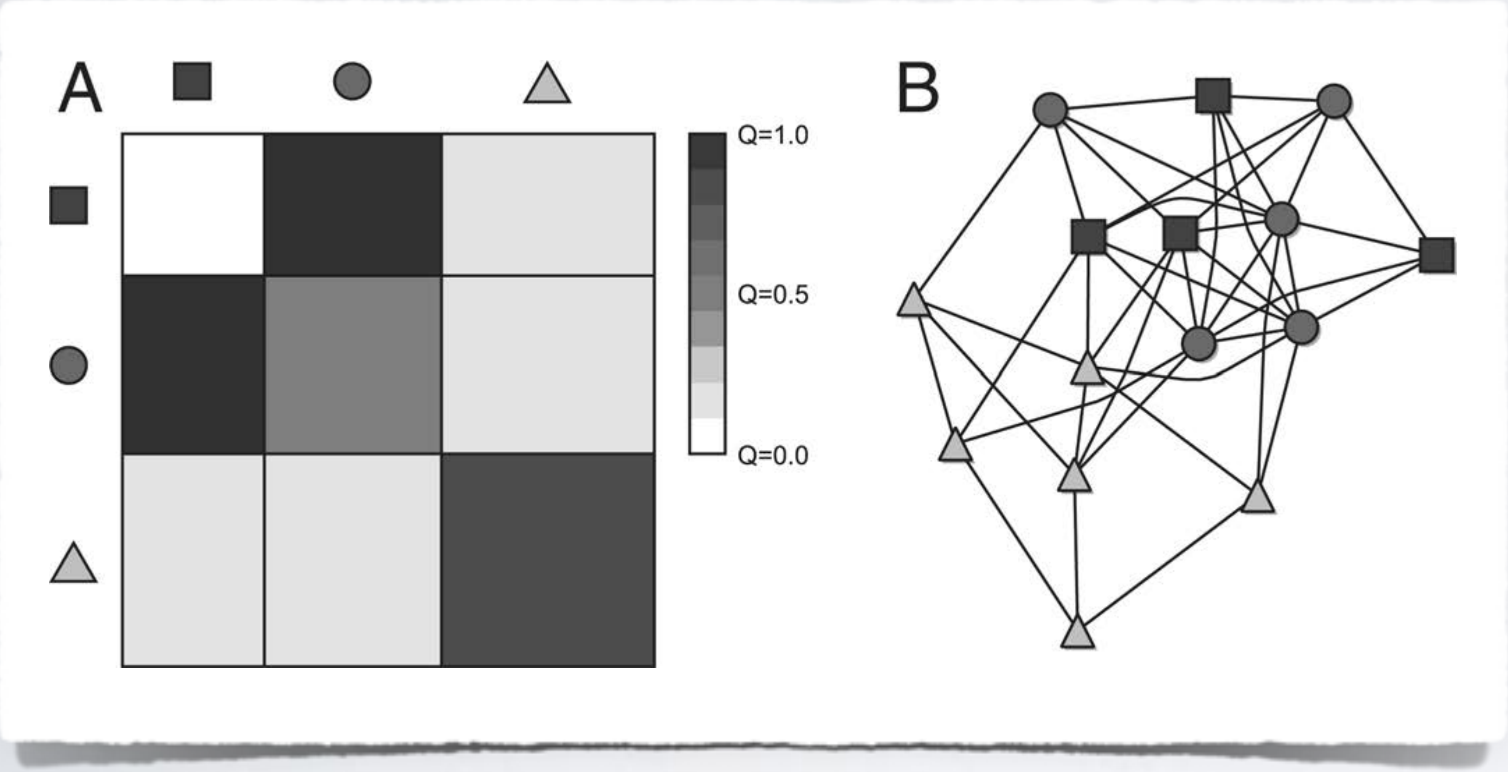


SCORING METHODS - III

Guimerà, Sales-Pardo, 2009

"Missing and spurious interactions and the reconstruction of complex networks"

Fig. 1. Stochastic block models. A stochastic block model is fully specified by a partition of nodes into groups and a matrix $\boldsymbol{Q}$ in which each element $Q_{\alpha\beta}$ represents the probability that a node in group α connects to a node in group β. (*A*), A simple matrix of probabilities $\boldsymbol{Q}$. Nodes are divided in three groups (which contain 4, 5, and 6 nodes, respectively) and are represented as squares, circles, and triangles depending on their group. The value of each element $Q_{\alpha\beta}$ is indicated by the shade of gray; for example, squares do not connect to other squares, and connect to triangles with small probability, but squares connect to circles with high probability. (*B*) A realization of the model in *A*. In this realization, the number of links between the square and the triangle group is $l_{\square\triangle} = 4$, whereas the maximum possible number of links between these groups is $r_{\square\triangle} = 24$.



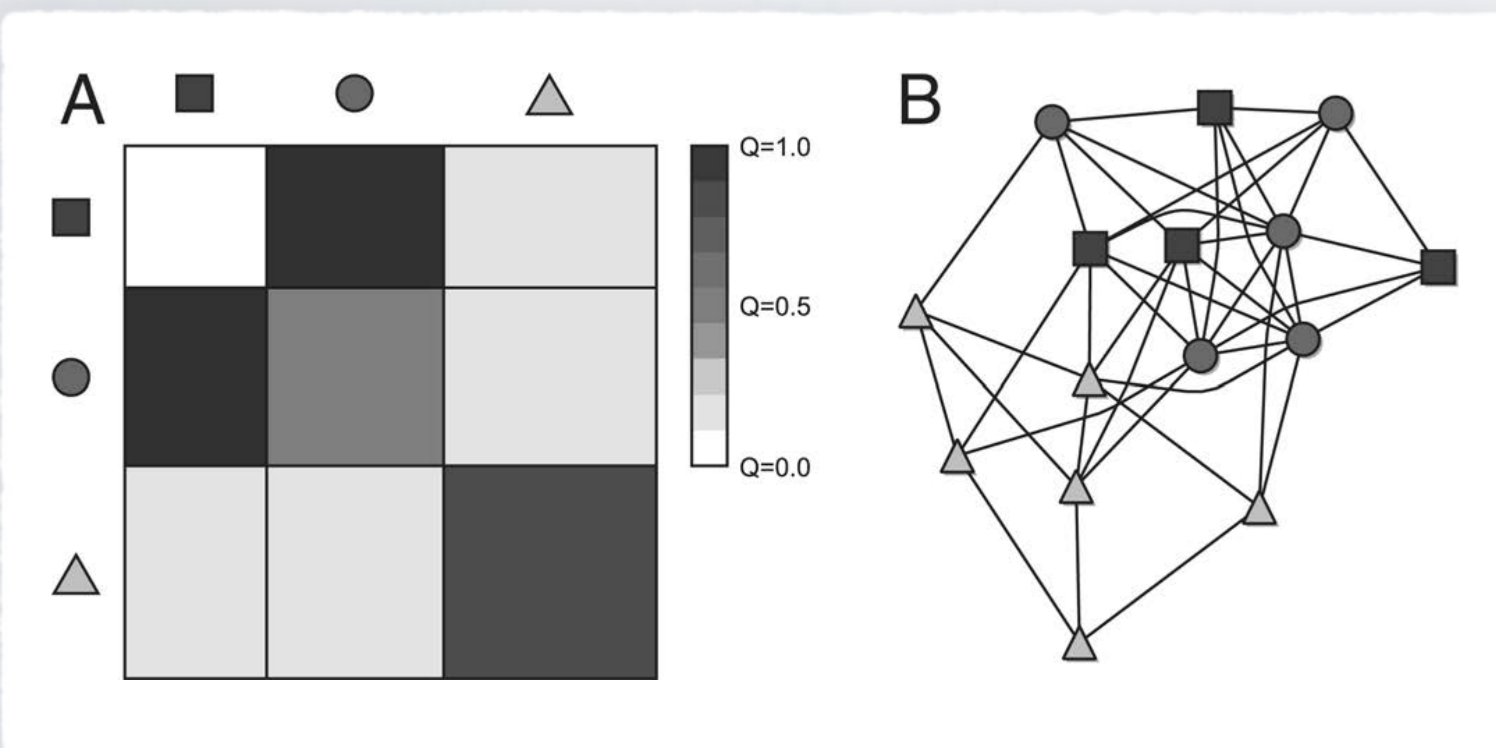
reliability of link *ij* : $$R^L_{ij} = \frac{1}{Z} \sum_{P \in \mathcal{P}} \left(\frac{l^O_{\sigma_i \sigma_j} + 1}{r_{\sigma_i \sigma_j} + 2} \right) \exp[-\mathcal{H}(P)]$$

SCORING METHO

Guimerà, Sales-Pardo, 2009

"Missing and spurious interactions and the reconstruction of complex networks"

Fig. 1. Stochastic block models. A stochastic block model is fully specified by a partition of nodes into groups and a matrix $\boldsymbol{Q}$ in which each element $Q_{\alpha\beta}$ represents the probability that a node in group α connects to a node in group β. (*A*), A simple matrix of probabilities $\boldsymbol{Q}$. Nodes are divided in three groups (which contain 4, 5, and 6 nodes, respectively) and are represented as squares, circles, and triangles depending on their group. The value of each element $Q_{\alpha\beta}$ is indicated by the shade of gray; for example, squares do not connect to other squares, and connect to triangles with small probability, but squares connect to circles with high probability. (*B*) A realization of the model in *A*. In this realization, the number of links between the square and the triangle group is $l_{\square\triangle} = 4$, whereas the maximum possible number of links between these groups is $r_{\square\triangle} = 24$.



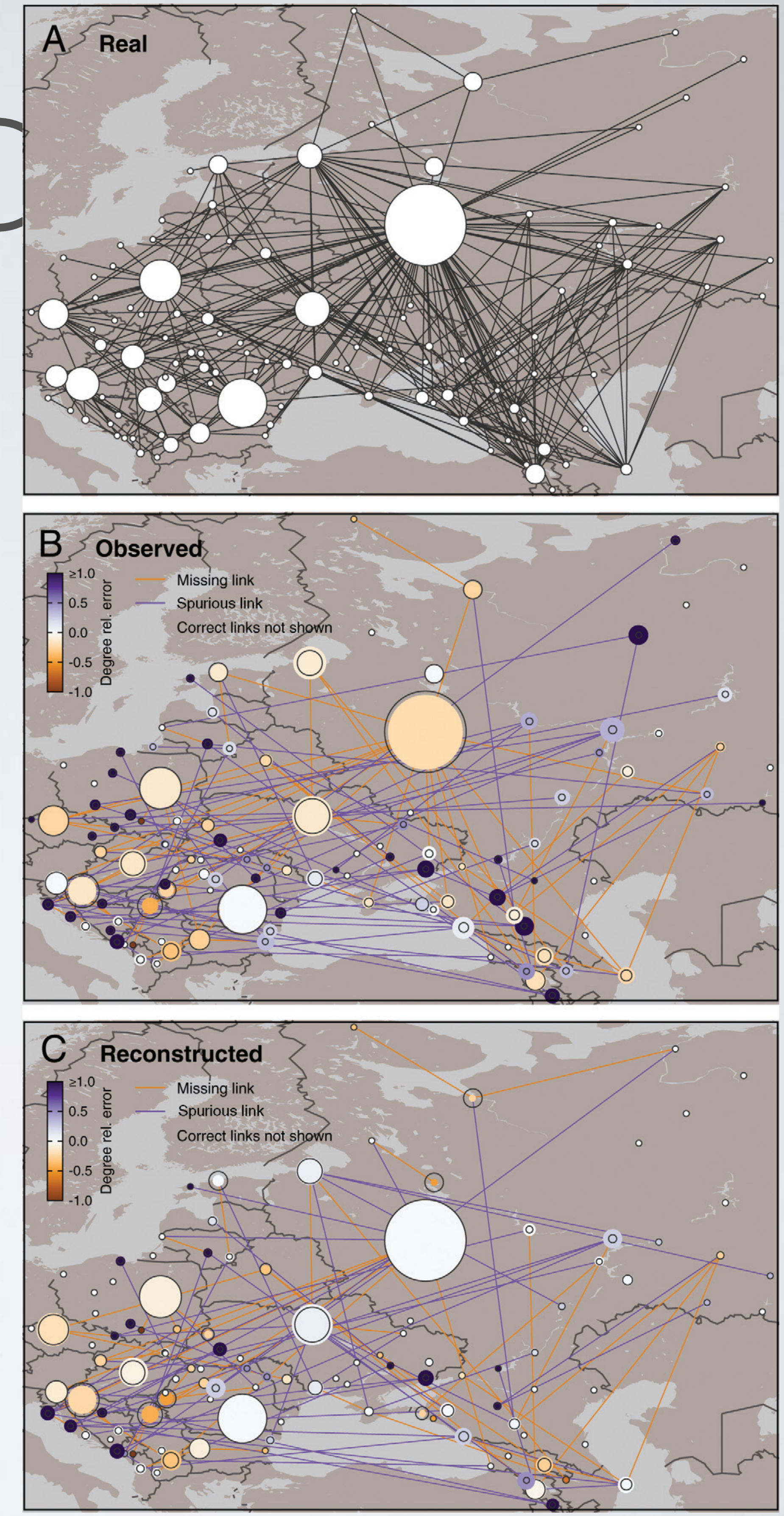


Fig. 3. Reconstruction of the air transportation network of Eastern Europe. (*A*) The true air transportation network. The area of each node is proportional to its betweenness centrality, with Moscow being the most central node in the network. (*B*) The observed air transportation network, which we build by randomly removing 20% of the real links and replacing them by random links. (*C*) The reconstructed air transportation network that we obtain, from the observed network, applying the heuristic reconstruction method described

reliability of link *ij* : $$R^L_{ij} = \frac{1}{Z}\sum_{P\in\mathcal{P}}\left(\frac{l^O_{\sigma_i\sigma_j}+1}{r_{\sigma_i\sigma_j}+2}\right)\exp[-\mathcal{H}(P)]$$

A BRIEF TAXONOMY...

reconstructing / using	processes	structure
processes	Preferential attachment Link prediction, classifiers Scoring methods	PA-based models Rewiring models Cost optimization Agent-based models
structure	ERGMs, p1, p* Markov graphs SOAMs	Prescribed structure, edge swaps Subgraph-based Kronecker graphs

ECONOMETRIC METHODS - I

Holland, Leinhardt, 1981

"An exponential family of probability distributions for directed graphs"

p_1 model

$$p_1(G) \sim \exp(\textstyle\sum_i \lambda_i v_i(G)) = \Pi_i \exp(\lambda_i v_i(G))$$

Dyadic		
Choice	ϕ	$L = \sum_{ij} X_{ij} = X_{++}$
Mutuality	ρ	$M = \sum_{i<j} X_{ij} X_{ji}$

Effect	Explanatory variable	Model parameter	Estimated value	Approximate standard error
Choice	L^{same}	ϕ^{same}	-2.17	1.15
	L^{differ}	ϕ^{differ}	-4.30	1.17
Mutual	M^{gg}	ρ^{gg}	3.15	0.69
	M^{bb}	ρ^{bb}	3.05	0.49
	M^{differ}	ρ^{differ}	3.95	0.72

ECONOMETRIC METHODS - I

Holland, Leinhardt, 1981

"An exponential family of probability distributions for directed graphs"

p_1 model

$$p_1(G) \sim \exp(\textstyle\sum_i \lambda_i v_i(G)) = \Pi_i \exp(\lambda_i v_i(G))$$

Dyadic		
Choice	ϕ	$L = \sum_{ij} X_{ij} = X_{++}$
Mutuality	ρ	$M = \sum_{i<j} X_{ij} X_{ji}$

- **p_1 assumes independence between dyads:**
  - limits the model to simple dyad-centric observables: principally, *degree* and *reciprocity*

Effect	Explanatory variable	Model parameter	Estimated value	Approximate standard error
Choice	L^{same}	ϕ^{same}	-2.17	1.15
	L^{differ}	ϕ^{differ}	-4.30	1.17
Mutual	M^{gg}	ρ^{gg}	3.15	0.69
	M^{bb}	ρ^{bb}	3.05	0.49
	M^{differ}	ρ^{differ}	3.95	0.72

ECONOMETRIC METHODS - I

Holland, Leinhardt, 1981

"An exponential family of probability distributions for directed graphs"

p_1 model

$$p_1(G) \sim \exp(\sum_i \lambda_i v_i(G)) = \Pi_i \exp(\lambda_i v_i(G))$$

Dyadic		
Choice	ϕ	$L = \sum_{ij} X_{ij} = X_{++}$
Mutuality	ρ	$M = \sum_{i<j} X_{ij} X_{ji}$

- **p_1 assumes independence between dyads:**
  - limits the model to simple dyad-centric observables: principally, *degree* and *reciprocity*

Effect	Explanatory variable	Model parameter	Estimated value	Approximate standard error
Choice	L^{same}	ϕ^{same}	-2.17	1.15
	L^{differ}	ϕ^{differ}	-4.30	1.17
Mutual	M^{gg}	ρ^{gg}	3.15	0.69
	M^{bb}	ρ^{bb}	3.05	0.49
	M^{differ}	ρ^{differ}	3.95	0.72

- **can nonetheless be applied to:**
  - a partition of the network into subgroups Fienberg, Meyer, Wasserman, 1985

ECONOMETRIC METHODS - I

Holland, Leinhardt, 1981

"An exponential family of probability distributions for directed graphs"

p_1 model

$$p_1(G) \sim \exp(\textstyle\sum_i \lambda_i v_i(G)) = \Pi_i \exp(\lambda_i v_i(G))$$

Dyadic		
Choice	ϕ	$L = \sum_{ij} X_{ij} = X_{++}$
Mutuality	ρ	$M = \sum_{i<j} X_{ij} X_{ji}$

- **p_1 assumes independence between dyads:**
  - limits the model to simple dyad-centric observables: principally, *degree* and *reciprocity*

Effect	Explanatory variable	Model parameter	Estimated value	Approximate standard error
Choice	L^{same}	ϕ^{same}	-2.17	1.15
	L^{differ}	ϕ^{differ}	-4.30	1.17
Mutual	M^{gg}	ρ^{gg}	3.15	0.69
	M^{bb}	ρ^{bb}	3.05	0.49
	M^{differ}	ρ^{differ}	3.95	0.72

- **can nonetheless be applied to:**
  - a partition of the network into subgroups — Fienberg, Meyer, Wasserman, 1985
  - stochastic block-models — Holland, Laskey, Leinhardt, 1983 — Anderson, Wasserman, Faust, 1992

ECONOMETRIC METHODS - I

Exponential Random Graph Models (ERGMs)

(Wasserman, Pattison, 1997; Anderson, Wasserman, Crouch, 1999)

Frank, Strauss, 1986

"Markov Graphs"

$$P(\mathbf{X} = \mathbf{x}) = \frac{\exp(\theta_1 z_1(\mathbf{x}) + \cdots + \theta_r z_r(\mathbf{x}))}{\kappa(\theta)}$$

$\log[\Pr(\mathbf{X} = \mathbf{x})]$ is proportional to $\theta_1 z_1(\mathbf{x}) + \cdots + \theta_r z_r(\mathbf{x})$

Dyadic		
Choice	ϕ	$L = \sum_{ij} X_{ij} = X_{++}$
Mutuality	ρ	$M = \sum_{i<j} X_{ij} X_{ji}$
Triadic		
Transitivity	τ_T	$T_T = \sum_{i,j,k} X_{ij} X_{jk} X_{ik}$

ECONOMETRIC METHODS - I

Exponential Random Graph Models (ERGMs)

(Wasserman, Pattison, 1997; Anderson, Wasserman, Crouch, 1999)

Frank, Strauss, 1986

"Markov Graphs"

$$P(\mathbf{X}=\mathbf{x}) = \frac{\exp(\theta_1 z_1(\mathbf{x}) + \cdots + \theta_r z_r(\mathbf{x}))}{\kappa(\theta)}$$

$\log[\Pr(\mathbf{X}=\mathbf{x})]$ is proportional to $\theta_1 z_1(\mathbf{x}) + \cdots + \theta_r z_r(\mathbf{x})$

Dyadic		
Choice	ϕ	$L = \sum_{ij} X_{ij} = X_{++}$
Mutuality	ρ	$M = \sum_{i<j} X_{ij} X_{ji}$
Triadic		
Transitivity	τ_T	$T_T = \sum_{i,j,k} X_{ij} X_{jk} X_{ik}$

ECONOMETRIC METHODS - I

Exponential Random Graph Models (ERGMs)

(Wasserman, Pattison, 1997; Anderson, Wasserman, Crouch, 1999)

Frank, Strauss, 1986

"Markov Graphs"

$$P(\mathbf{X}=\mathbf{x}) = \frac{\exp(\theta_1 z_1(\mathbf{x}) + \cdots + \theta_r z_r(\mathbf{x}))}{\kappa(\theta)}$$

$\log[\Pr(\mathbf{X}=\mathbf{x})]$ is proportional to $\theta_1 z_1(\mathbf{x}) + \cdots + \theta_r z_r(\mathbf{x})$

In practice estimating κ(θ) is generally untractable

- logit models:
  considering the odds
  that link i-j is present

$$\omega_{ij} = \log\left(\frac{P(\mathbf{x}_{ij}^{+})}{P(\mathbf{x}_{ij}^{-})}\right) = \sum_{p=1}^{r} \theta_p (z_p(\mathbf{x}_{ij}^{+}) - z_p(\mathbf{x}_{ij}^{-}))$$

Dyadic		
Choice	ϕ	$L = \sum_{ij} X_{ij} = X_{++}$
Mutuality	ρ	$M = \sum_{i<j} X_{ij} X_{ji}$
Triadic		
Transitivity	τ_T	$T_T = \sum_{i,j,k} X_{ij} X_{jk} X_{ik}$

ECONOMETRIC METHODS - I

Exponential Random Graph Models (ERGMs)

(Wasserman, Pattison, 1997; Anderson, Wasserman, Crouch, 1999)

Frank, Strauss, 1986

"Markov Graphs"

$\cdots + \theta_r z_r(\mathbf{x})$

In pr

lo

co

th

$p(\mathbf{x}_{ij}^{-}))$

Variable	Parameter	Estimated value	Standard error
Choice	$\phi_{3\text{rd}}$	0.54	0.68
	$\phi_{4\text{th}}$	2.56	0.58
	$\phi_{5\text{th}}$	1.44	0.74
Mutuality	$\rho_{3\text{rd}} = \rho_{4\text{th}} = \rho_{5\text{th}}$	1.81	0.20
	$\rho_{5\text{th,gg}}$	2.74	1.10
Degree Centralization	$\alpha_{5\text{th}}$	4.37	1.78
Acceptance	$\gamma_{3\text{rd}} = \gamma_{5\text{th}}$	1.32	0.17
Ratings	$\gamma_{4\text{th}}$	0.62	0.17
Transitivity	$\tau_{T,3\text{rd,gg}} = \tau_{T,3\text{rd,bb}} = \tau_{T,3\text{rd,gb}} = \tau_{T,4\text{th}}$	0.28	0.02
	$\tau_{T,5\text{th}}$	0.55	0.06

Dyadic		
Choice	ϕ	$L = \sum_{ij} X_{ij} = X_{++}$
Mutuality	ρ	$M = \sum_{i<j} X_{ij} X_{ji}$
Triadic		
Transitivity	τ_T	$T_T = \sum_{i,j,k} X_{ij} X_{jk} X_{ik}$

ECONOMETRIC METHODS - II

"Stochastic actor-oriented model":

(Snijders, 2001; see also the SIENA package at http://www.stats.ox.ac.uk/~snijders/siena)

In the case of dynamic networks,
we assume an objective function which agents try to optimize:

- which depends on each agent **i** and a set of agent-centered parameterized observables $\mathbf{s_{i,p}(X)}$

$$\boxed{f_i(\mathbf{X}, \theta) = \sum_{p=1}^{r} \theta_p s_{i,p}(\mathbf{X})}$$

- assuming the process is a Markov Chain:
  at each step, an actor may (myopically) change an outgoing link, optimizing her objective function (plus an i.i.d. "random utility" component)

- estimate the parameter vector $\boldsymbol{\theta}$ that explains best **relation changes**

ECONOMETRIC METHODS - II

(Snijders, 2001; see also the SIENA package at http://www.stats.ox.ac.uk/~snijders/siena)

TABLE 1

Parameters and Standard Errors for Models Estimated Using Observations at t_1, t_2, t_3

	Model 1		Model 2		Model 3	
Effect	Par.	(s.e.)	Par.	(s.e.)	Par.	(s.e.)
Rate (period 1)	3.87		3.78		3.91	
Rate (period 2)	3.10		3.14		3.07	
Density	−1.48	(0.30)	−1.05	(0.19)	−1.13	(0.22)
Reciprocity	1.98	(0.31)	2.44	(0.40)	2.52	(0.37)
Transitivity	0.21	(0.11)	—		—	
Balance	−0.33	(0.66)	—		—	
Indirect relations	−0.347	(0.074)	−0.557	(0.083)	−0.502	(0.084)
Gender activity	—		—		−0.60	(0.28)
Gender popularity	—		—		0.64	(0.24)
Gender dissimilarity	—		—		−0.42	(0.24)

A BRIEF TAXONOMY...

reconstructing / using	processes	structure
processes	Preferential attachment Link prediction, classifiers Scoring methods	PA-based models Rewiring models Cost optimization Agent-based models
structure	ERGMs, p1, p* Markov graphs SOAMs	Prescribed structure, edge swaps Subgraph-based Kronecker graphs

PRESCRIBED STRUCTURAL FEATURES

Random graphs with **prescribed degree distributions**

- a.k.a. "configuration model"
- using generating functions

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k$$

$$G_0(1) = \sum_k p_k = 1$$

$$G_0'(1) = \sum_k k p_k = \langle k \rangle$$

(Newman, Strogatz, Watts, 2001)

$$\boxed{\langle s \rangle = 1 + \frac{G_0'(1)}{1 - G_1'(1)}}$$

mean component size

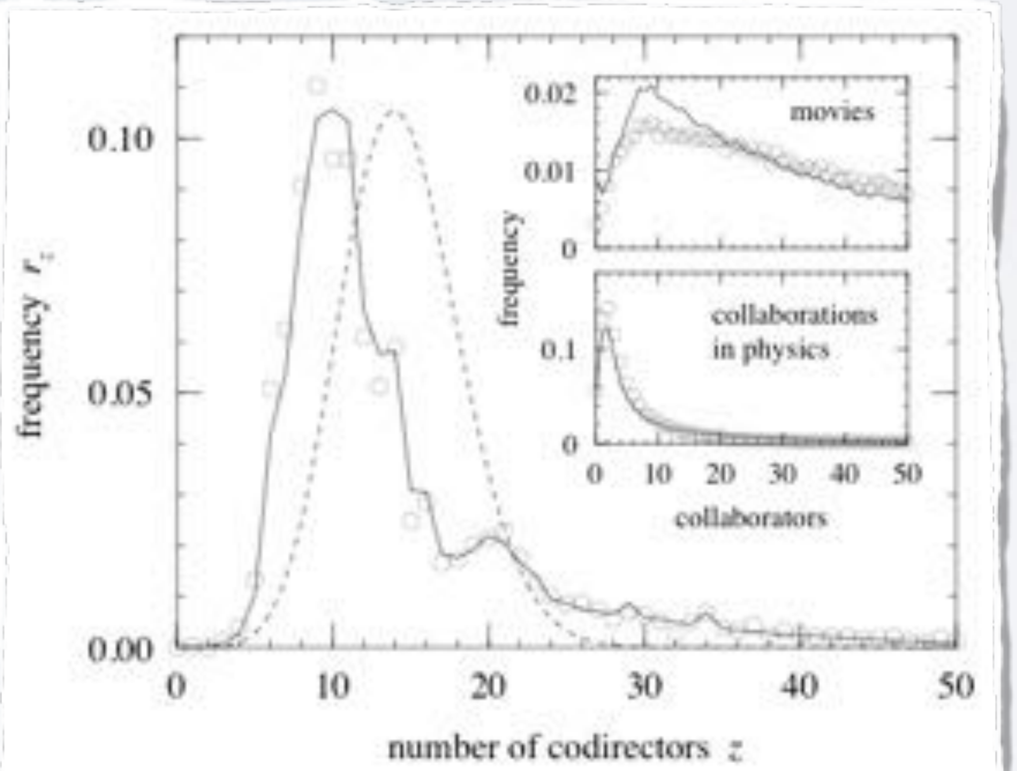


FIG. 9. The probability distribution of numbers of co-directors in the Fortune 1000 graph. The points are the real-world data, the solid line is the bipartite graph model, and the dashed line is the Poisson distribution with the same mean. Insets: the equivalent distributions for the numbers of collaborators of movie actors and physicists.

PRESCRIBED STRUCTURAL FEATURES

Random graphs with **prescribed degree distributions**

- a.k.a. "configuration model"
- using generating functions

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k$$

$$G_0(1) = \sum_k p_k = 1$$

$$G_0'(1) = \sum_k k p_k = \langle k \rangle$$

(Newman, Strogatz, Watts, 2001)

$$\boxed{\langle s \rangle = 1 + \frac{G_0'(1)}{1 - G_1'(1)}}$$

mean component size

Random graphs with **prescribed subgraph distributions**

- closing the loop by reusing the generation function formalism

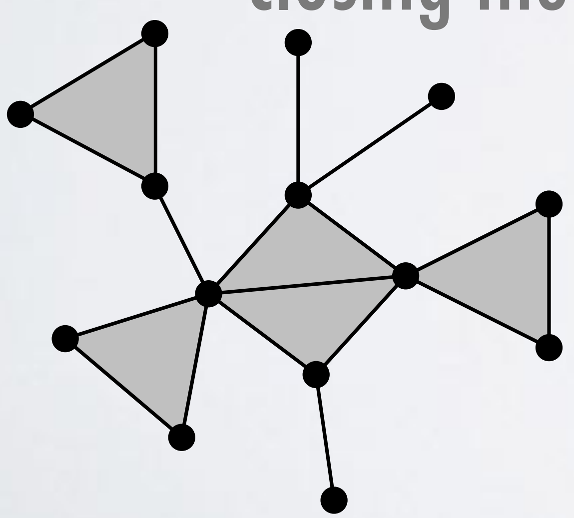


FIG. 1: A small network made of single edges, triangles, and "diamond" subgraphs composed of two overlapping triangles.

(Karrer, Newman, 2010)

generating function

$$G_0(z_1, z_2) = \sum_{st} p(s,t)\, z_1^s z_2^t$$

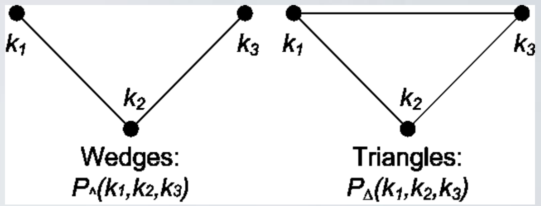
mean component size

$$\boxed{S = 1 - (1-2a)\mathrm{e}^{-c_1 S - c_2 S(2-S)} - 2a\mathrm{e}^{-4c_1 S - 4c_2 S(2-S)}}$$

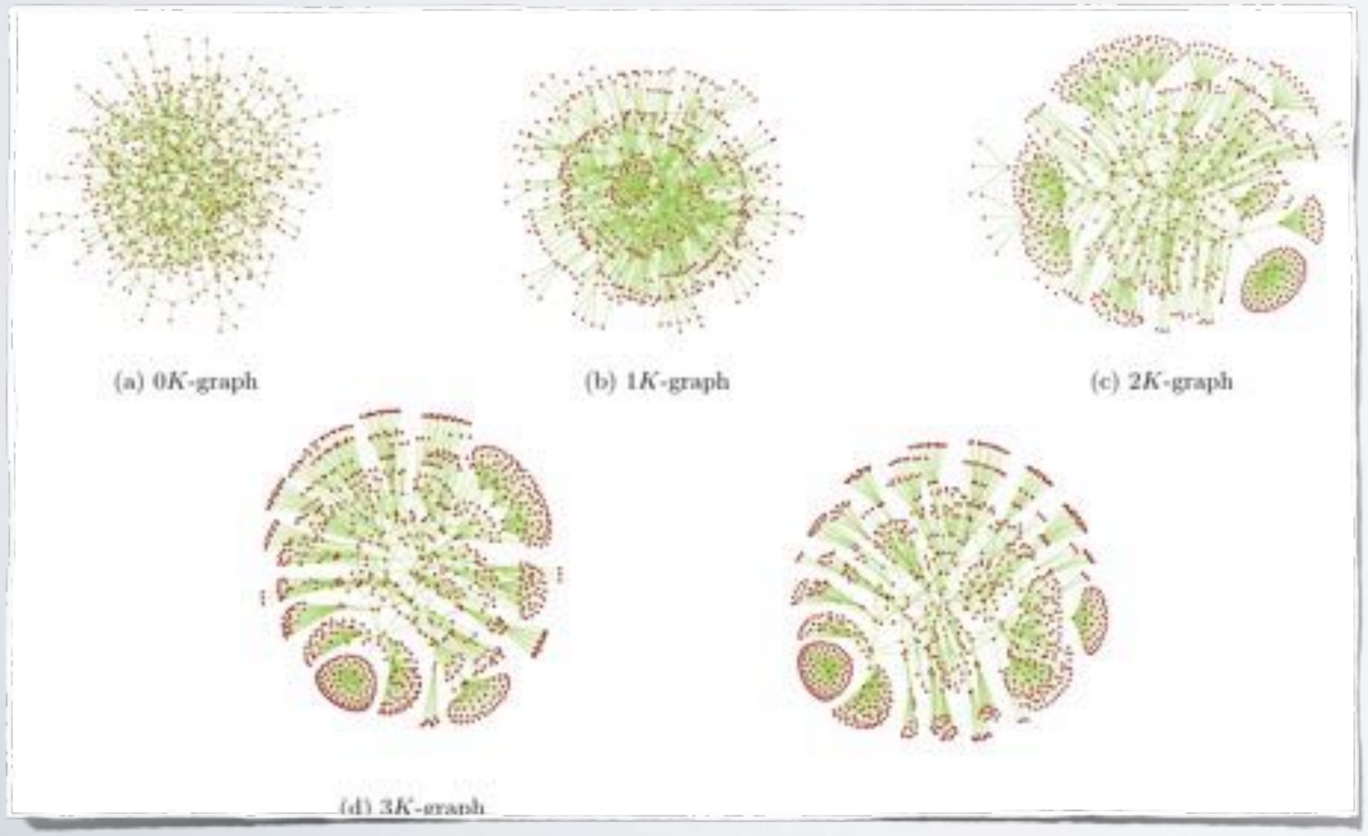
PRESCRIBED STRUCTURAL FEATURES

Random graphs with **prescribed degree correlations**

3K



- class of **dK-graphs** preserving node degree correlations
  within subgraphs of size d: 0K preserves average degree, 1K degree distribution, 2K degree correlations, etc.
- increasingly precise, and reproduces assortativity, clustering, distance and Laplacian eigenvalues as early as 2K



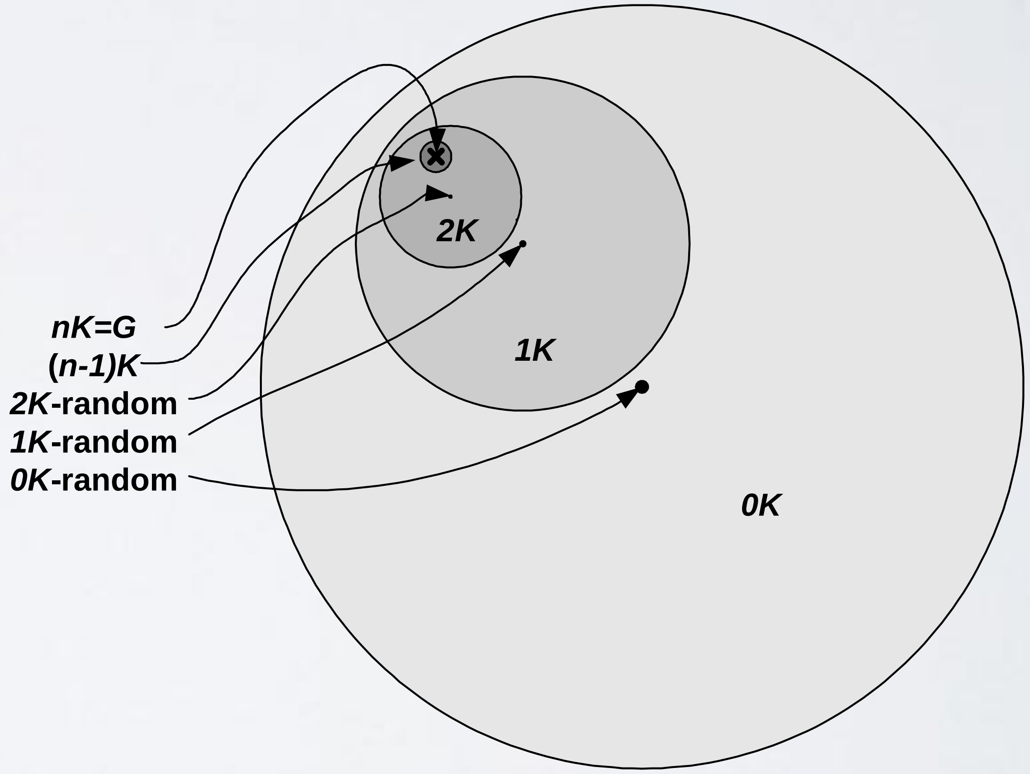
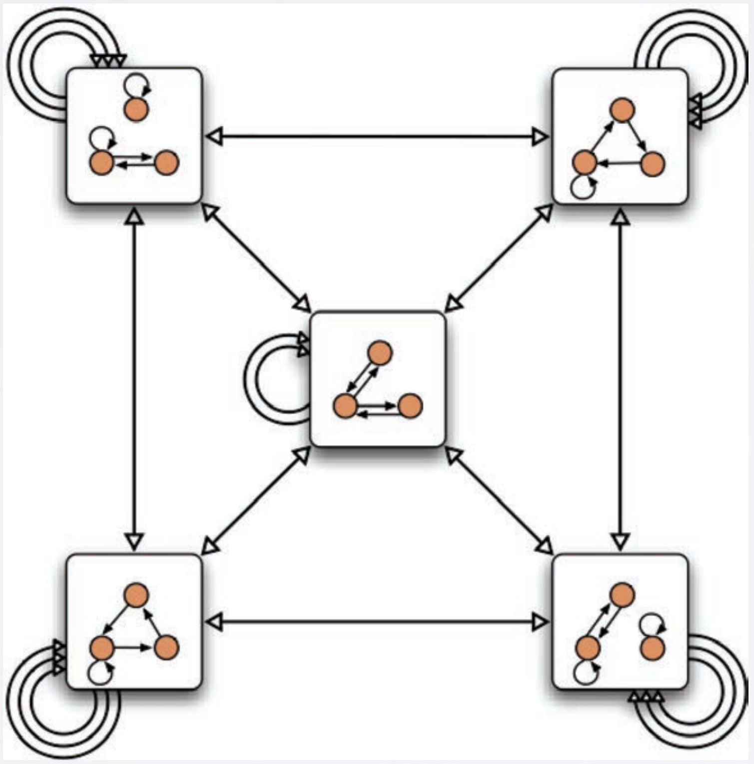


Figure 2: The dK- and dK-random graph hierarchy.

(Mahadevan, Krioukov, Fall, Vahdat, 2006)

PRESCRIBED GRAPH CONSTRAINTS

- Exploring a graph space with **prescribed constraints**
  - typically using edge swaps for degree-preserving constraints



Rao et al. 1996; Kannan et al. 1997;
Stauffer and Barbosa 2005; Cooper et al. 2006;
Feder et al. 2006; Mahadevan et al. 2006;
Bansal et al. 2008 ...

Fig. 1. Simple Markov graph for a constraint on a graph of (i) three nodes with (ii) given in- and out-degree distributions and (iii) without multiple edges but possibly self-loops. Nonvalid swaps are represented by self-loops in this Markov graph, which has thus a constant degree.

PRESCRIBED GRAPH CONSTRAINTS

- Exploring a graph space with **prescribed constraints**
  - typically using edge swaps for degree-preserving constraints
  - or higher-level constraints using so-called “k-edge swaps”

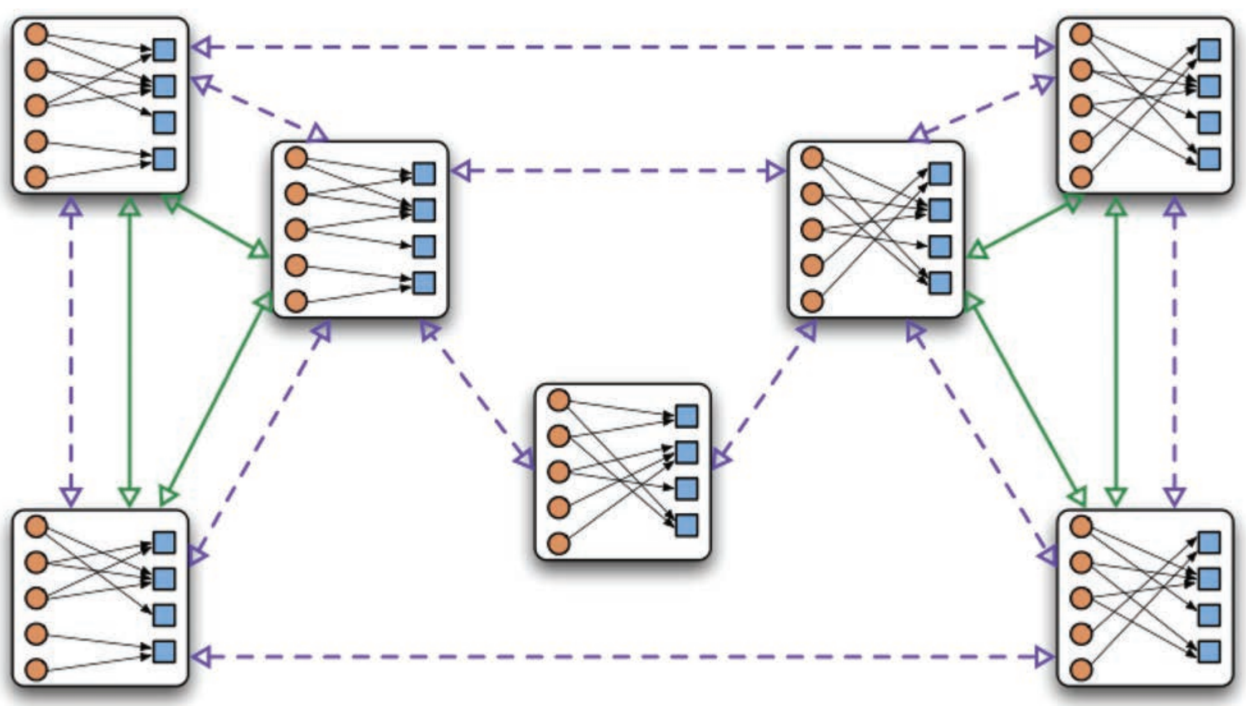


Fig. 3. Markov graph of $\mathcal{G}_{\mathbf{C}_0}$ for various k-switching procedures: dashed blue arrows correspond to $k = 2$, plain green arrows to $k = 4$. For readability purposes, we simplified the representation by discarding self-loops and multiple edges of the Markov graph.

(Tabourier, Cointet, Roth, 2011, 2016)

PRESCRIBED GRAPH CONSTRAINTS

Tabourier, Cointet, Roth, 2011, 2016

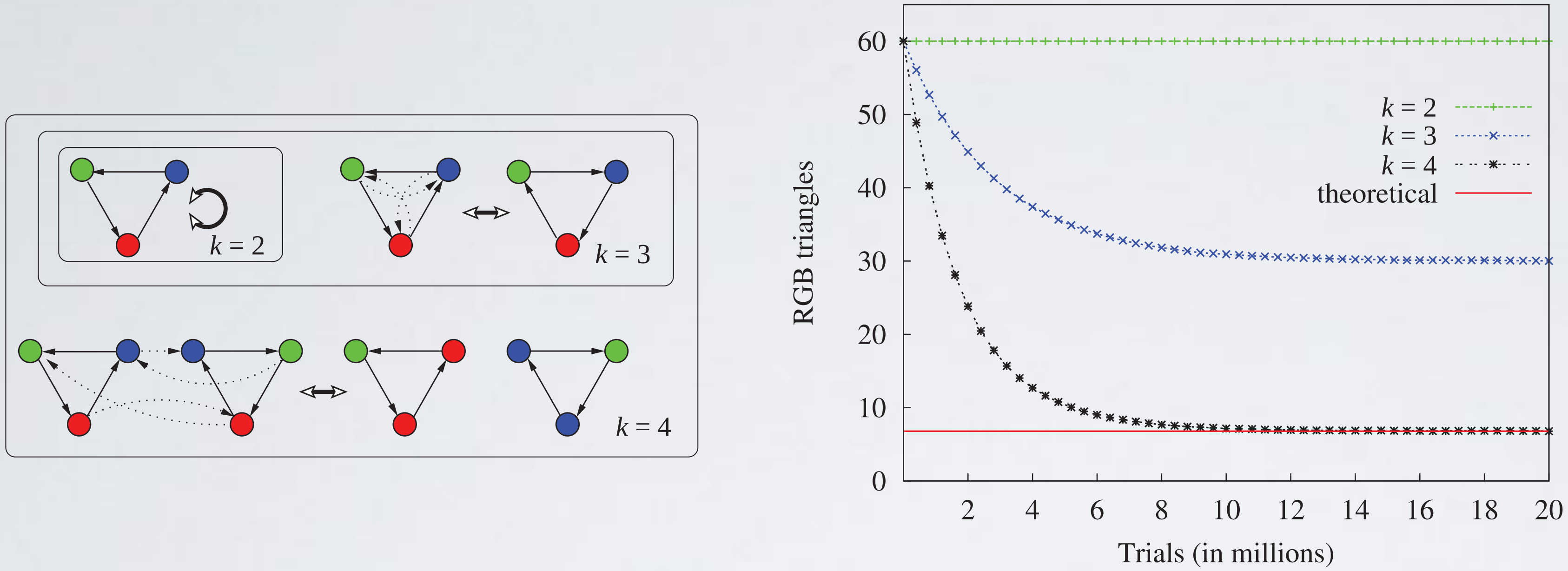


Fig. 4. Left: Illustration of the increasing possibilities of k-switches for $k \in \{2, 3, 4\}$ in the case of "R-B-G" triangles. Right: Number of "R-B-G" triangles with respect to the number of k-switch trials, for $k \in \{2, 3, 4\}$ (averages and corresponding confidence intervals computed over 10,000 simulations for each k).

PRESCRIBED GRAPH CONSTRAINTS

Tabourier, Cointet, Roth, 2011, 2016

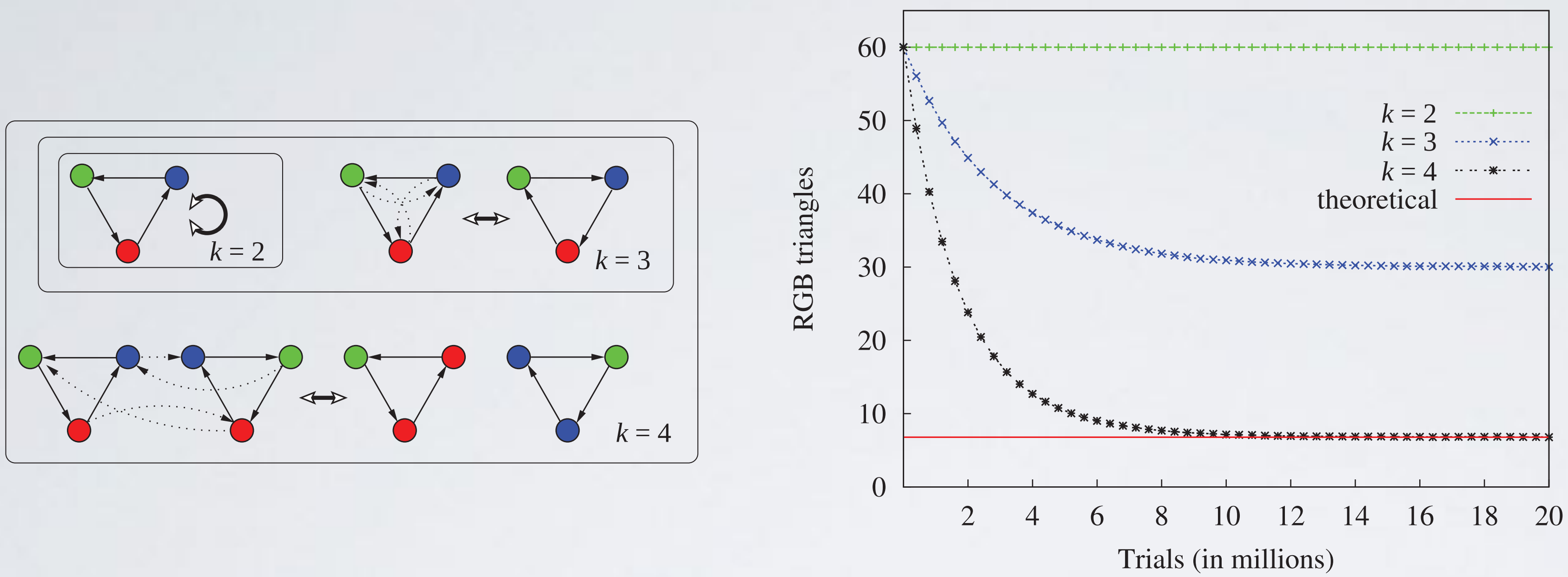


Fig. 4. Left: Illustration of the increasing possibilities of k-switches for $k \in \{2, 3, 4\}$ in the case of "R-B-G" triangles. Right: Number of "R-B-G" triangles with respect to the number of k-switch trials, for $k \in \{2, 3, 4\}$ (averages and corresponding confidence intervals computed over 10,000 simulations for each k).

—$\mathbf{C}_3^{\emptyset}$. The graph is undirected, with a fixed degree distribution, has no multiple edges nor self-loops.
—$\mathbf{C}_3^{+}$. The number of (undirected) triangles remains the same.

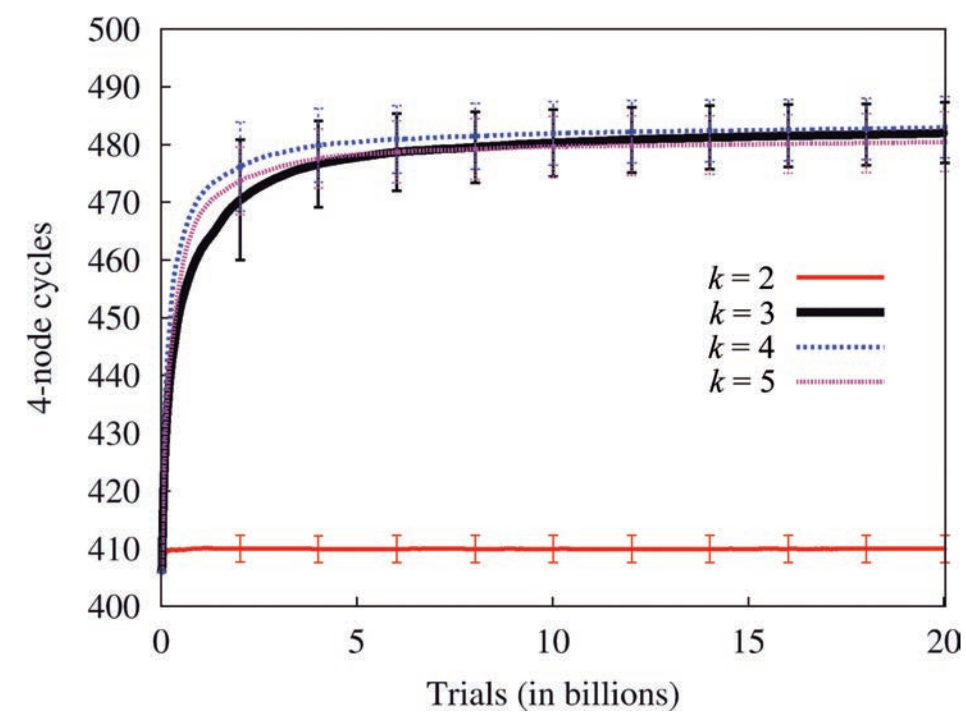


Fig. 6. Cumulative mean number of 4-nodes cycles for $\mathbf{C}_3$.

KRONECKER GRAPHS

$N_k = (N_1)^k$ and $E_k = (E_1)^k$

$$\mathbf{C} = \mathbf{A} \otimes \mathbf{B} \doteq \begin{pmatrix} a_{1,1}\mathbf{B} & a_{1,2}\mathbf{B} & \dots & a_{1,m}\mathbf{B} \\ a_{2,1}\mathbf{B} & a_{2,2}\mathbf{B} & \dots & a_{2,m}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1}\mathbf{B} & a_{n,2}\mathbf{B} & \dots & a_{n,m}\mathbf{B} \end{pmatrix}$$

Kronecker product of two matrices

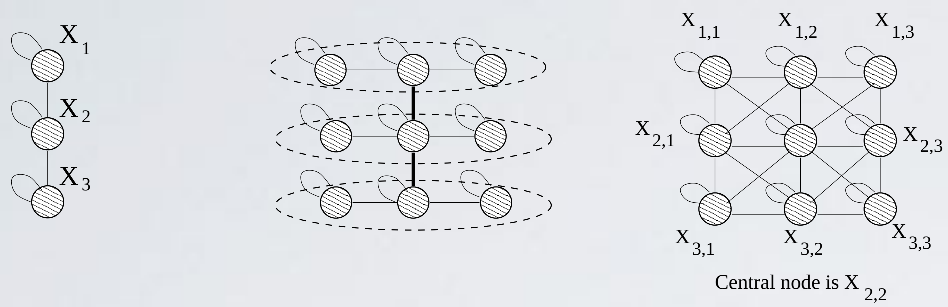
(Leskovec, Chakrabarti, Kleinberg, Faloutsos, Ghahramin, 2010)

KRONECKER GRAPHS

$N_k=(N_1)^k$ and $E_k=(E_1)^k$

$$\mathbf{C} = \mathbf{A} \otimes \mathbf{B} \doteq \begin{pmatrix} a_{1,1}\mathbf{B} & a_{1,2}\mathbf{B} & \dots & a_{1,m}\mathbf{B} \\ a_{2,1}\mathbf{B} & a_{2,2}\mathbf{B} & \dots & a_{2,m}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1}\mathbf{B} & a_{n,2}\mathbf{B} & \dots & a_{n,m}\mathbf{B} \end{pmatrix}$$

Kronecker product of two matrices



(a) Graph K_1 (b) Intermediate stage (c) Graph $K_2 = K_1 \otimes K_1$

1	1	0
1	1	1
0	1	1

(d) Adjacency matrix of K_1

K_1	K_1	0
K_1	K_1	K_1
0	K_1	K_1

(e) Adjacency matrix of $K_2 = K_1 \otimes K_1$

Figure 1: *Example of Kronecker multiplication:* Top: a “3-chain” initiator graph and its Kronecker product with itself. Each of the X_i nodes gets expanded into 3 nodes, which are then linked using Observation 1. Bottom row: the corresponding adjacency matrices. See Figure 2 for adjacency matrices of K_3 and K_4.

(Leskovec, Chakrabarti, Kleinberg, Faloutsos, Ghahramin, 2010)

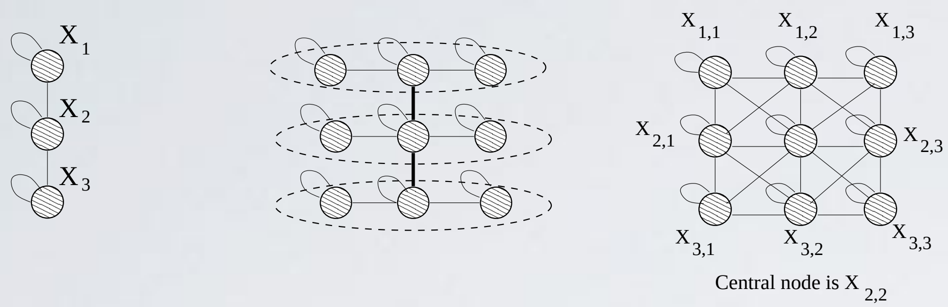
KRONECKER GRAPHS

$N_k = (N_1)^k$ and $E_k = (E_1)^k$

$$\mathbf{C} = \mathbf{A} \otimes \mathbf{B} \doteq \begin{pmatrix} a_{1,1}\mathbf{B} & a_{1,2}\mathbf{B} & \dots & a_{1,m}\mathbf{B} \\ a_{2,1}\mathbf{B} & a_{2,2}\mathbf{B} & \dots & a_{2,m}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1}\mathbf{B} & a_{n,2}\mathbf{B} & \dots & a_{n,m}\mathbf{B} \end{pmatrix}$$

Kronecker product of two matrices

(Leskovec, Chakrabarti, Kleinberg, Faloutsos, Ghahramin, 2010)



(a) Graph K_1 (b) Intermediate stage (c) Graph $K_2 = K_1 \otimes K_1$

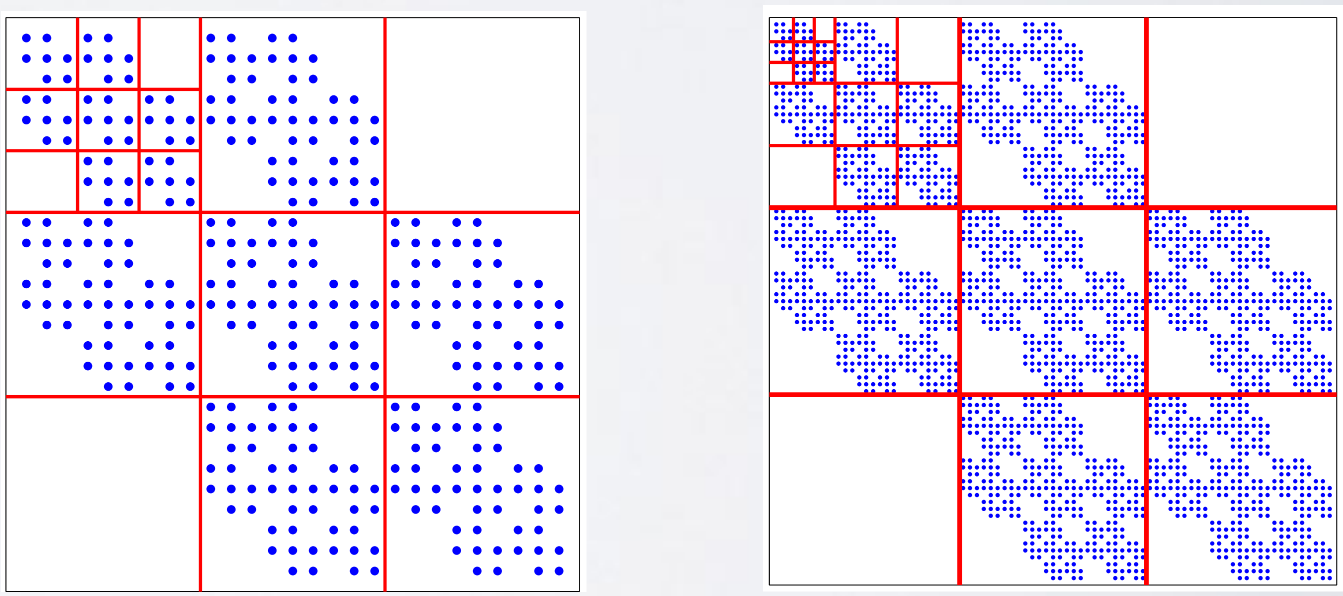
1	1	0
1	1	1
0	1	1

(d) Adjacency matrix of K_1

K_1	K_1	0
K_1	K_1	K_1
0	K_1	K_1

(e) Adjacency matrix of $K_2 = K_1 \otimes K_1$

Figure 1: *Example of Kronecker multiplication:* Top: a "3-chain" initiator graph and its Kronecker product with itself. Each of the X_i nodes gets expanded into 3 nodes, which are then linked using Observation 1. Bottom row: the corresponding adjacency matrices. See Figure 2 for adjacency matrices of K_3 and K_4.



(a) K_3 adjacency matrix (27×27) (b) K_4 adjacency matrix (81×81)

Figure 2: Adjacency matrices of K_3 and K_4, the 3^{rd} and 4^{th} Kronecker power of K_1 matrix as defined in Figure 1. Dots represent non-zero matrix entries, and white space represents zeros. Notice the recursive self-similar structure of the adjacency matrix.

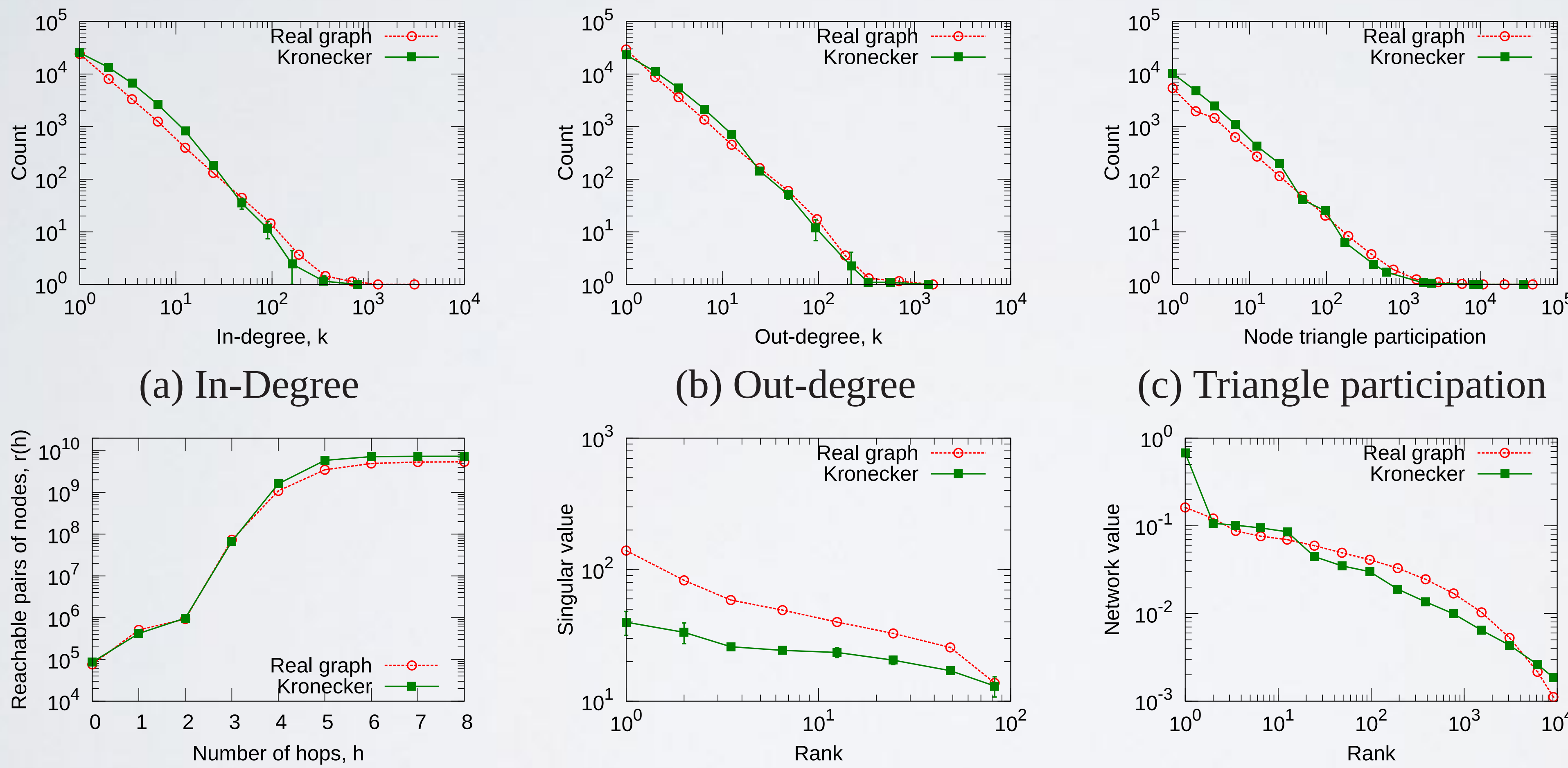
KRONECKER GRAPHS

$N_k=(N_1)^k$ and $E_k=(E_1)^k$

$$\mathbf{C} = \mathbf{A} \otimes \mathbf{B} \doteq \begin{pmatrix} a_{1,1}\mathbf{B} & a_{1,2}\mathbf{B} & \dots & a_{1,m}\mathbf{B} \\ a_{2,1}\mathbf{B} & a_{2,2}\mathbf{B} & \dots & a_{2,m}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1}\mathbf{B} & a_{n,2}\mathbf{B} & \dots & a_{n,m}\mathbf{B} \end{pmatrix}$$

Kronecker product of two matrices

(Leskovec, Chakrabarti, Kleinberg, Faloutsos, Ghahramin, 2010)



(a) In-Degree (b) Out-degree (c) Triangle participation

(d) Hop plot (e) Scree plot (f) "Network" value

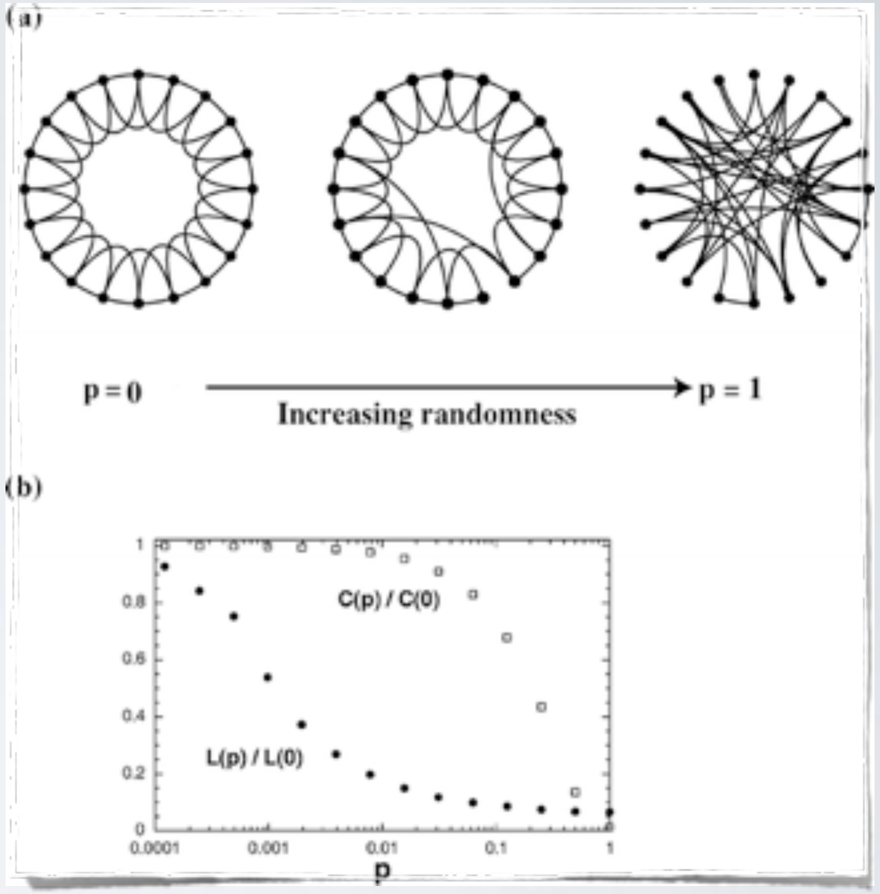
Figure 21: EPINIONS *who-trusts-whom social network:* Overlayed patterns of real network and the fitted Kronecker graph using only 4 parameters (2×2 initiator matrix). Again, the synthetic Kronecker graph matches all the properties of the real network.

A BRIEF TAXONOMY...

reconstructing / using	processes	structure
processes	Preferential attachment Link prediction, classifiers Scoring methods	PA-based models Rewiring models Cost optimization Agent-based models
structure	ERGMs, p1, p* Markov graphs SOAMs	Prescribed structure, edge swaps Subgraph-based Kronecker graphs

REWIRING / OPTIMIZATION MODELS

- Watts-Strogatz' small-world model: prescribed fixed degree, rewiring



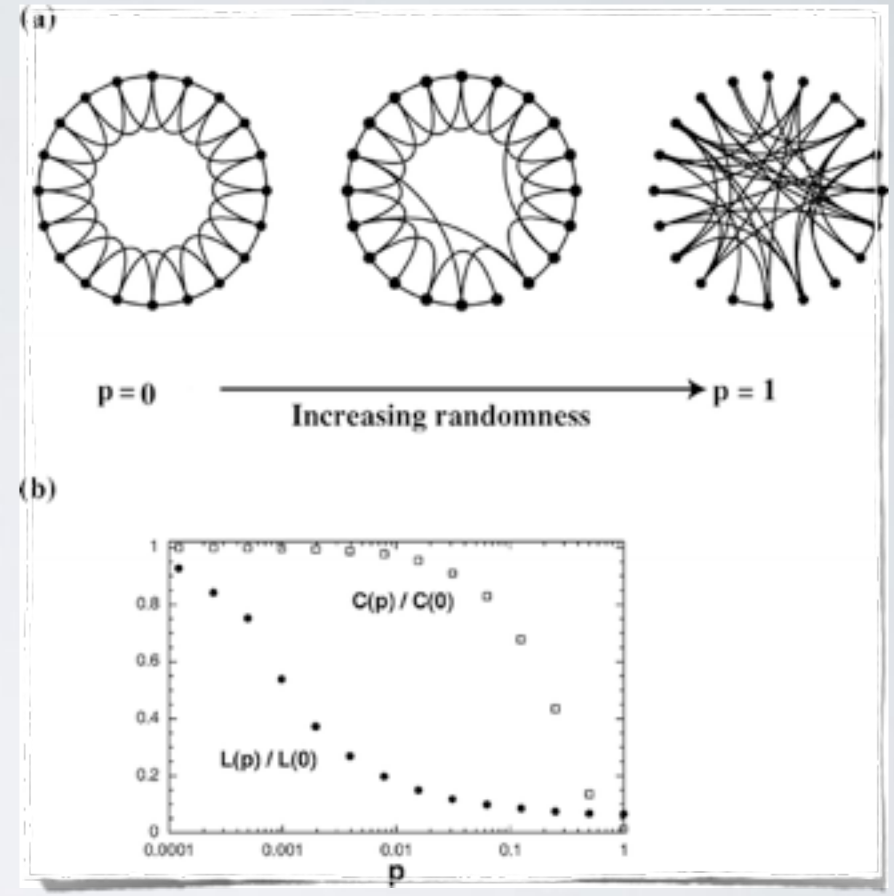
(Watts, Strogatz, 1999)

REWIRING / OPTIMIZATION MODELS

- Watts-Strogatz' small-world model:
  prescribed fixed degree, rewiring

- Fabrikant et al.' heuristically-optimized trade-off model (HOT):
  competition-based preferential attachment $min_{j<i}\alpha \cdot d_{ij} + h_j$

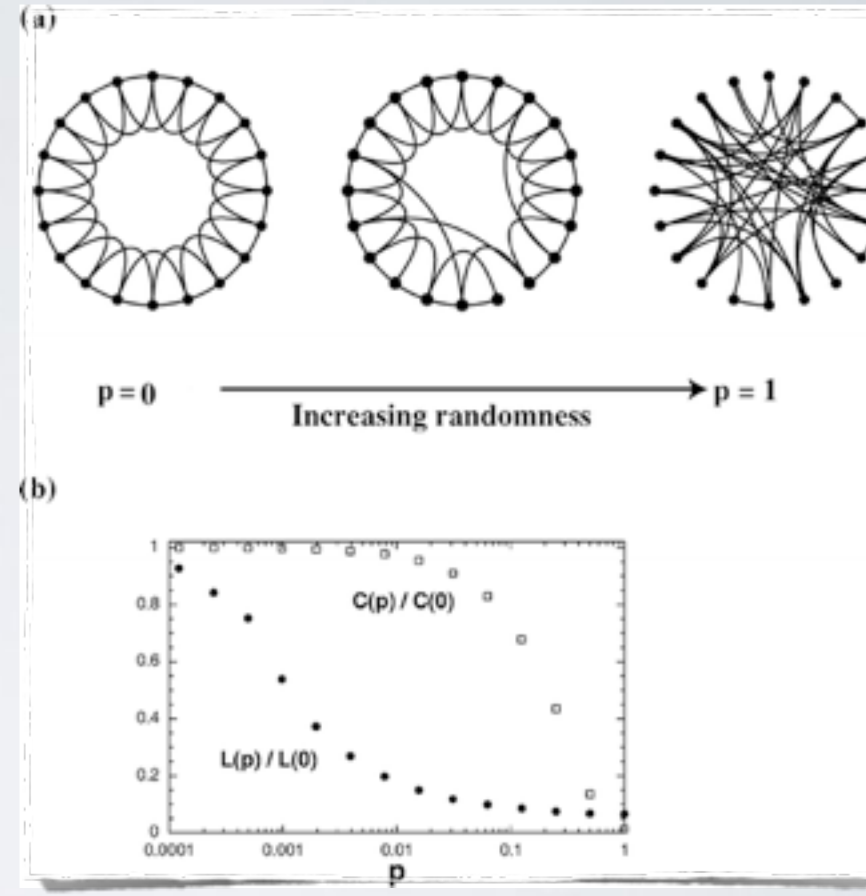
(Fabrikant et al., 2002)



(Watts, Strogatz, 1999)

REWIRING / OPTIMIZATION MODELS

- Watts-Strogatz' small-world model:
  prescribed fixed degree, rewiring

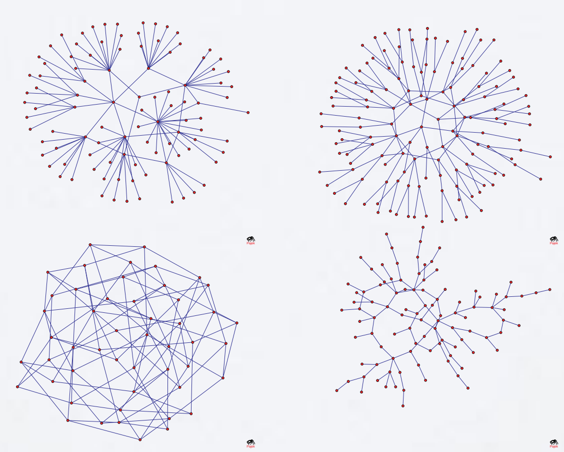


(Watts, Strogatz, 1999)

- Fabrikant et al.' heuristically-optimized trade-off model (HOT):
  competition-based preferential attachment $min_{j<i} \alpha \cdot d_{ij} + h_j$

(Fabrikant et al., 2002)

- Colizza et al.' "Network structure from selection principles"
  rewiring according to a global cost function

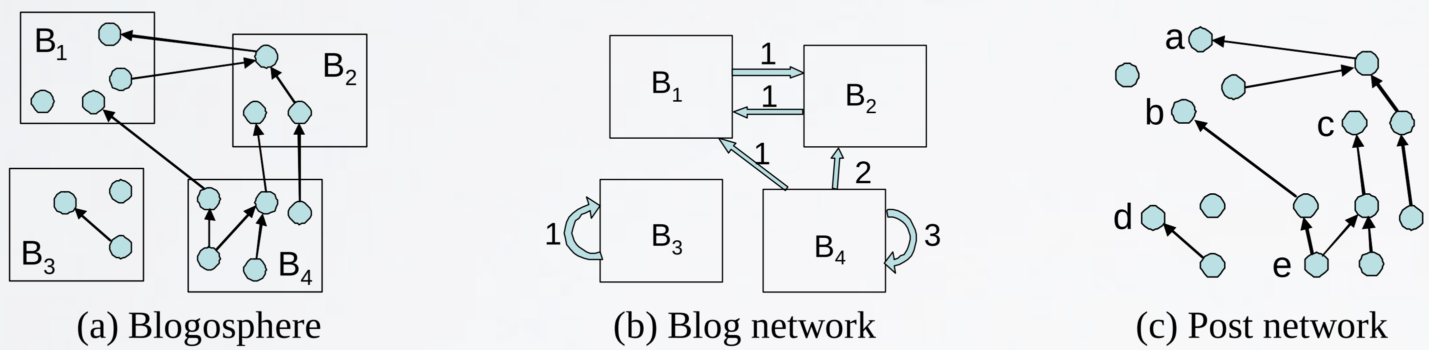


(Colizza, Banavar, Maritan, Rinaldo, 2004)

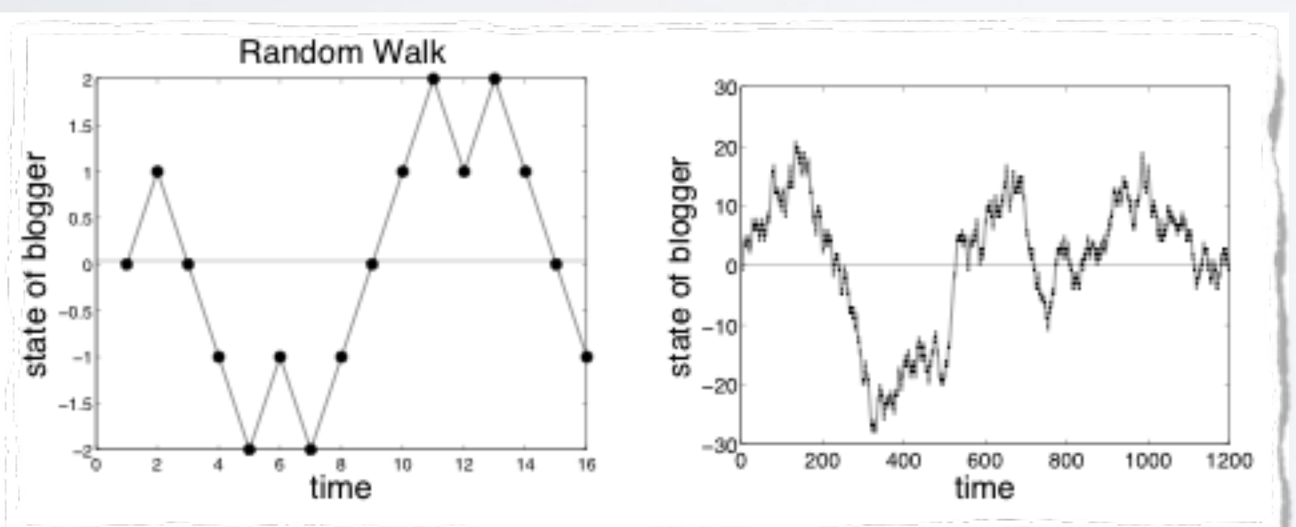
AGENT-BASED MODELS

- Specific (stochastic) rules
  - see Zero-Crossing Model

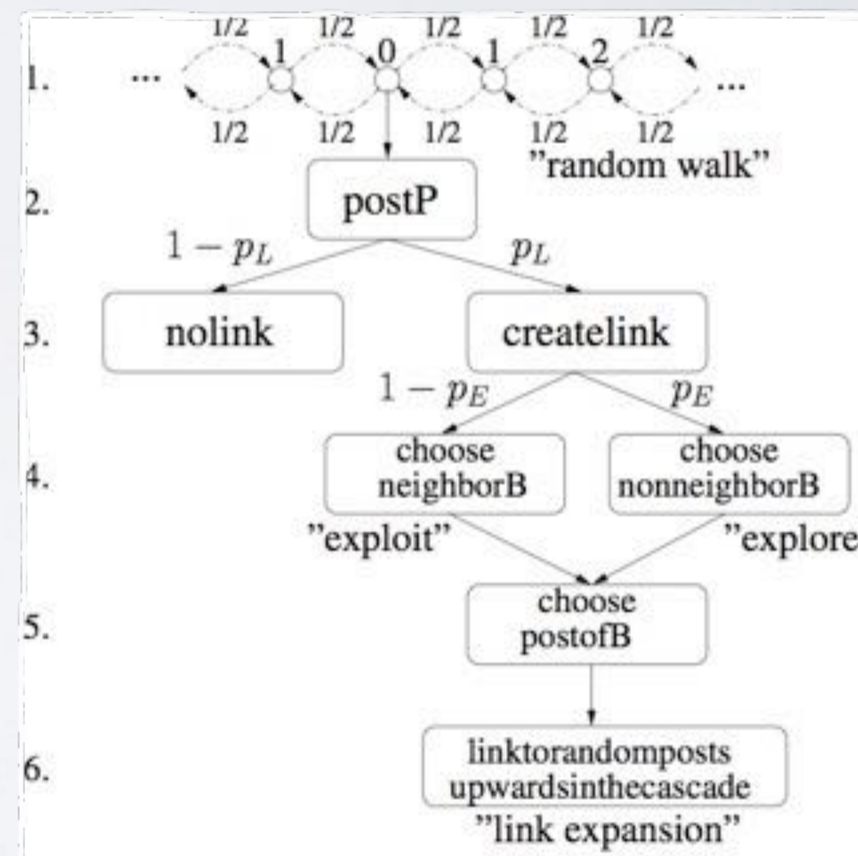
(Gotz, Leskovec, McGlohon, Faloutsos, 2009)



Network of posts and blogs



Burstiness of blog posting behavior



Model decision tree

$$\omega_i^{out}(T) \equiv \frac{\sum_j w_{ij}(T)}{k_i^{out}}$$

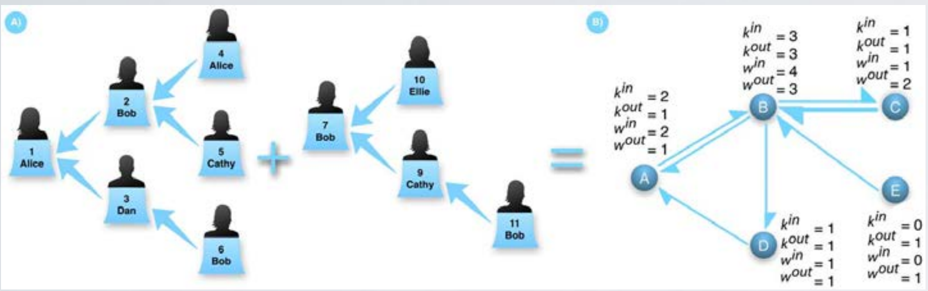


Figure 1. Reply trees and user network. A) The set of all trees is a forest. Each time a user replies, the corresponding tweet is connected to another one, resulting in a tree structure. B) Combining all the trees in the forest and projecting them onto the users results in a directed and weighted network that can be used as a proxy for relationships between users. The number of outgoing (incoming) connections of a given user is called the out (in) degree and is represented by k^{out} (k^{in}). The number of messages flowing along each edge is called the degree, w. The probability density function $P(k^{out})$ ($P(k^{int})$) indicates the probability that any given node has k^{out} (k^{in}) out (in) degree and it is called the out (in) degree distribution and is a measure of node diversity on the network.

OPEN ACCESS Freely available online

PLOS one

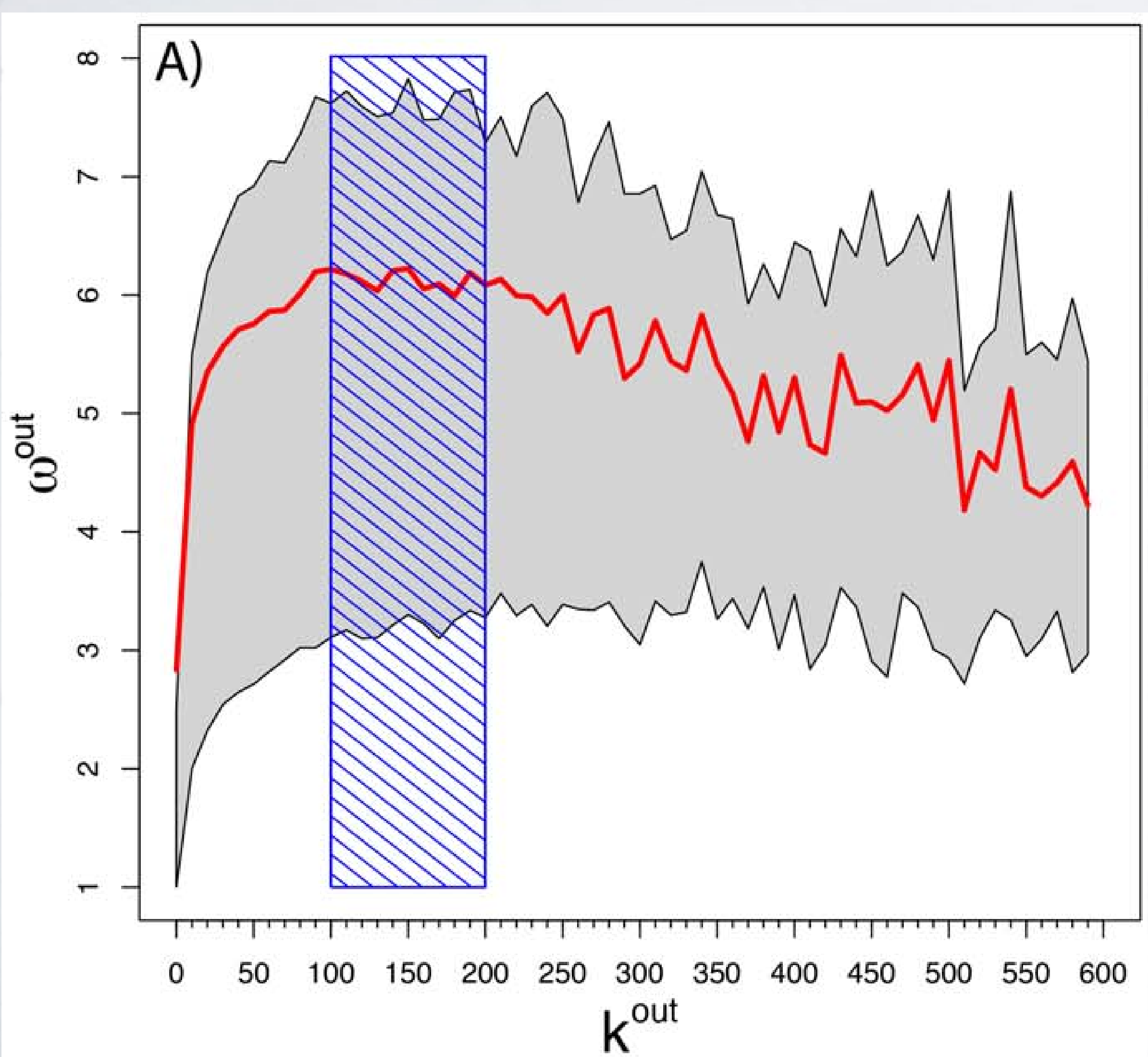
Modeling Users' Activity on Twitter Networks: Validation of Dunbar's Number

Bruno Gonçalves[1,2], Nicola Perra[1,3]*, Alessandro Vespignani[1,2,4]

1 School of Informatics and Computing, Center for Complex Networks and Systems Research, Indiana University, Bloomington, Indiana, United States of America, **2** Pervasive Technology Institute, Indiana University, Bloomington, Indiana, United States of America, **3** Complex Systems Computational Lab, Linkalab, Cagliari, Italy, **4** Institute for Scientific Interchange, Turin, Italy

Abstract

Microblogging and mobile devices appear to augment human social capabilities, which raises the question whether they remove cognitive or biological constraints on human communication. In this paper we analyze a dataset of Twitter conversations collected across six months involving 1.7 million individuals and test the theoretical cognitive limit on the number of stable social relationships known as Dunbar's number. We find that the data are in agreement with Dunbar's result; users can entertain a maximum of 100–200 stable relationships. Thus, the 'economy of attention' is limited in the online world by cognitive and biological constraints as predicted by Dunbar's theory. We propose a simple model for users' behavior that includes finite priority queuing and time resources that reproduces the observed social behavior.



$$\omega_i^{out}(T) \equiv \frac{\sum_j w_{ij}(T)}{k_i^{out}}$$

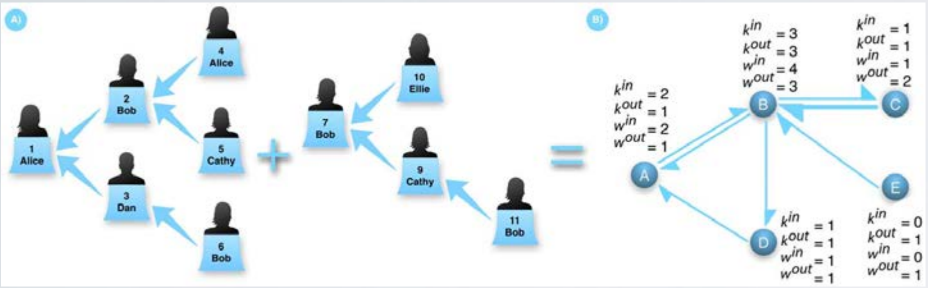


Figure 1. Reply trees and user network. A) The set of all trees is a forest. Each time a user replies, the corresponding tweet is connected to another one, resulting in a tree structure. B) Combining all the trees in the forest and projecting them onto the users results in a directed and weighted network that can be used as a proxy for relationships between users. The number of outgoing (incoming) connections of a given user is called the out (in) degree and is represented by k^{out} (k^{in}). The number of messages flowing along each edge is called the degree, w. The probability density function $P(k^{out})$ ($P(k^{int})$) indicates the probability that any given node has k^{out} (k^{in}) out (in) degree and it is called the out (in) degree distribution and is a measure of node diversity on the network.

OPEN ACCESS Freely available online

PLoS one

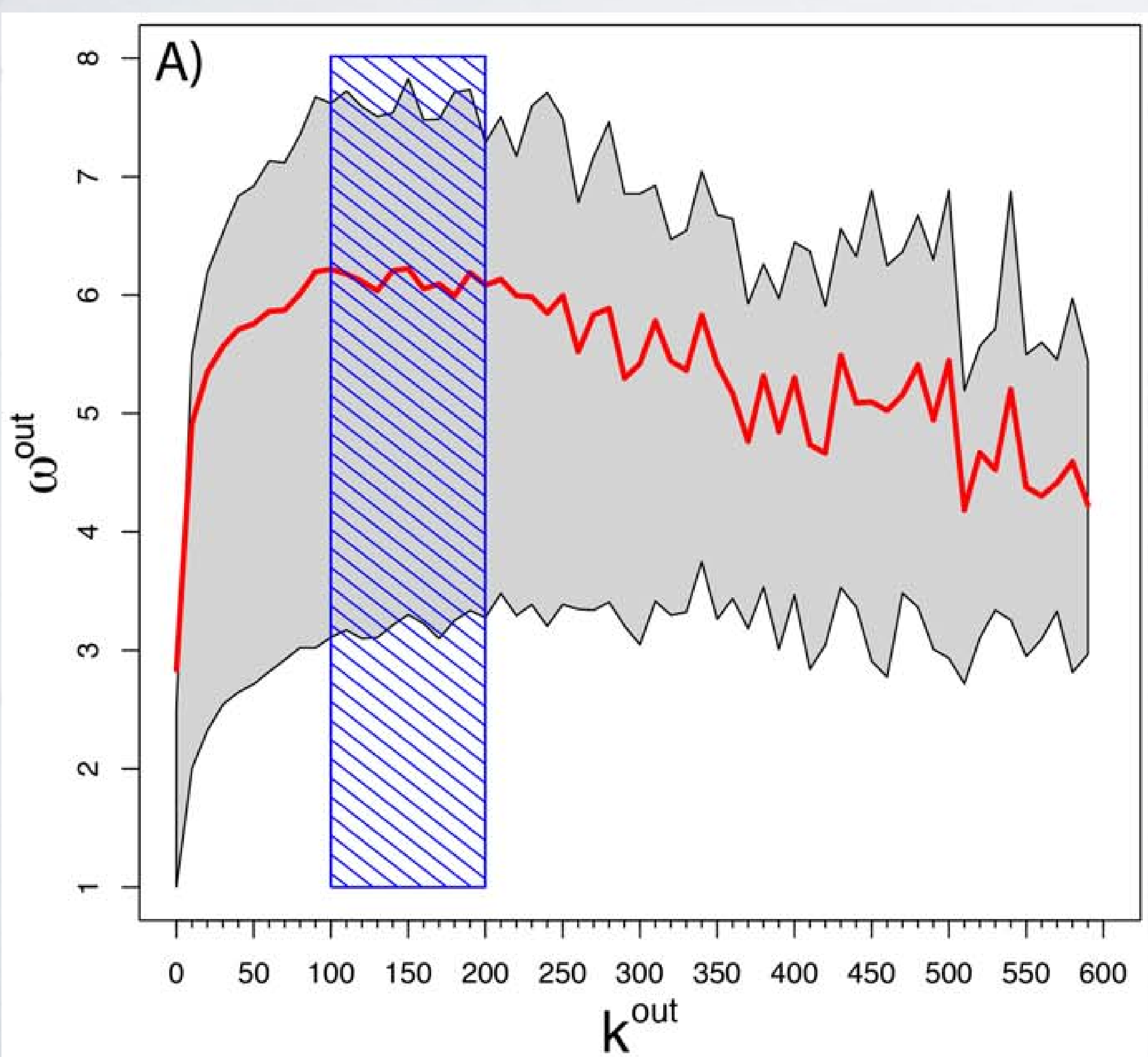
Modeling Users' Activity on Twitter Networks: Validation of Dunbar's Number

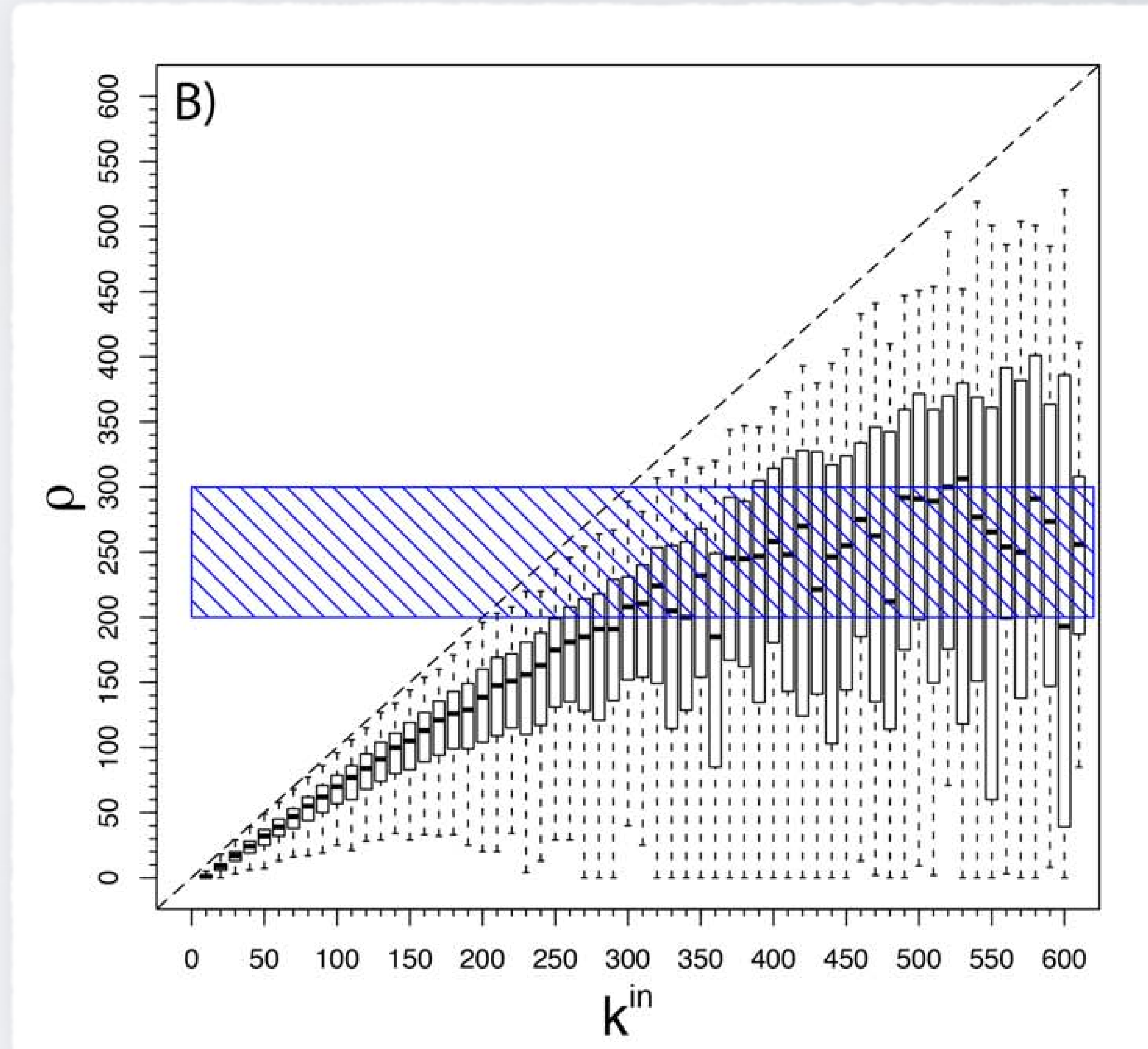
Bruno Gonçalves[1,2], **Nicola Perra**[1,3]*, **Alessandro Vespignani**[1,2,4]

1 School of Informatics and Computing, Center for Complex Networks and Systems Research, Indiana University, Bloomington, Indiana, United States of America, 2 Pervasive Technology Institute, Indiana University, Bloomington, Indiana, United States of America, 3 Complex Systems Computational Lab, Linkalab, Cagliari, Italy, 4 Institute for Scientific Interchange, Turin, Italy

Abstract

Microblogging and mobile devices appear to augment human social capabilities, which raises the question whether they remove cognitive or biological constraints on human communication. In this paper we analyze a dataset of Twitter conversations collected across six months involving 1.7 million individuals and test the theoretical cognitive limit on the number of stable social relationships known as Dunbar's number. We find that the data are in agreement with Dunbar's result; users can entertain a maximum of 100–200 stable relationships. Thus, the 'economy of attention' is limited in the online world by cognitive and biological constraints as predicted by Dunbar's theory. We propose a simple model for users' behavior that includes finite priority queuing and time resources that reproduces the observed social behavior.





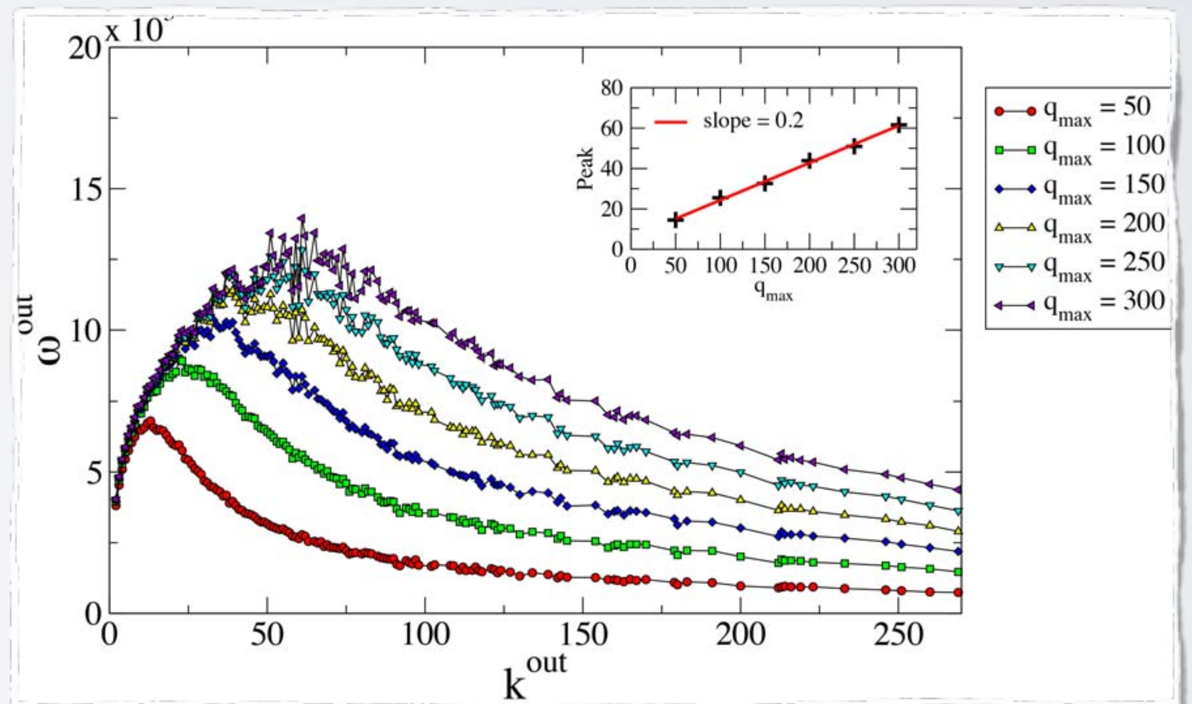
AGENT-BASED MODELS

Specific stochastic rules

see Message Queuing Models

1. each user has a message queue of some maximum size
2. they reply to a random number of messages, proportionally to the out-degree of the sender
3. the model features a simple, uniform initialization process

(Gonçalves, Perra, Vespignani, 2011)



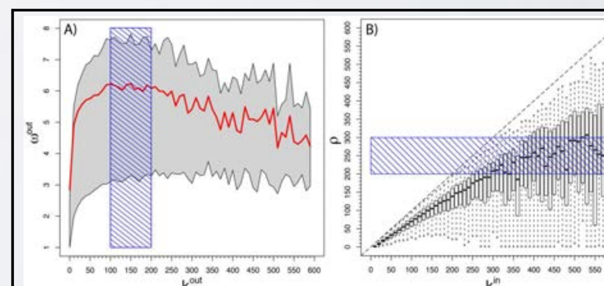


Figure 2. Connection weight and Reciprocated connections. A) Out-weight as a function of the out-degree. The average weight of each outward connection gradually increases until it reaches a maximum near 150–200 contacts, signaling that a maximum level of social activity has been reached. Above this point, an increase in the number of contacts can no longer be sustained with the same amount of dedication to each. The red line corresponds to the average out-weight, while the gray shaded area illustrates the 50% confidence interval. B) Number of reciprocated connections, ρ, as a function of k^{in}. As the number of people demanding our attention increases, it will eventually saturate our ability to reply leading to the flat behavior displayed in the dashed region.

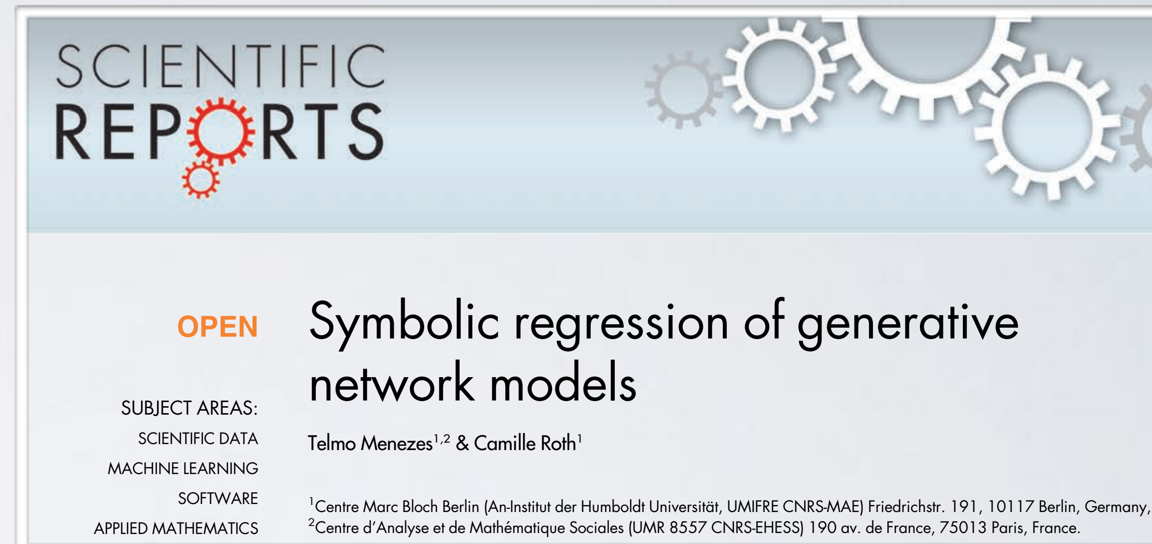
A BRIEF TAXONOMY…

reconstructing / using	processes	structure
processes	Preferential attachment Link prediction, classifiers Scoring methods	PA-based models Rewiring models Cost optimization Agent-based models
structure	ERGMs, p1, p* Markov graphs SOAMs	Prescribed structure, edge swaps Subgraph-based Kronecker graphs

Symbolic regression

SYMBOLIC REGRESSION

HOW TO PROPOSE PLAUSIBLE **GENERATIVE MODELS** FOR A GIVEN **COMPLEX**



SCIENTIFIC REPORTS

OPEN

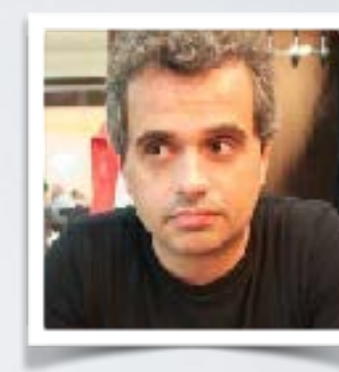
SUBJECT AREAS:
SCIENTIFIC DATA
MACHINE LEARNING
SOFTWARE
APPLIED MATHEMATICS

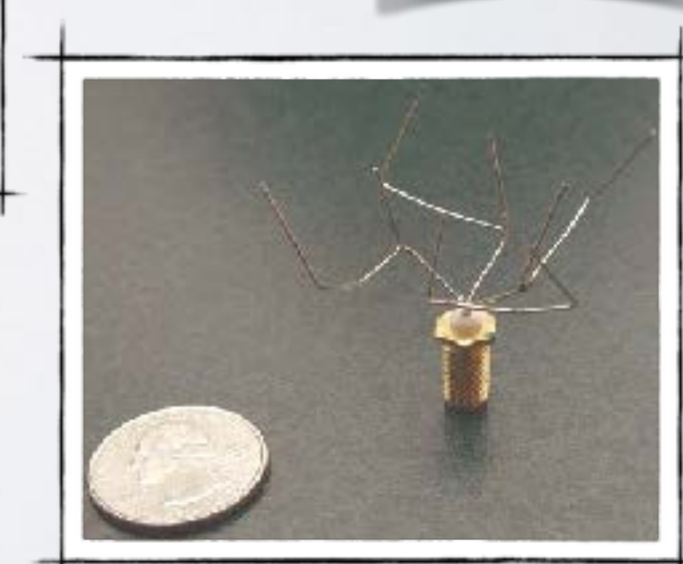
Symbolic regression of generative network models

Telmo Menezes[1,2] & Camille Roth[1]

[1]Centre Marc Bloch Berlin (An-Institut der Humboldt Universität, UMIFRE CNRS-MAE) Friedrichstr. 191, 10117 Berlin, Germany,
[2]Centre d'Analyse et de Mathématique Sociales (UMR 8557 CNRS-EHESS) 190 av. de France, 75013 Paris, France.

(Menezes, Roth, 2014)

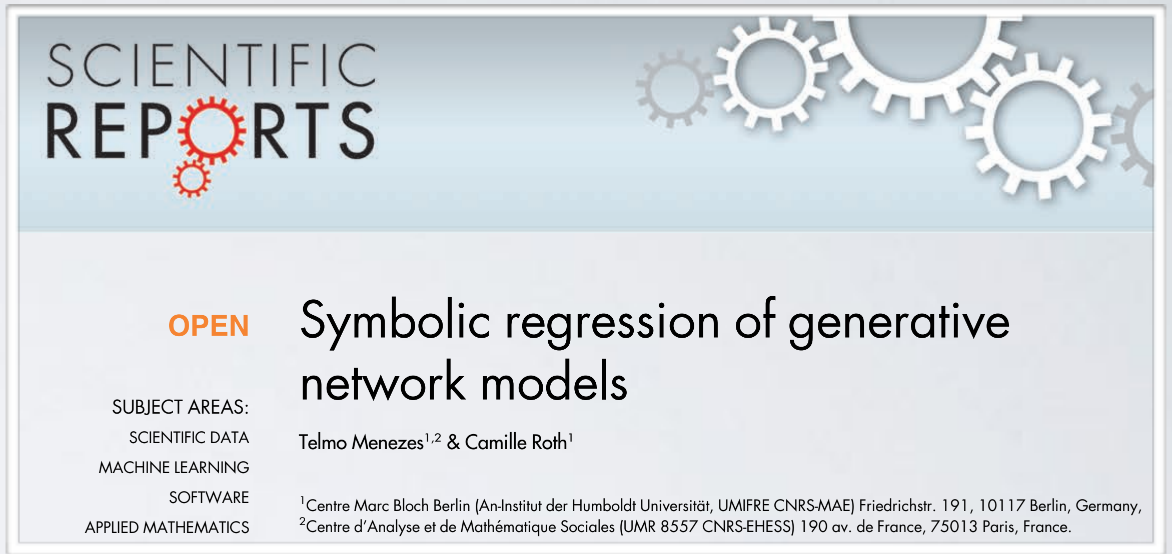




SYMBOLIC REGRESSION

HOW TO PROPOSE PLAUSIBLE **GENERATIVE MODELS** FOR A GIVEN **COMPLEX**

- designing **network models** is challenging – stylized hypotheses based on intuition
- approach based on stochastic simulation driven by **tree-based programs**
- automatic discovery (a.k.a. symbolic regression) through **genetic programming**


SCIENTIFIC REPORTS

OPEN

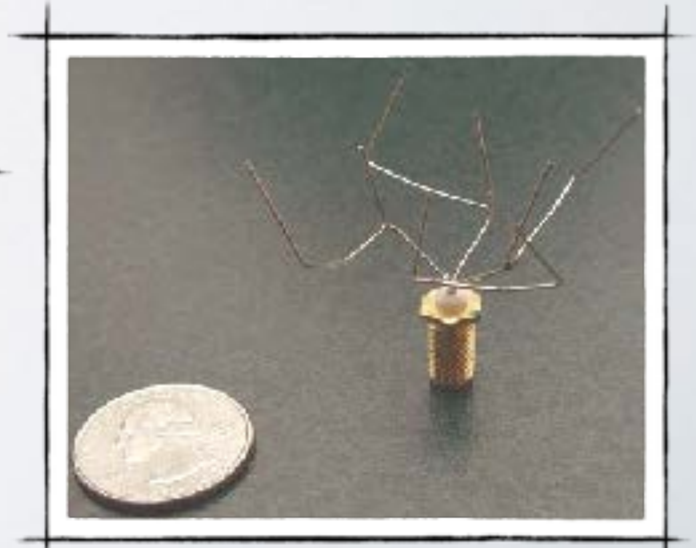
SUBJECT AREAS:
SCIENTIFIC DATA
MACHINE LEARNING
SOFTWARE
APPLIED MATHEMATICS

Symbolic regression of generative network models

Telmo Menezes[1,2] & Camille Roth[1]

[1]Centre Marc Bloch Berlin (An-Institut der Humboldt Universität, UMIFRE CNRS-MAE) Friedrichstr. 191, 10117 Berlin, Germany,
[2]Centre d'Analyse et de Mathématique Sociales (UMR 8557 CNRS-EHESS) 190 av. de France, 75013 Paris, France.

(Menezes, Roth, 2014)

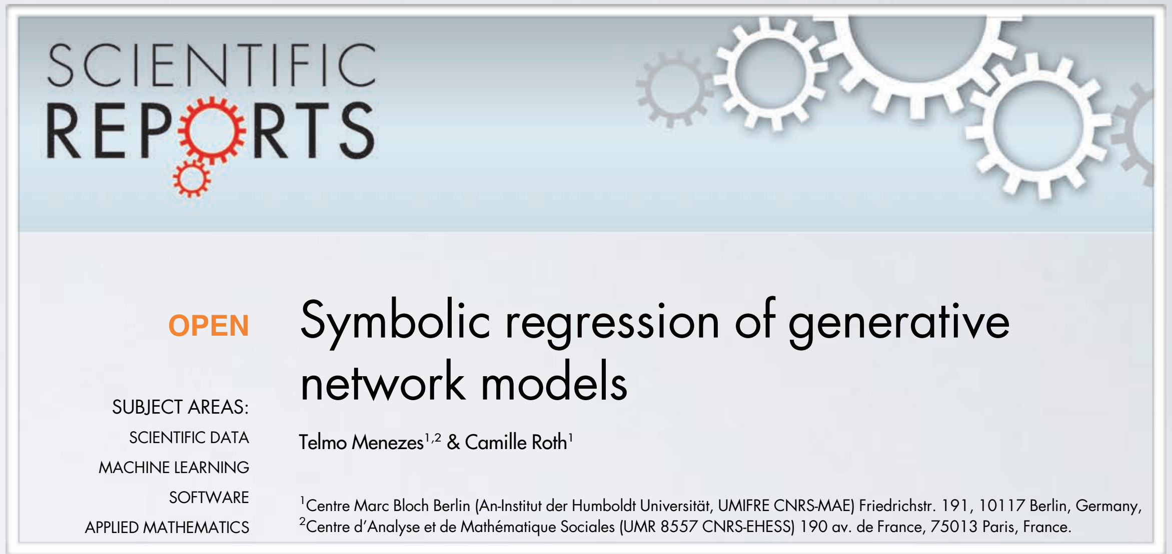


SYMBOLIC REGRESSION

HOW TO PROPOSE PLAUSIBLE **GENERATIVE MODELS** FOR A GIVEN **COMPLEX**

We need:

- a **grammar** and **vocabulary**
- comparison **metrics**
- an **evolutionary process**


SCIENTIFIC REPORTS

OPEN

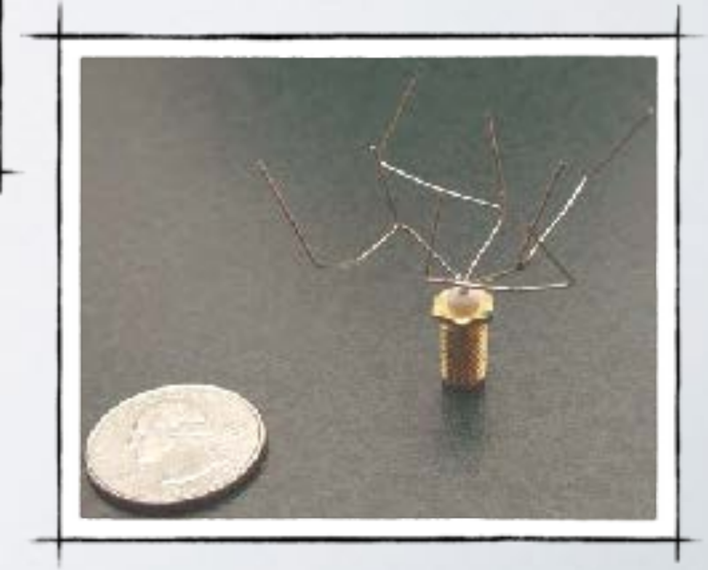
Symbolic regression of generative network models

SUBJECT AREAS:
SCIENTIFIC DATA
MACHINE LEARNING
SOFTWARE
APPLIED MATHEMATICS

Telmo Menezes[1,2] & Camille Roth[1]

[1]Centre Marc Bloch Berlin (An-Institut der Humboldt Universität, UMIFRE CNRS-MAE) Friedrichstr. 191, 10117 Berlin, Germany,
[2]Centre d'Analyse et de Mathématique Sociales (UMR 8557 CNRS-EHESS) 190 av. de France, 75013 Paris, France.

(Menezes, Roth, 2014)



NETWORK MODELS AS **TREE-BASED PROGRAMS**

$$P_{ij} = \frac{w_{ij}}{\sum_{(i',j') \in S} w_{i'j'}}$$

$w_{i,j} = \exp(4-2d)$

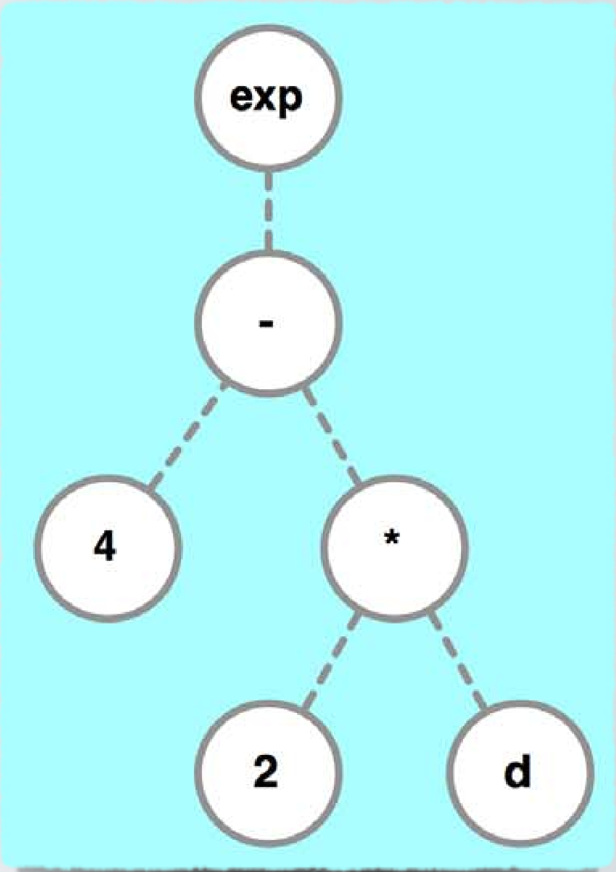
Vocabulary: the usual suspects

- in- and out-degrees **k**, **k'**
- undirected, directed and reverse distances **d**, $\mathbf{d_D}$ and $\mathbf{d_R}$
- sequential identifier **i**

Grammar: trees and operators

- +, -, *, /
- x^y, e^x, log, abs, min, max
- >, <, =, =0
- affinity function ψ

$$\psi(i,j,g,a,b) = \begin{cases} a, & \text{if } i \bmod g = j \bmod g \\ b, & \text{otherwise,} \end{cases}$$

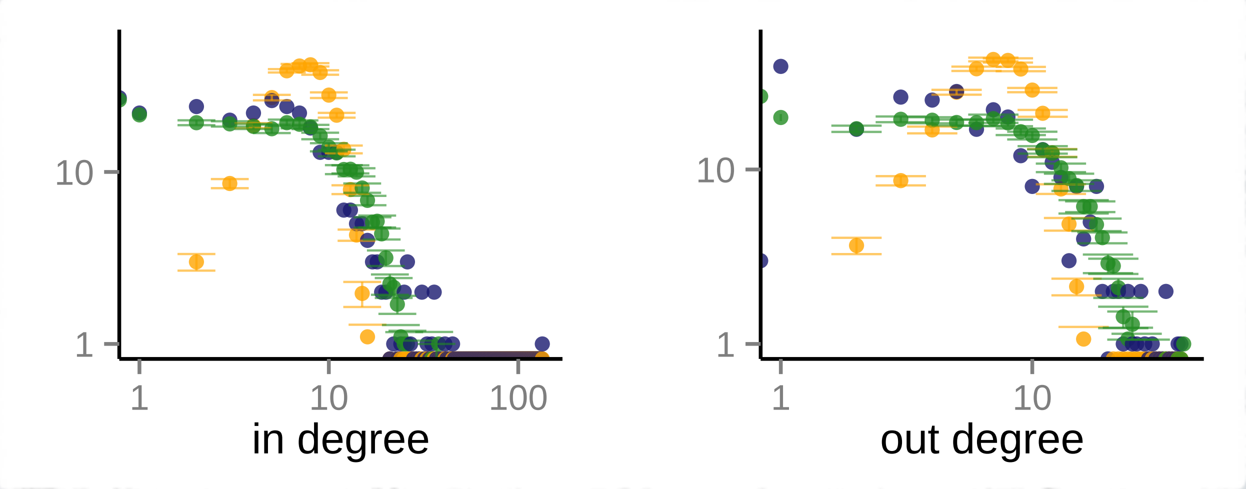


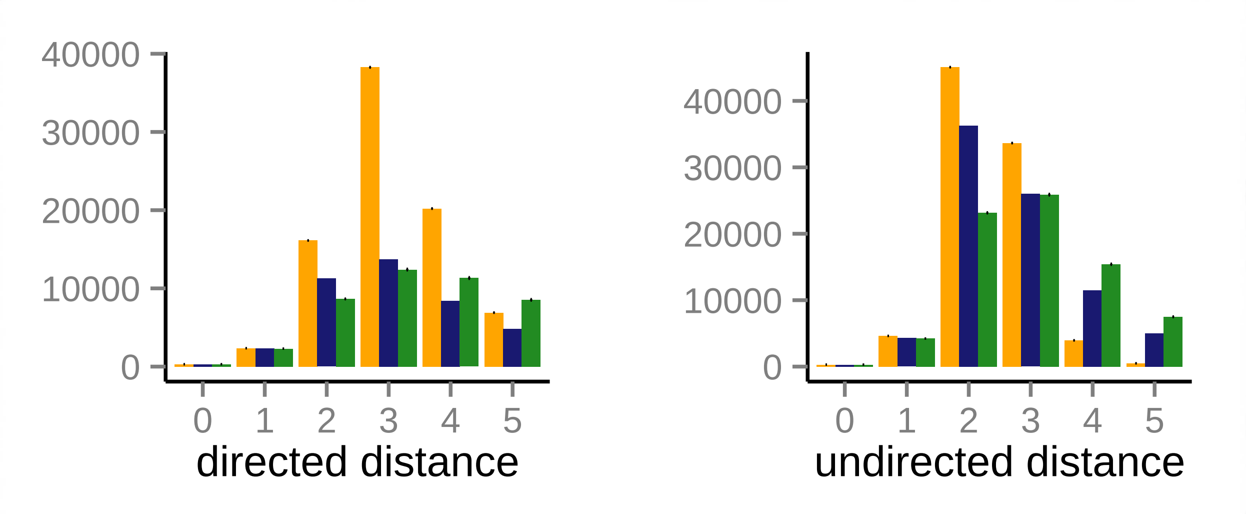
bottom-up evaluation of the tree

FITNESS FUNCTION AS **NORMALIZED NETWORK METRICS**

Metrics set: the usual suspects

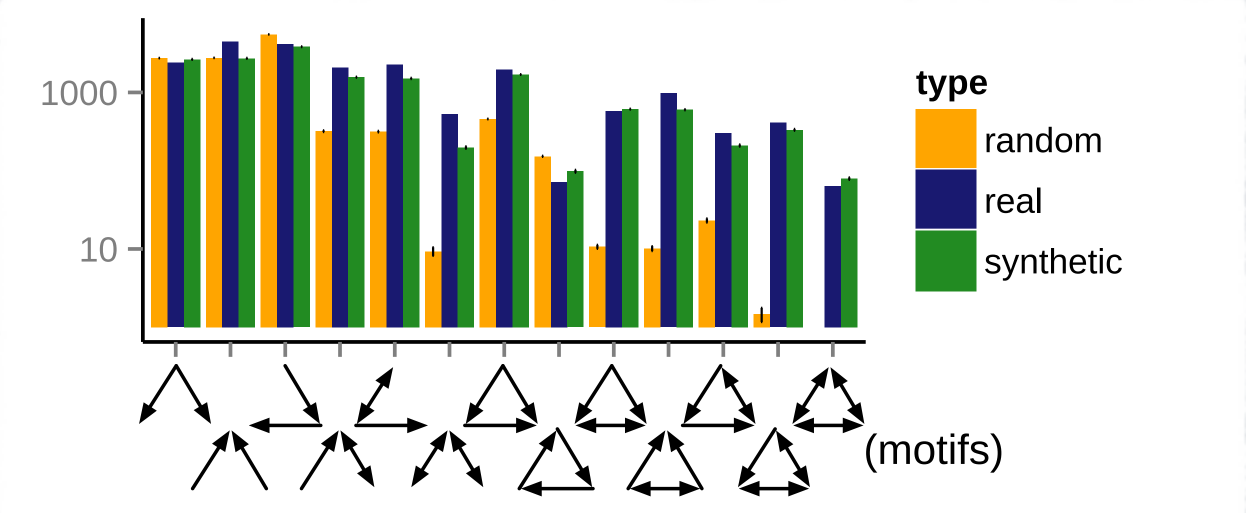
- in- and out-**degree distributions**
- directed and undirected **PageRank distributions**
- distance distributions
- triadic profiles (Milo et al., 2005)





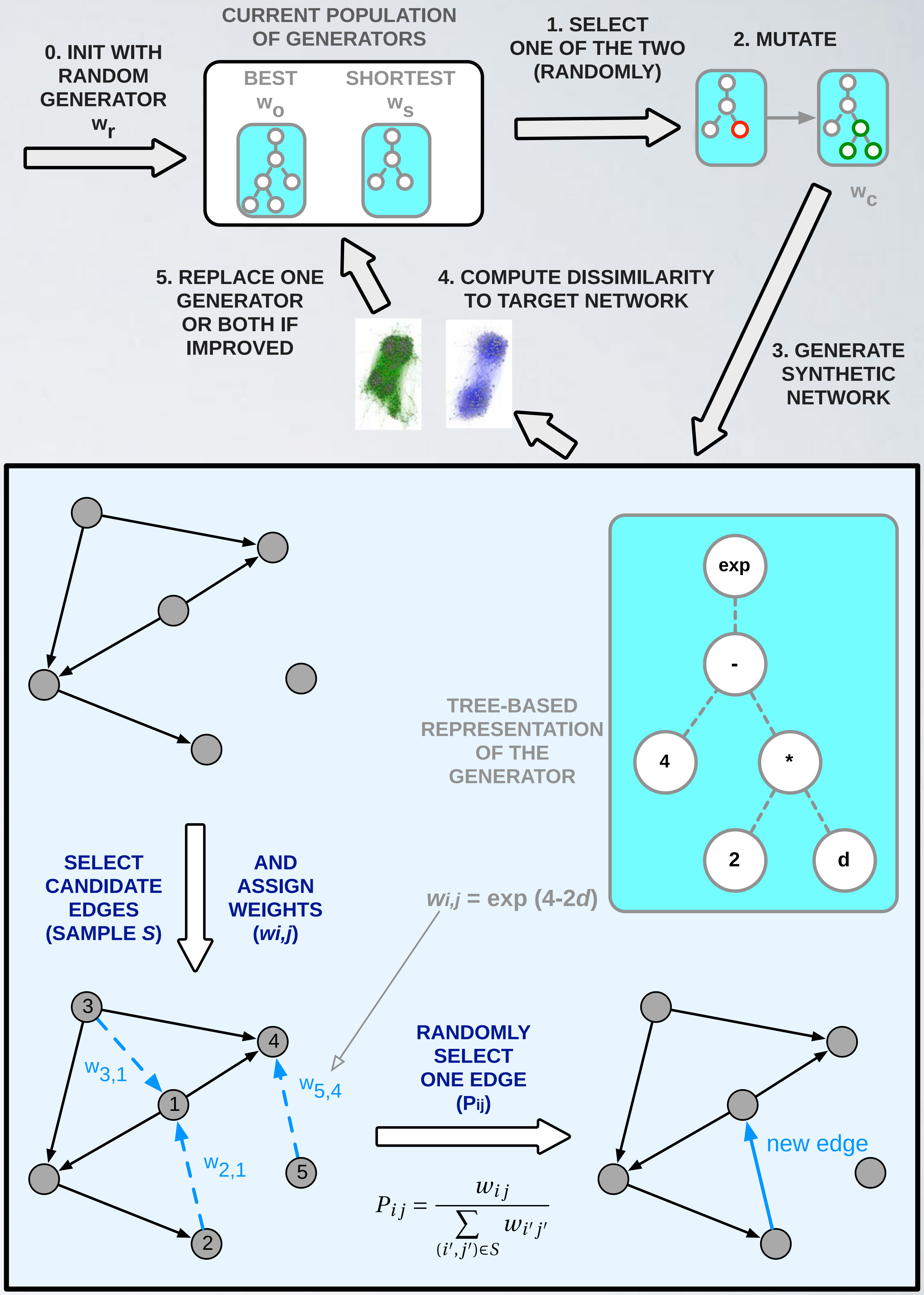
Grammar:

- Earth Mover's distance for distributions
- Improvement against random
- Worst improvement



EVOLUTIONARY PROCESS

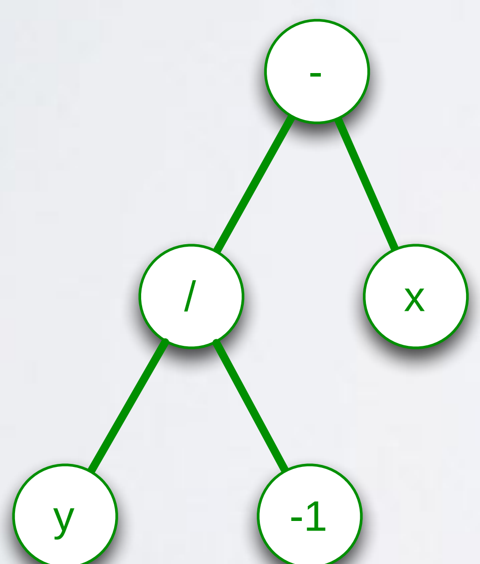
- Evolutionary algorithm iteratively improves generator

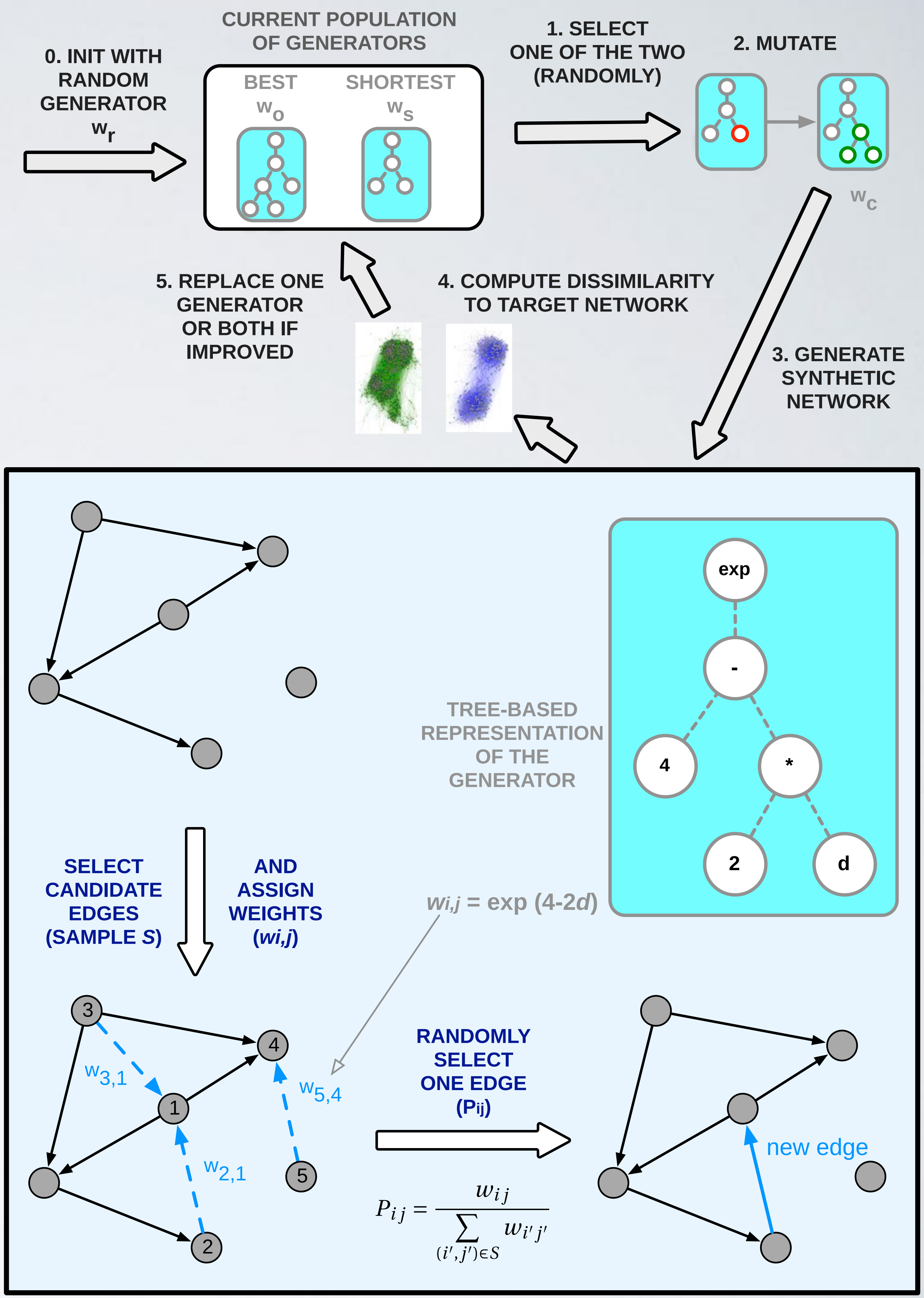


EVOLUTIONARY PROCESS

- Evolutionary algorithm iteratively improves generator

Random generator:

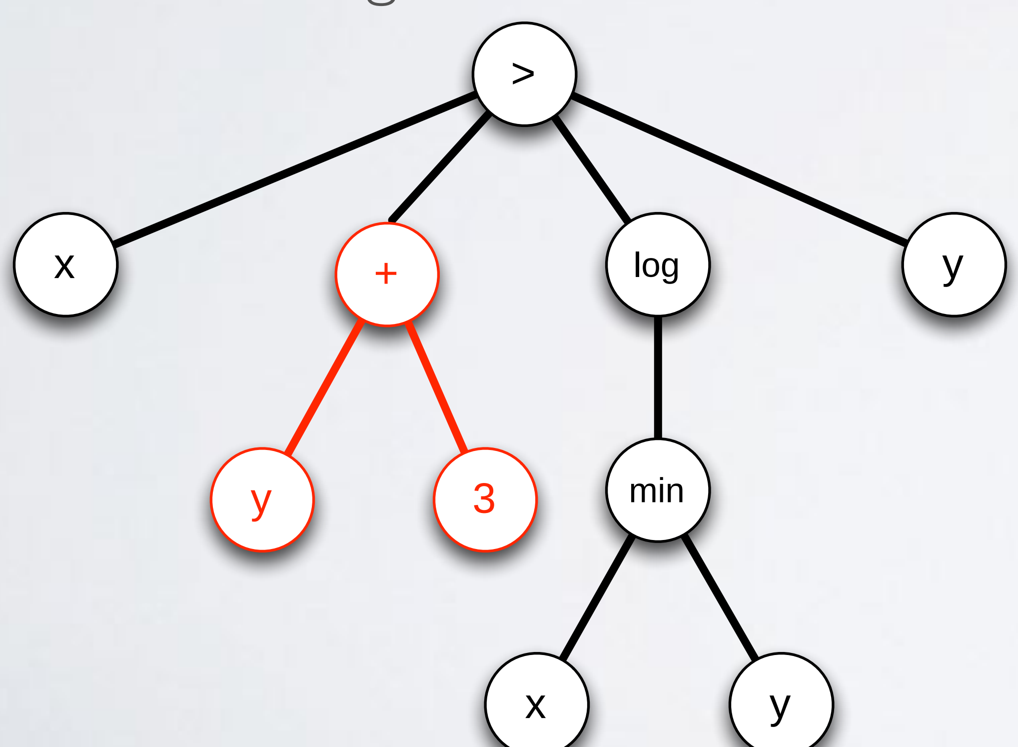


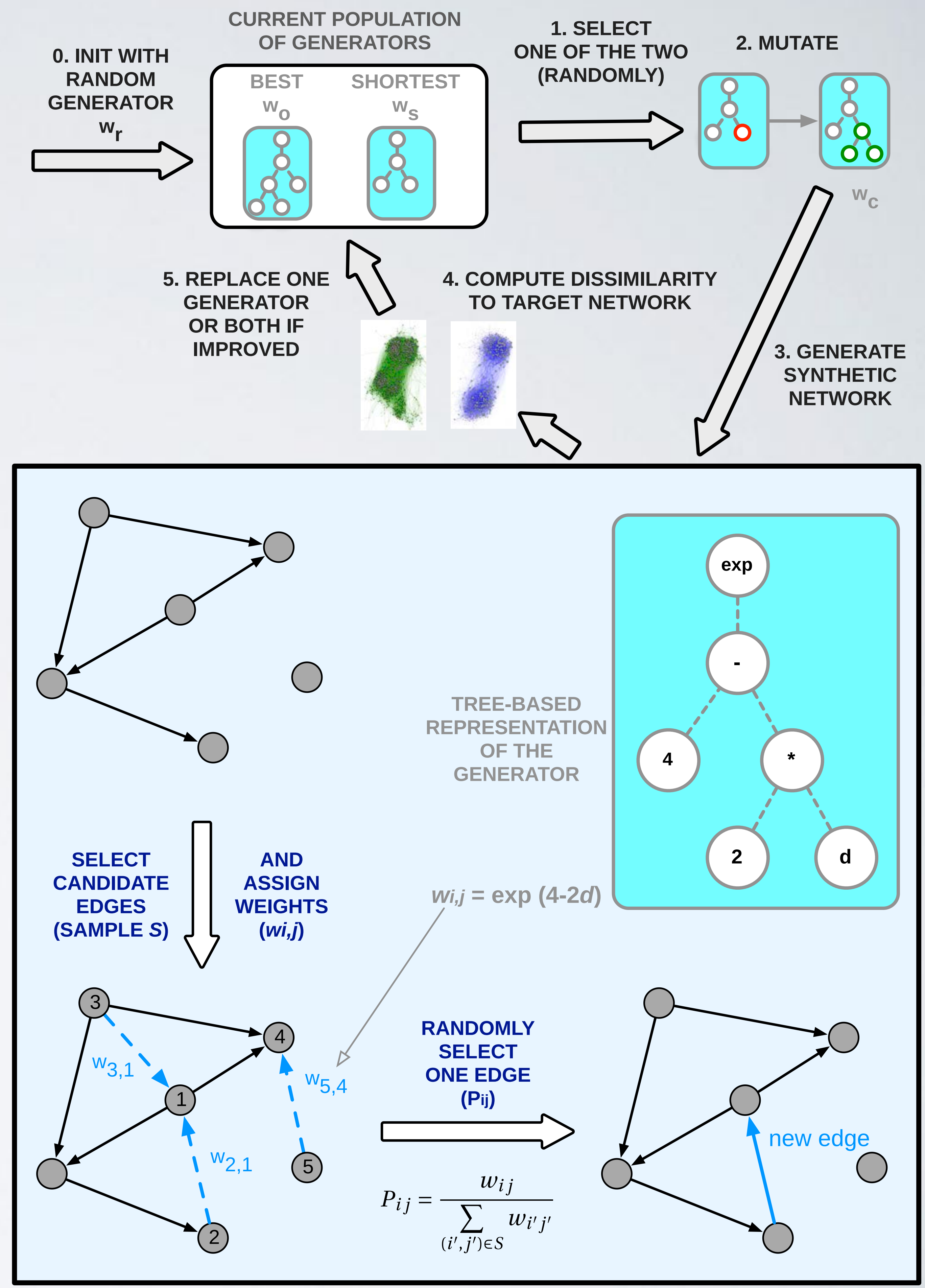


EVOLUTIONARY PROCESS

- Evolutionary algorithm iteratively improves generator

Program mutation:

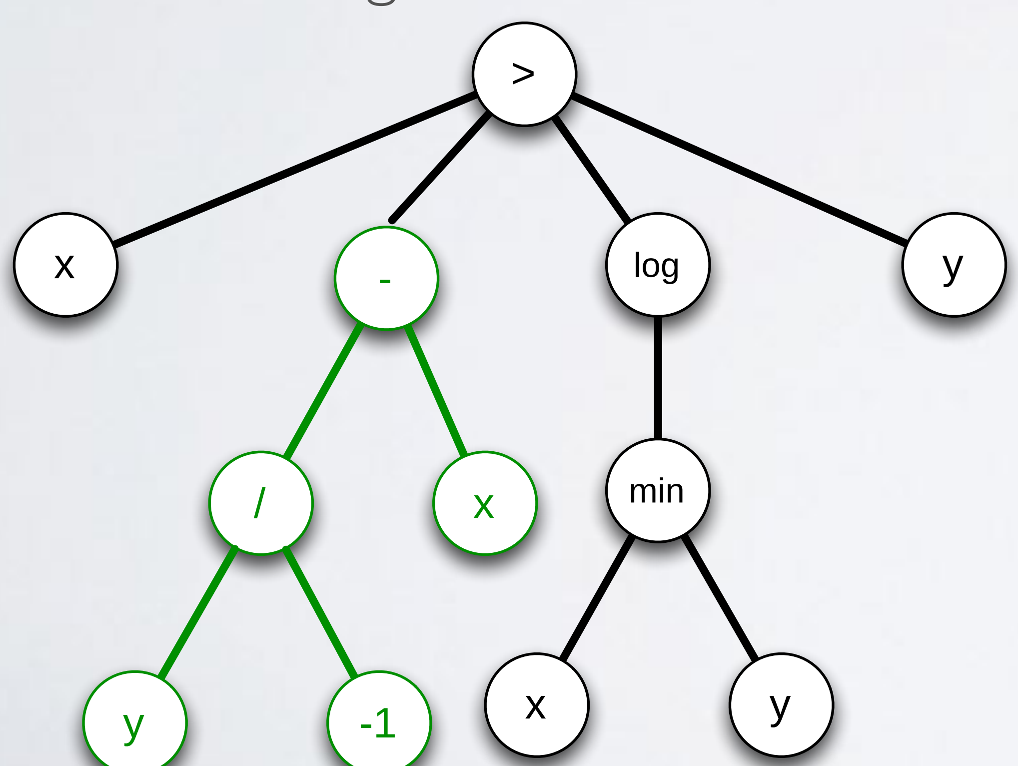


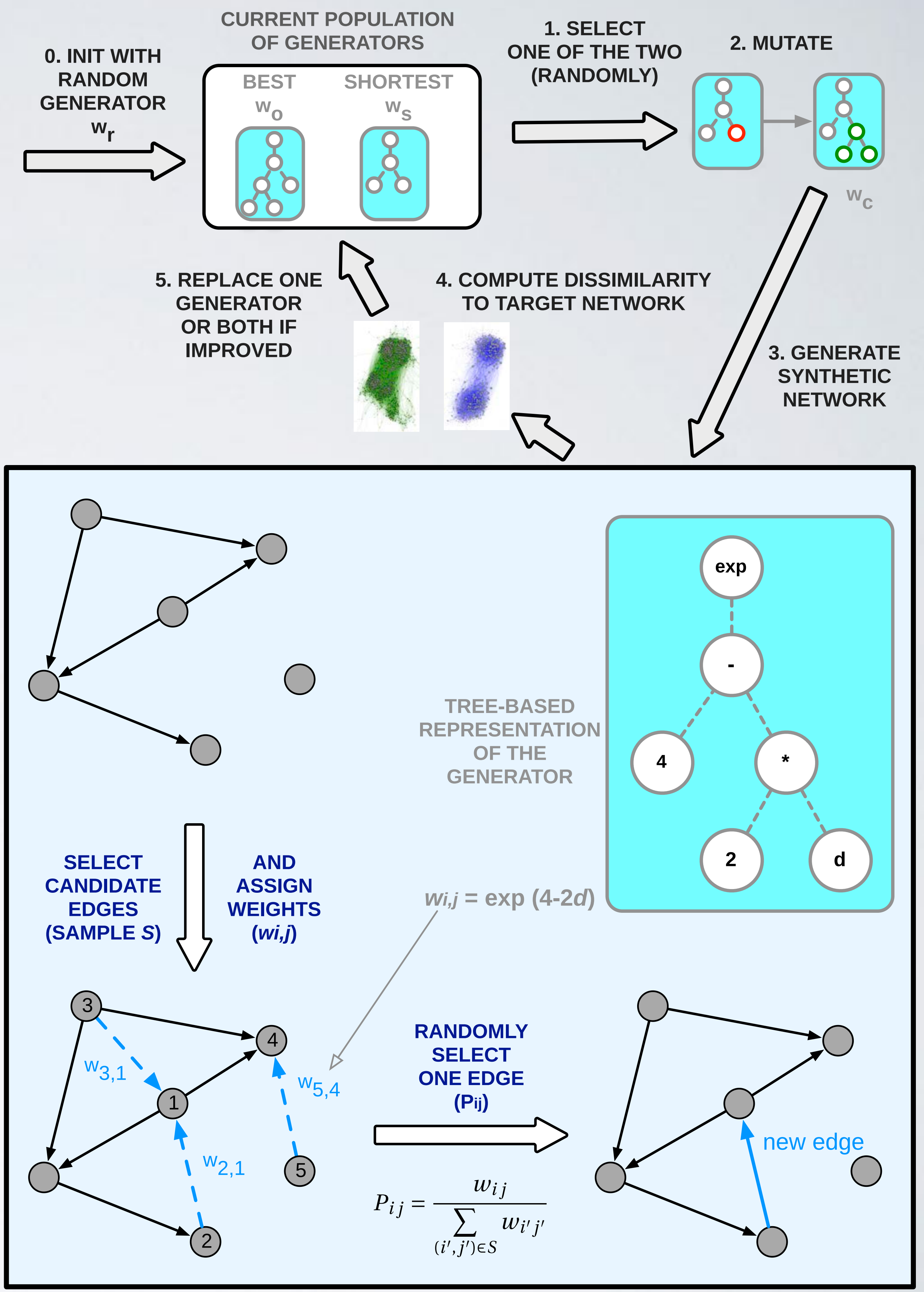


EVOLUTIONARY PROCESS

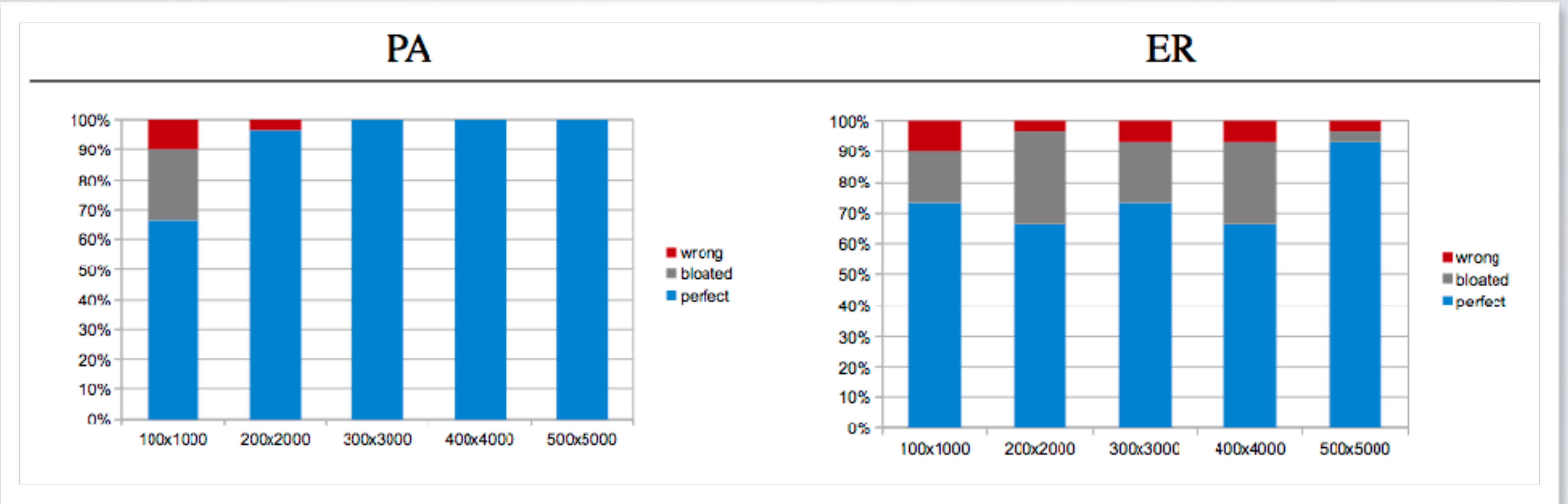
- Evolutionary algorithm iteratively improves generator

Program mutation:





ARTIFICIAL CASES



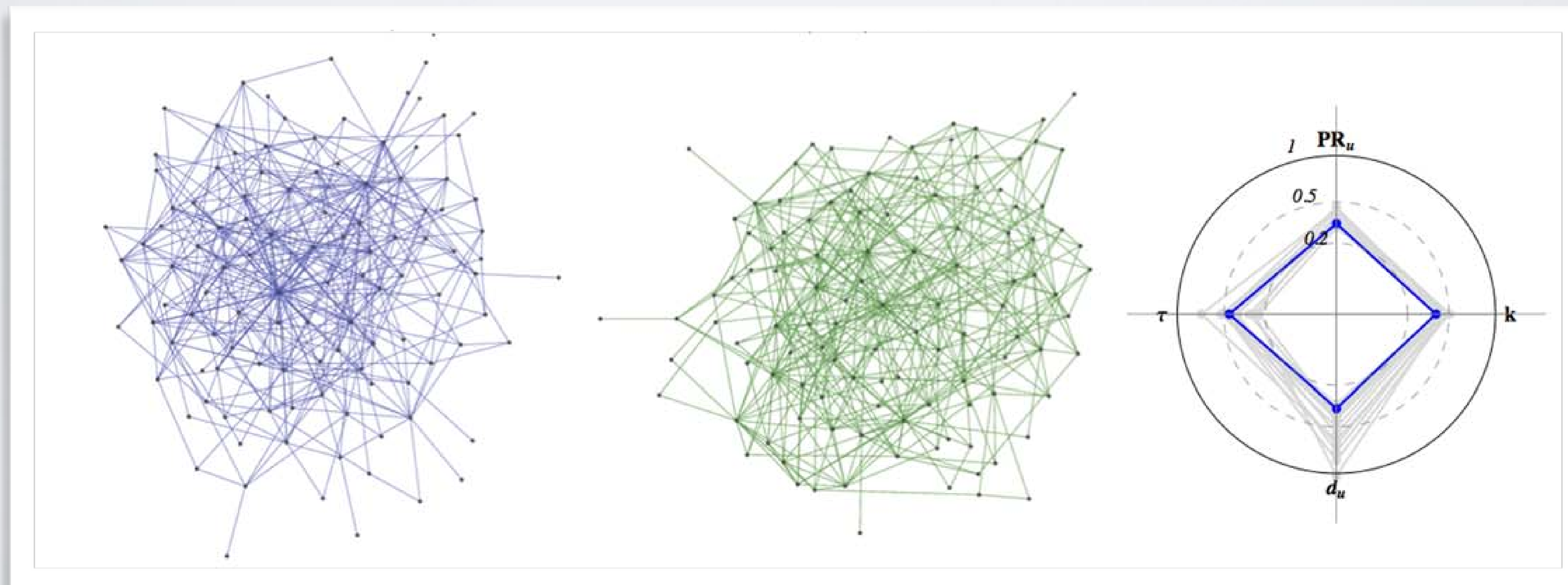
$$\mathrm{w}(i,j) = \mathrm{k}(j)$$

$$\mathrm{w}(i,j) = 1$$

Word adjacencies

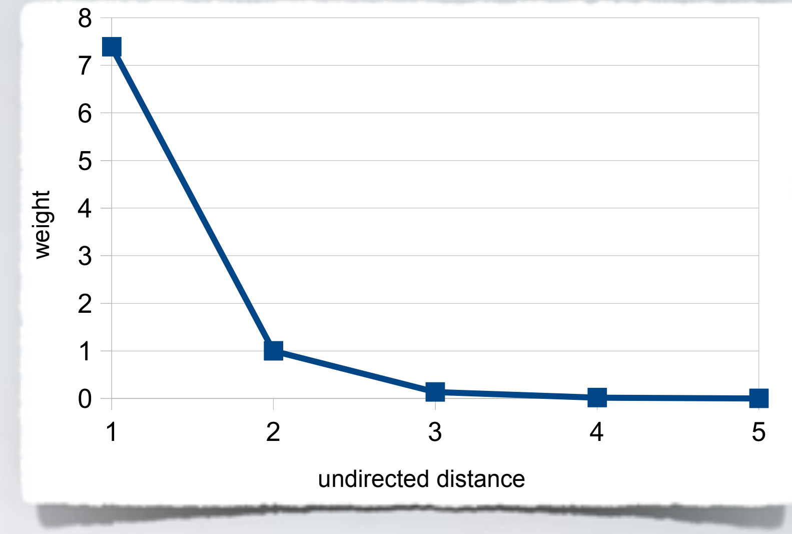
$$\boxed{\mathrm{w}(i, j) = \ k(i) - \mathrm{d}}$$

k, yet not too far



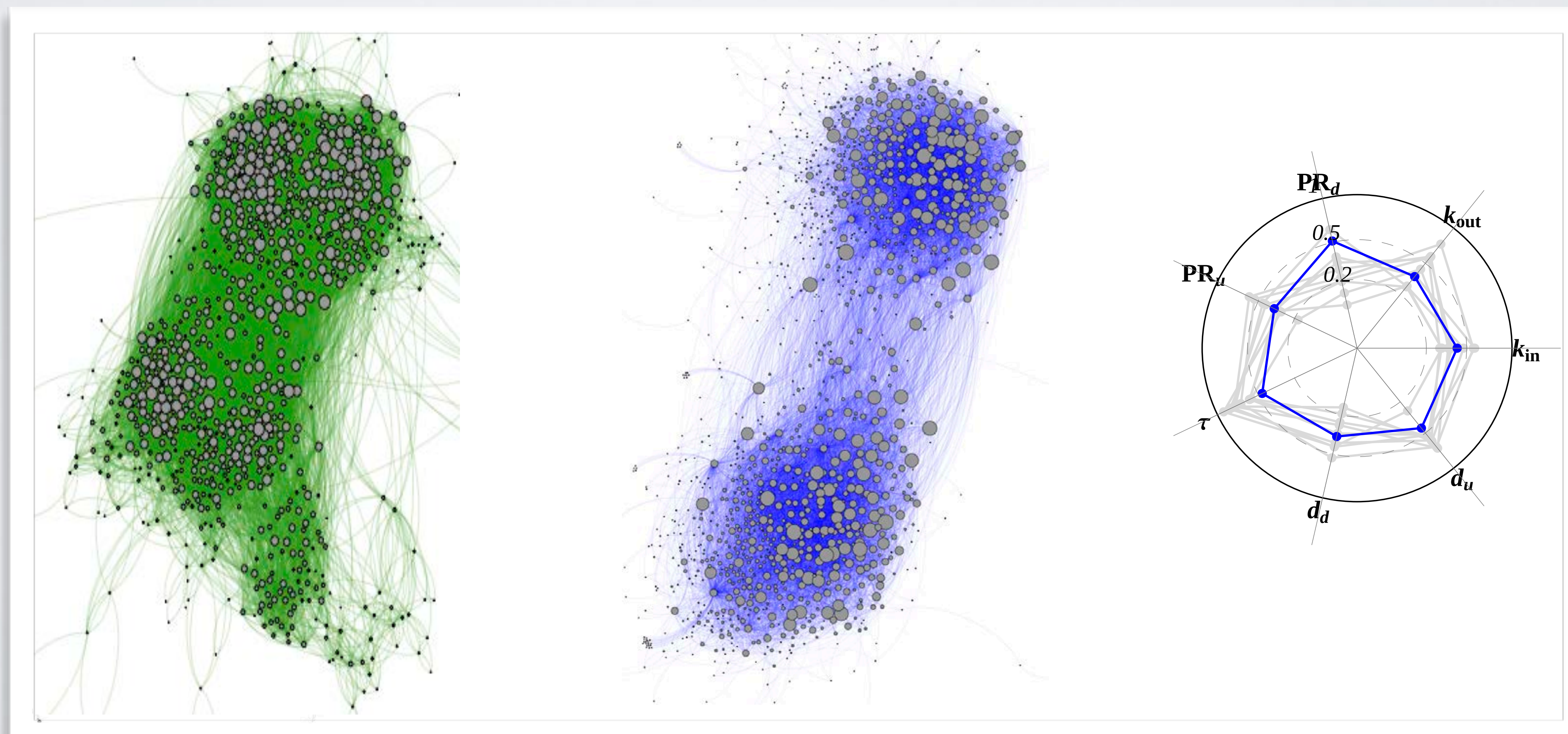
(dataset from Newman, 2006)

Political blogs



$$w(i, j) = \exp(4 - 2d)$$

close, reciprocal

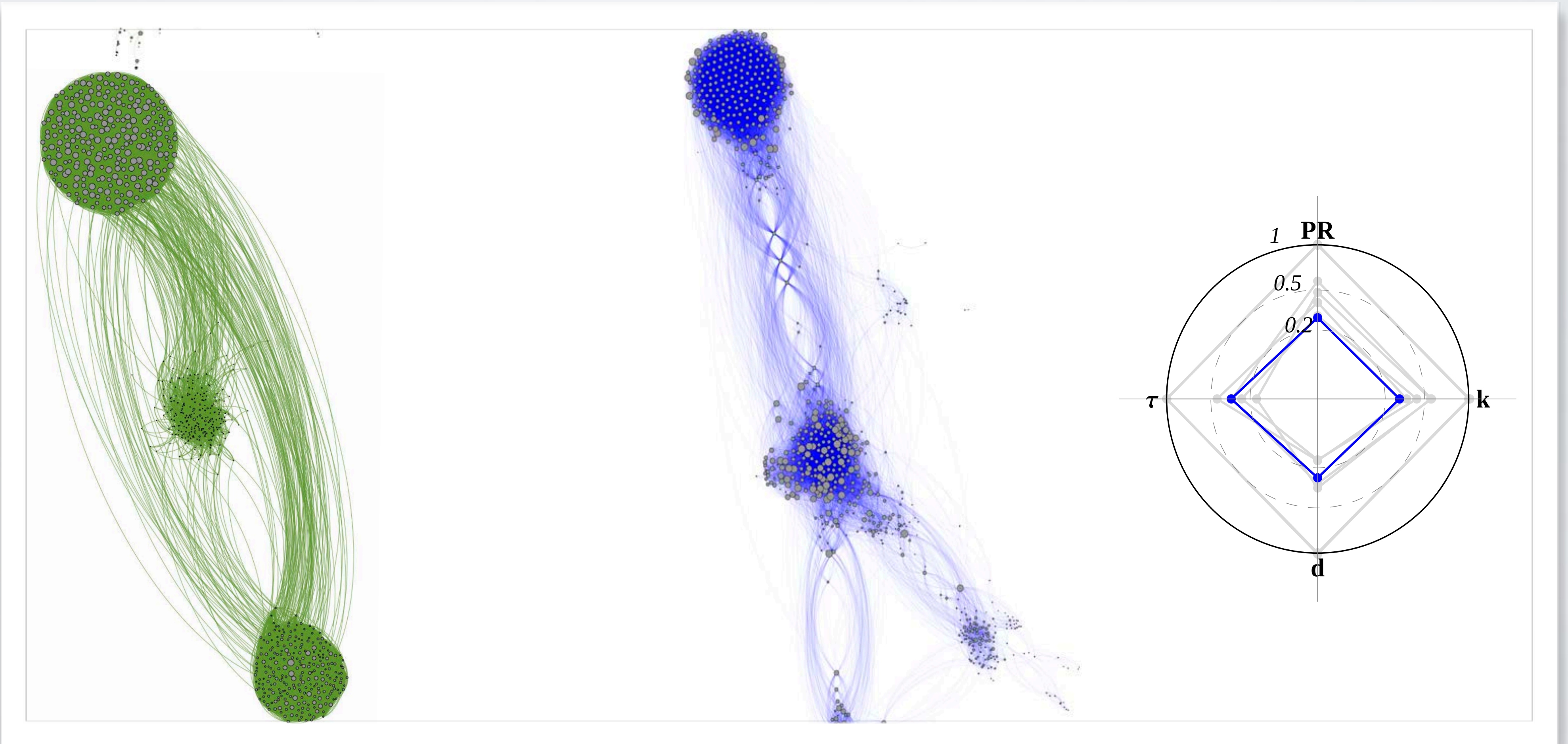


(dataset from Adamic & Glance, 2005)

Facebook

$$\boxed{\mathrm{w}(i, j) = \psi\,(3, i \cdot \mathrm{k}(i), \mathrm{k}(i))}$$

3 groups, local PA

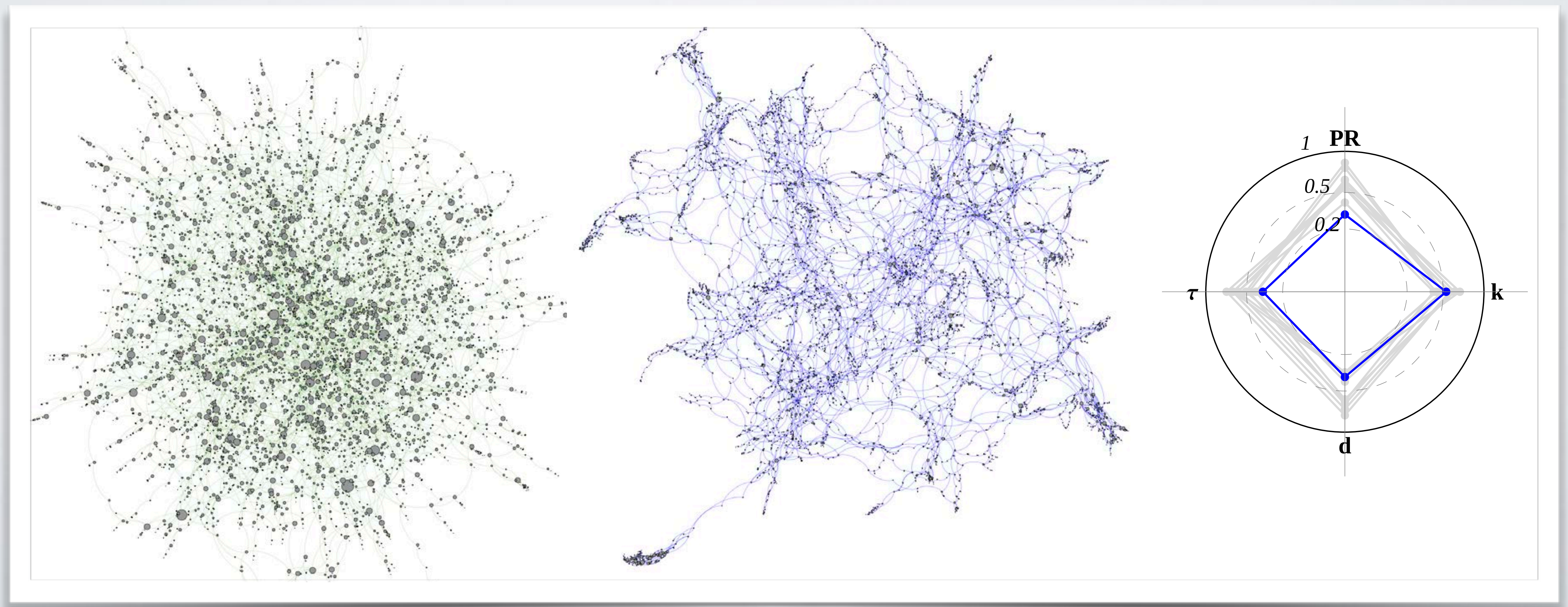


(dataset from Leskovec & Mc Auley, 2012)

Power grid

$$w(i,j) = \psi\left(d^{\,d \cdot \begin{cases} i-1, & \text{if } k(j)=0 \\ k(i), & \text{otherwise} \end{cases}}, 1, 0 \right)$$

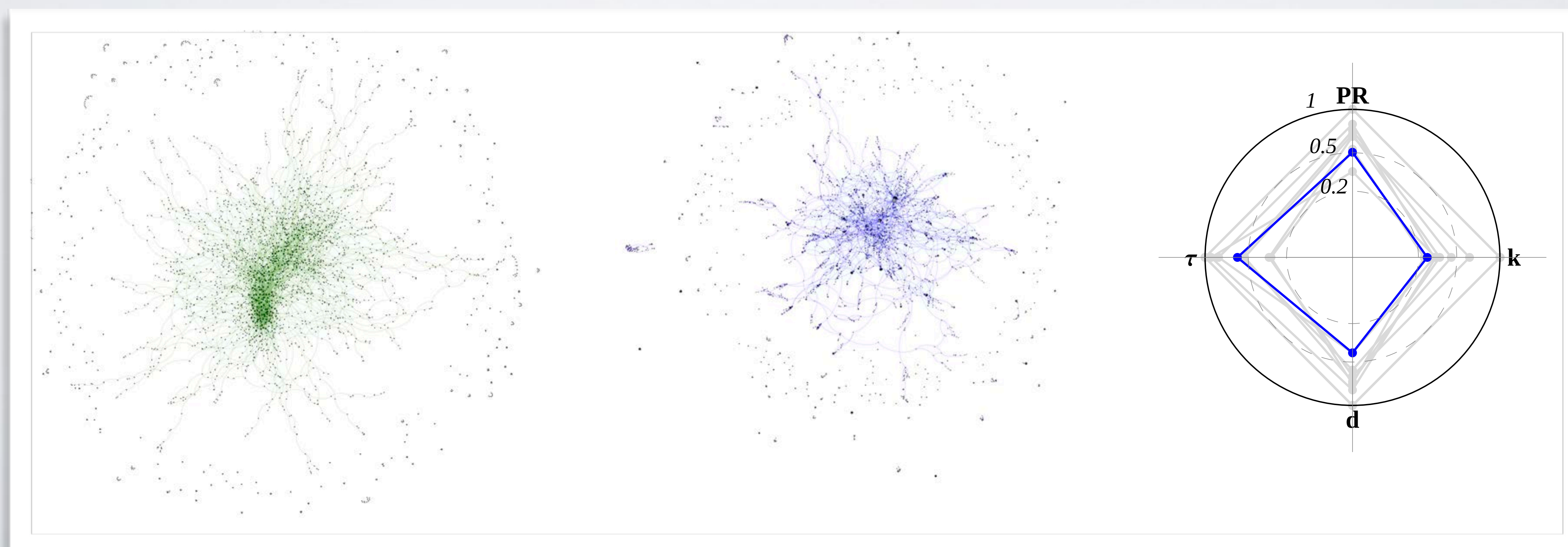
affinity exponentially
less likely with distance,
unless newcomer to central node
or origin node unconnected

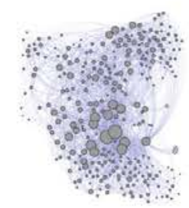
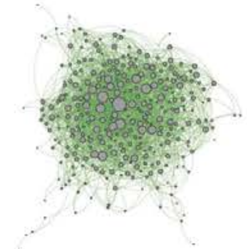
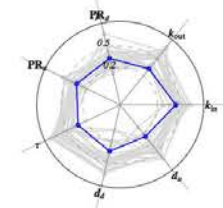
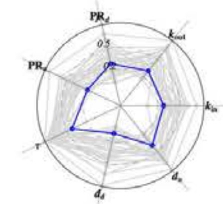
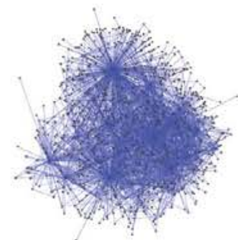
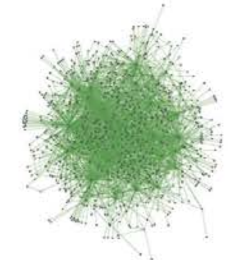
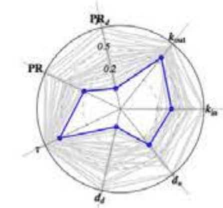
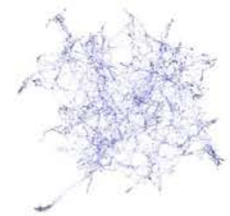
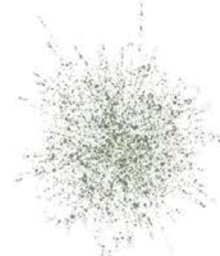
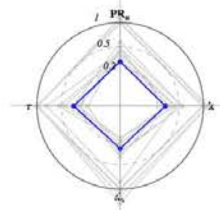
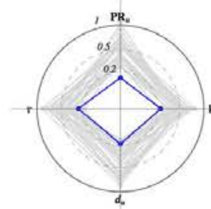
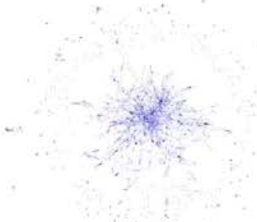
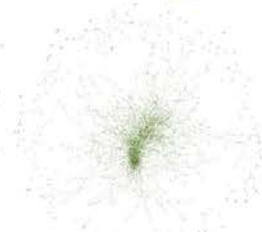
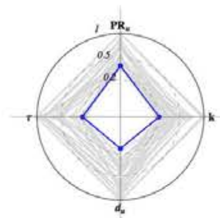
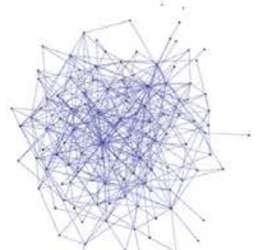
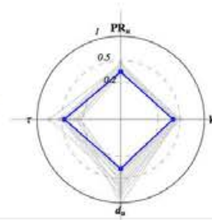


Protein interactions

$$
\begin{cases}
\log(i), & \text{if } \mathrm{k}(j) = \begin{cases} 0, & \text{if } \mathrm{k}(i) < 4 \\ \mathrm{k}(i), & \text{otherwise} \end{cases} \\
-1, & \text{otherwise}
\end{cases}
$$

a priori logarithmic distribution of affinity, for some pairs



Network	*Program* $w(i,j)$	*Real*	*Synthetic*	*Metrics*
C. elegans	$\log(d_D)^{d_D-7} - \min(j+0.52, 0.77)$			
Political blogs	$\exp(4-2d)$			
Software collaborations	$\psi\left(\frac{k'(i)}{d}, k'(j), 0.7k(j)\right)\cdot\min\left(k(j), 9\right)$			
Power grid	$\psi\left(d^{d\cdot\begin{cases} i-1, & \text{if } k(j)=0 \\ k(i), & \text{otherwise}\end{cases}}, 1, 0\right)$			
Facebook	$\psi(3, i\cdot k(i), k(i))$			
Proteins	$\begin{cases}\log(i), & \text{if } k(j) = \begin{cases} 0, & \text{if } k(i) < 4 \\ k(i), & \text{otherwise}\end{cases} \\ -1, & \text{otherwise}\end{cases}$			
Word adjacencies	$k(i) - d$			

GENERATOR SIMILARITIES

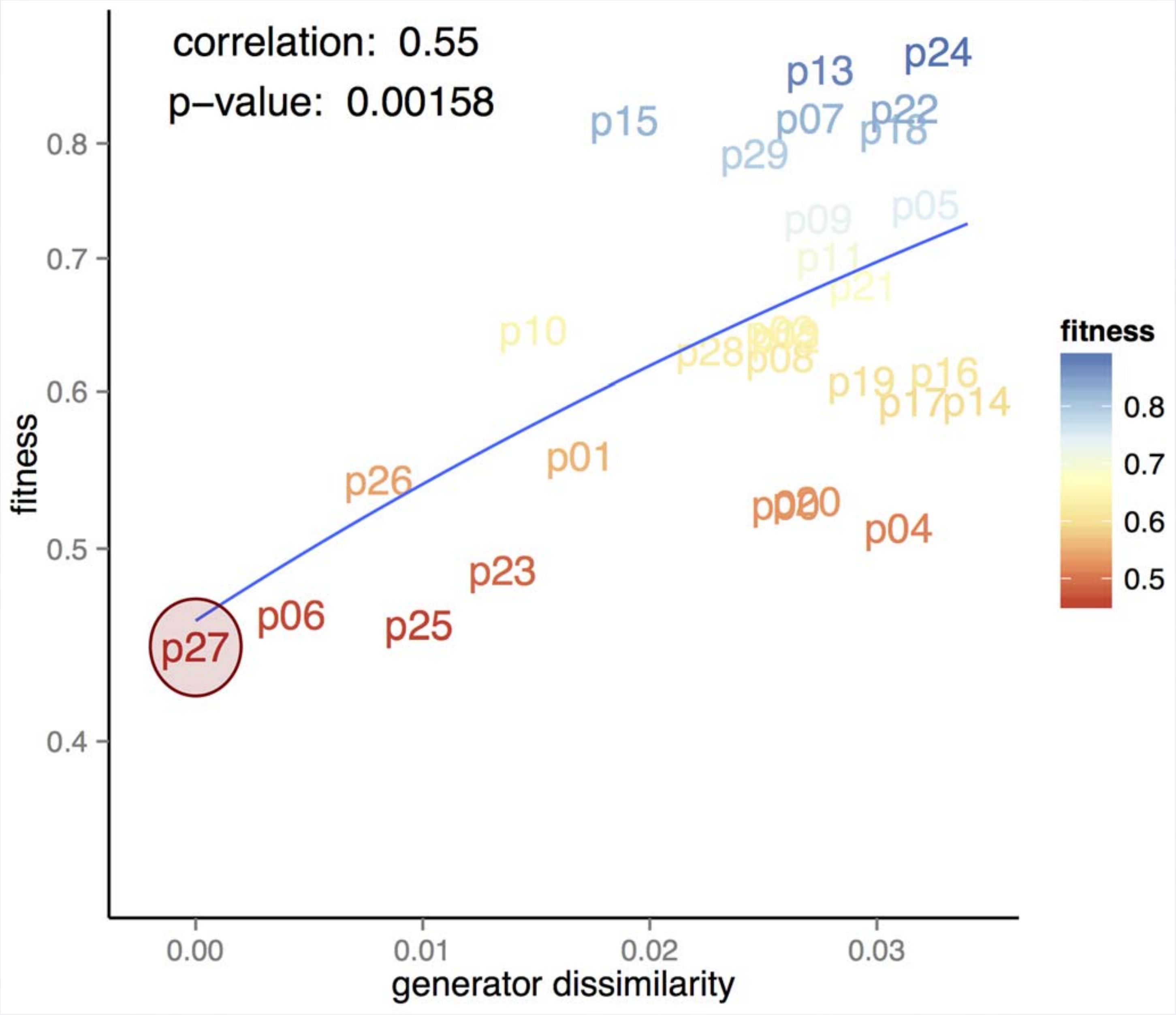


Figure 4 | Similarity of generators. Comparison between generator similarity to the optimal generator (p27) and fitness.

TAKE-HOME MESSAGE

- Propose an artificial scientist to **guide hypothesis search**
- **Decipher the genotype of networks** from their phenotype

TAKE-HOME PAPER

SCIENTIFIC REPORTS

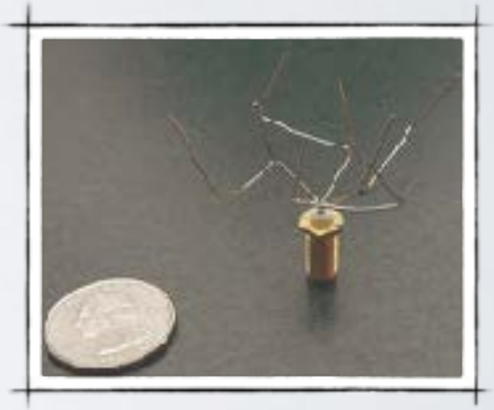
OPEN

Symbolic regression of generative network models

SUBJECT AREAS:
SCIENTIFIC DATA
MACHINE LEARNING
SOFTWARE
APPLIED MATHEMATICS

Telmo Menezes[1,2] & Camille Roth[1]

[1]Centre Marc Bloch Berlin (An-Institut der Humboldt Universität, UMIFRE CNRS-MAE) Friedrichstr. 191, 10117 Berlin, Germany, [2]Centre d'Analyse et de Mathématique Sociales (UMR 8557 CNRS-EHESS) 190 av. de France, 75013 Paris, France.



Networks are a powerful abstraction with applicability to a variety of scientific fields. Models explaining their morphology and growth processes permit a wide range of phenomena to be more systematically analysed and understood. At the same time, creating such models is often challenging and requires insights that may be counter-intuitive. Yet there currently exists no general method to arrive at better models. We have developed an approach to automatically detect realistic decentralised network growth models from empirical data, employing a machine learning technique inspired by natural selection and defining a unified formalism to describe such models as computer programs. As the proposed method is completely general and does not assume any pre-existing models, it can be applied "out of the box" to any given network. To validate our approach empirically, we systematically rediscover pre-defined growth laws underlying several canonical network generation models and credible laws for diverse real-world networks. We were able to find programs that are simple enough to lead to an actual understanding of the mechanisms proposed, namely for a simple brain and a social network.

TAKE-HOME MESSAGE

- Propose an artificial scientist to **guide hypothesis search**
- **Decipher the genotype of networks** from their phenotype

TAKE-HOME SOFTWARE

- **Synthetic** open-source tool
- https://github.com/telmomenezes/synthetic

TAKE-HOME PAPER

SCIENTIFIC REPORTS

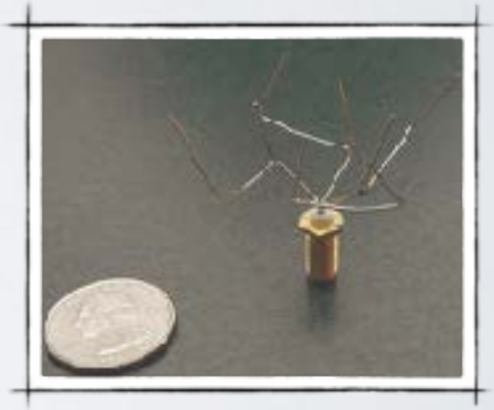
OPEN

SUBJECT AREAS:
SCIENTIFIC DATA
MACHINE LEARNING
SOFTWARE
APPLIED MATHEMATICS

Symbolic regression of generative network models

Telmo Menezes[1,2] & Camille Roth[1]

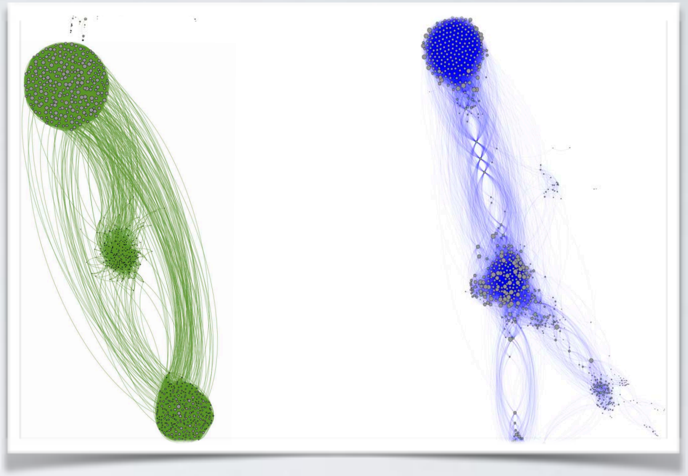
[1]Centre Marc Bloch Berlin (An-Institut der Humboldt Universität, UMIFRE CNRS-MAE) Friedrichstr. 191, 10117 Berlin, Germany,
[2]Centre d'Analyse et de Mathématique Sociales (UMR 8557 CNRS-EHESS) 190 av. de France, 75013 Paris, France.



Networks are a powerful abstraction with applicability to a variety of scientific fields. Models explaining their morphology and growth processes permit a wide range of phenomena to be more systematically analysed and understood. At the same time, creating such models is often challenging and requires insights that may be counter-intuitive. Yet there currently exists no general method to arrive at better models. We have developed an approach to automatically detect realistic decentralised network growth models from empirical data, employing a machine learning technique inspired by natural selection and defining a unified formalism to describe such models as computer programs. As the proposed method is completely general and does not assume any pre-existing models, it can be applied "out of the box" to any given network. To validate our approach empirically, we systematically rediscover pre-defined growth laws underlying several canonical network generation models and credible laws for diverse real-world networks. We were able to find programs that are simple enough to lead to an actual understanding of the mechanisms proposed, namely for a simple brain and a social network.

GENOTYPE FAMILIES

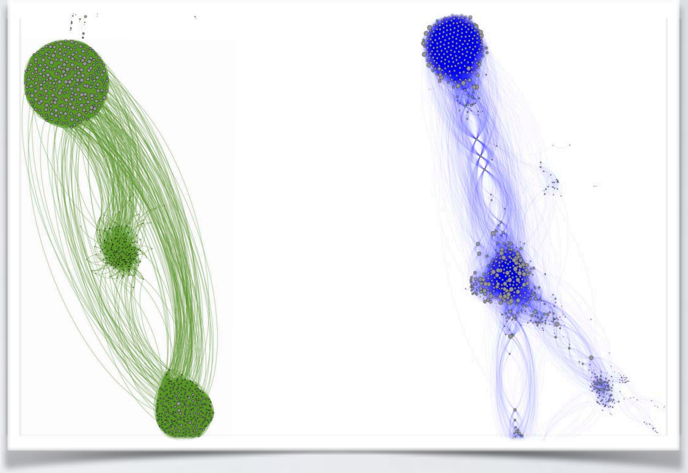
using 238 anonymized Facebook ego-centered friendship networks

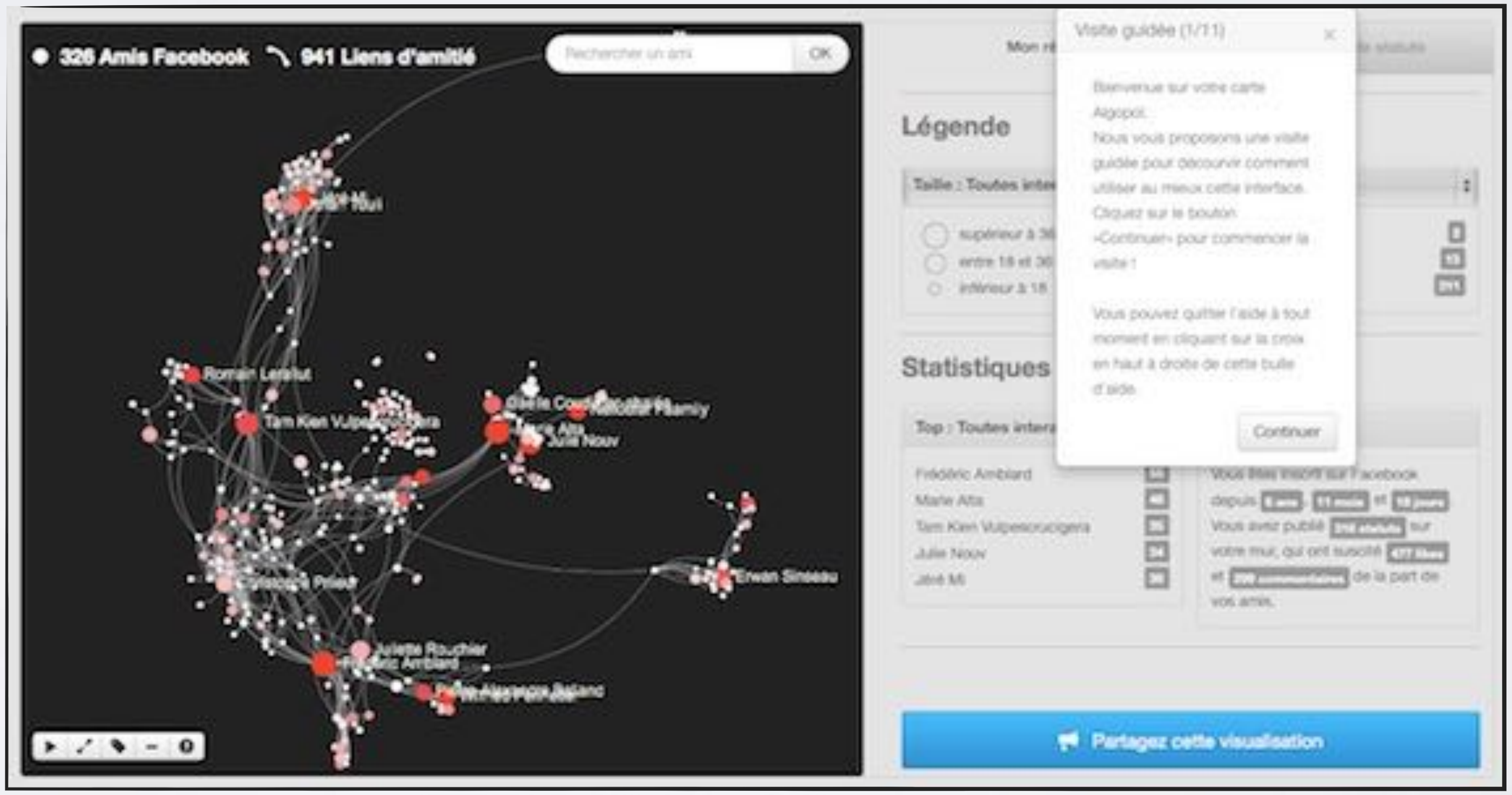


"Algopol" application

GENOTYPE FAMILIES

using 238 anonymized
Facebook ego-centered
friendship networks

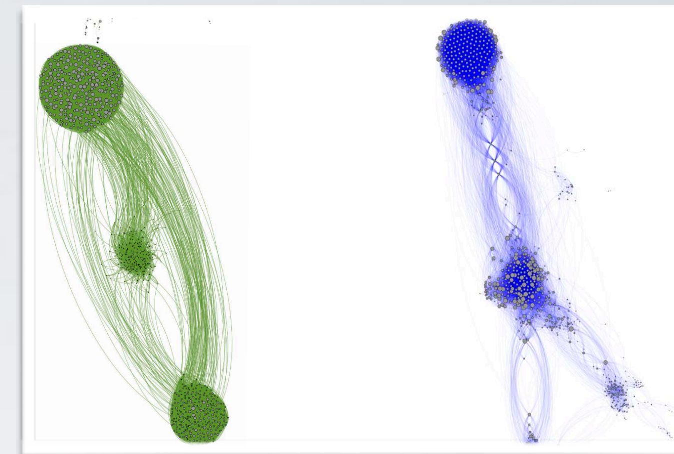


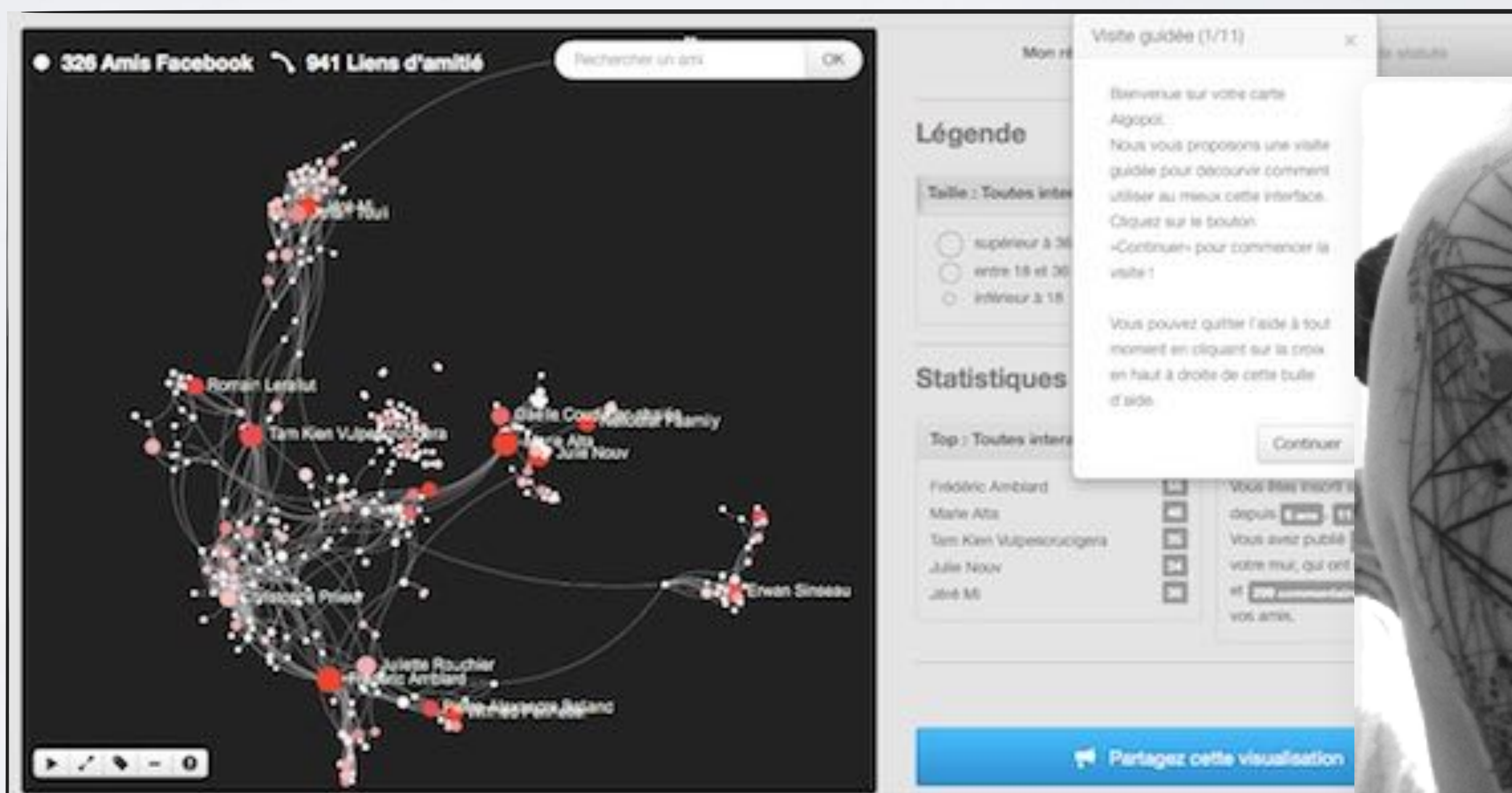


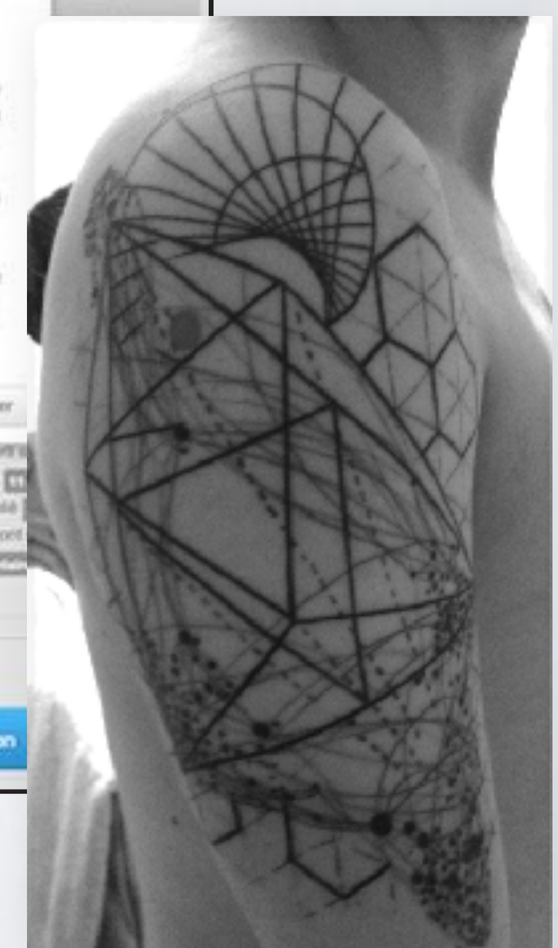
"Algopol" application

GENOTYPE FAMILIES

using 238 anonymized
Facebook ego-centered
friendship networks

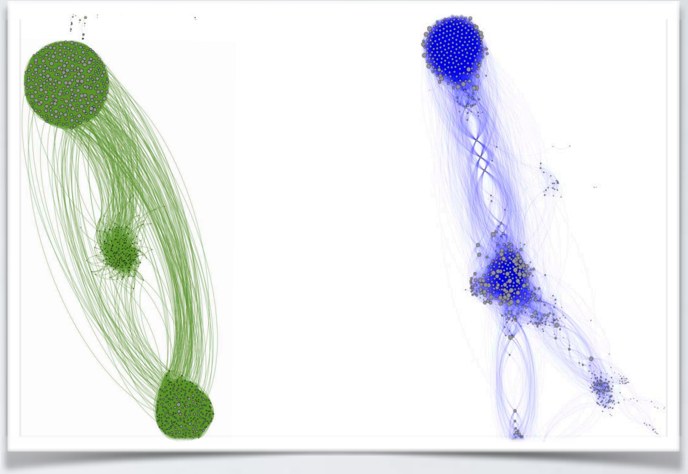






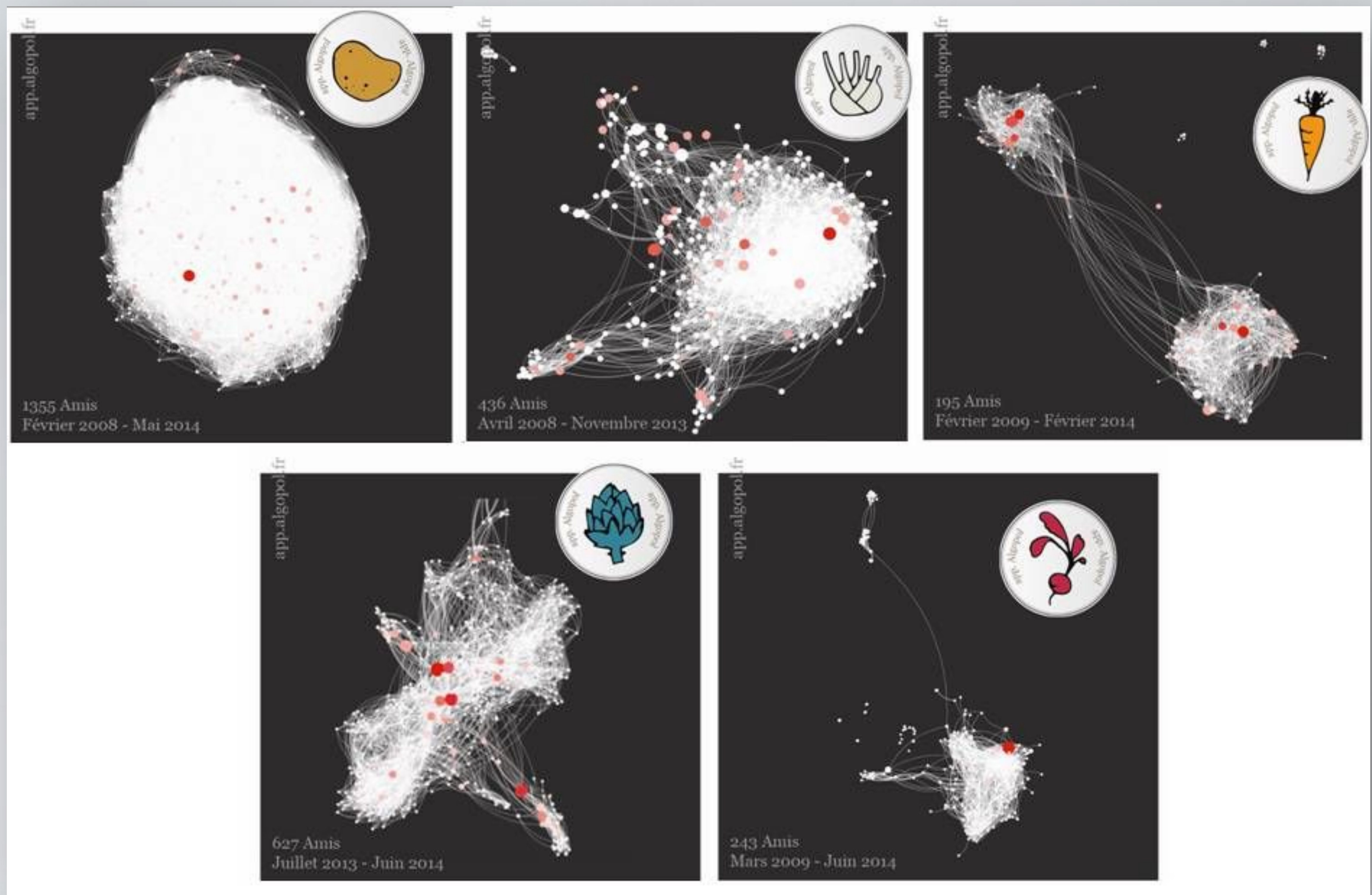
"Algopol" application

GENOTYPE FAMILIES



using 238 anonymized
Facebook ego-centered
friendship networks

"Algopol" application



(Algopol, 2012-16)

Family	*List of generator functions and corresponding network number* ⟨ID⟩					
ER c	0.08 ⟨14⟩	0.88 ⟨50⟩	0.95 ⟨78⟩	54.6 ⟨82⟩	0.62 ⟨108⟩	6.0 ⟨124⟩
	$(\max(k_i,i)=0\rightarrow 0,0.63)$ ⟨198⟩					
ID i	i ⟨58⟩	i ⟨109⟩				
ID' e^i	e^i ⟨18⟩	e^i ⟨139⟩				
PA k	k ⟨26⟩	k ⟨81⟩	k ⟨100⟩	k ⟨105⟩	k ⟨111⟩	k ⟨134⟩
	k ⟨145⟩	k ⟨170⟩	k ⟨227⟩			
PA' ${k_i}^{k_j}$	${k_j}^{k_i}$ ⟨0⟩	$(\min(j,.66)>k_i\rightarrow j,e^{k_j})^{(\min((j=0,k_j,k_i),e^{k_j}))}$ ⟨47⟩				${k_i}^{k_j}$ ⟨193⟩
SC-α $\psi_g(k^s,c)$	$\psi_8(k_j^2,.62)-k_i$ ⟨69⟩		$\psi_7(k^3,4)$ ⟨126⟩			
SC-β $\psi_g(e^k,>\frac{1}{2})$	$\psi_3(2^k,.48)$ ⟨3⟩	$\psi_9(e^{k_i},.49)$ ⟨36⟩	$\psi_4(e^k,1.1)$ ⟨39⟩	$\psi_5(\frac{e^{\max(k_i,k_j)}}{k_i},k_i)$ ⟨80⟩		$\psi_5(e^k,1)$ ⟨90⟩
	$\psi_4(e^k,1)$ ⟨110⟩	$\psi_8(e^k,d)$ ⟨138⟩	$\psi_4(k_i,.67)^{k_i}$ ⟨153⟩	$\psi_5(e^k,1.7)$ ⟨213⟩	$\psi_3(e^k,2)$ ⟨224⟩	
SC-γ $\psi_g(k^B,\sim 0)$	$\psi_9(k^k,0)$ ⟨23⟩	$\psi_6(3^k,0)$ ⟨31⟩	$\psi_4(4\cdot k^5,0)$ ⟨41⟩	$\psi_8(k^k,0)$ ⟨57⟩	$\psi_3(e^{k_i+k_j},.05)$ ⟨97⟩	
	$\psi_3(e^k,0)$ ⟨104⟩	$\psi_3(2^k,0)$ ⟨127⟩	$\psi_6(e^{\psi_5(1,k)},0)+.07$ ⟨141⟩		$\psi_7(e^k,0)$ ⟨155⟩	$\psi_4(e^k,.06)$ ⟨157⟩
	$\psi_2(k_i\cdot e^{k_j},0)$ ⟨164⟩	$\psi_4(e^k,0)$ ⟨177⟩	$\psi_5(k^7,.01)$ ⟨235⟩	$\psi_5(e^k,.03)$ ⟨236⟩		
SC-δ $\psi_g(e^i,*)$	$\psi_4(e^i,e^{k_i})$ ⟨6⟩	$\psi_4(i^j,k_j)$ ⟨89⟩	$\psi_2(j^i,k_i)$ ⟨92⟩	$\psi_3(e^i,k_i)$ ⟨121⟩	$\psi_3(e^i,e^7)$ ⟨137⟩	$\psi_3(e^i,1)$ ⟨148⟩
	$\psi_2(9^i,9^9)$ ⟨181⟩	$\psi_3(e^i,j)$ ⟨184⟩	$\psi_3(e^{i+j-d},e^5)$ ⟨196⟩		$\psi_4(9^i,9)$ ⟨202⟩	
SC-ε $\psi_g(ik,*)$	$9\psi_3(ik_i,2k_i)$ ⟨9⟩	$\psi_4(ik_j,6k_j)$ ⟨24⟩	$\psi_5(jk_j,k_j)$ ⟨25⟩	$\psi_9(ik_i,.1k_i)$ ⟨37⟩	$\psi_2(jk_j,k_j)$ ⟨75⟩	$\psi_7(jk_j,7k_j)$ ⟨91⟩
	$\psi_6(ik_i,.44k_i)$ ⟨106⟩	$\psi_4(jk_i,.38)$ ⟨107⟩	$\psi_3(jk_i,k_j)$ ⟨115⟩	$\psi_4(i\log(k_i),0)$ ⟨165⟩		$\psi_3(jk_i,\frac{k_i}{4})$ ⟨166⟩
	$(\frac{k_jk_i}{.66}+d)\psi_4(j,.61)$ ⟨188⟩		$\psi_3(ik_j,2k_j)$ ⟨194⟩	$\psi_3(ik_j,k_j)$ ⟨206⟩	$\psi_3(ik_i,0)$ ⟨209⟩	$\psi_4(ik_i,3k_i)$ ⟨218⟩
SC-ζ $\psi_g(i^k,*)$	$\psi_7(i,0)^{k_j}$ ⟨68⟩	$\frac{7}{d}\psi_4(i^{k_i},.48)$ ⟨93⟩	$\psi_4(\frac{i^{k_j}}{k_j},.18)$ ⟨95⟩	$\psi_8(i^{k_i},2)$ ⟨125⟩	$\psi_4(i^{k_i},0)$ ⟨156⟩	$\psi_4(\frac{1}{6}i^{k_i},d)$ ⟨179⟩
	$\psi_9(dj^{k_i},0)$ ⟨185⟩	$\psi_{\min(i,4)}(i^{k_i},0)$ ⟨195⟩		$\psi_5(9j^{k_i},.03)$ ⟨219⟩		
SC-η $\psi_g(ik^2,*)$	$\psi_5((ik_i)^2,i)$ ⟨16⟩	$\psi_5(ik_i{}^2,6)$ ⟨128⟩	$\psi_4(2980.96k^2,2k)$ ⟨132⟩		$\psi_2(ik_j{}^2,k_j{}^2)$ ⟨163⟩	
	$\psi_7(\psi_i(.5,k_j{}^2),0)$ ⟨182⟩					
SC-θ $\psi_g(k,0)-1$	$\psi_4(k,0)-.99$ ⟨8⟩		$\psi_7(k,0)-.93$ ⟨83⟩			

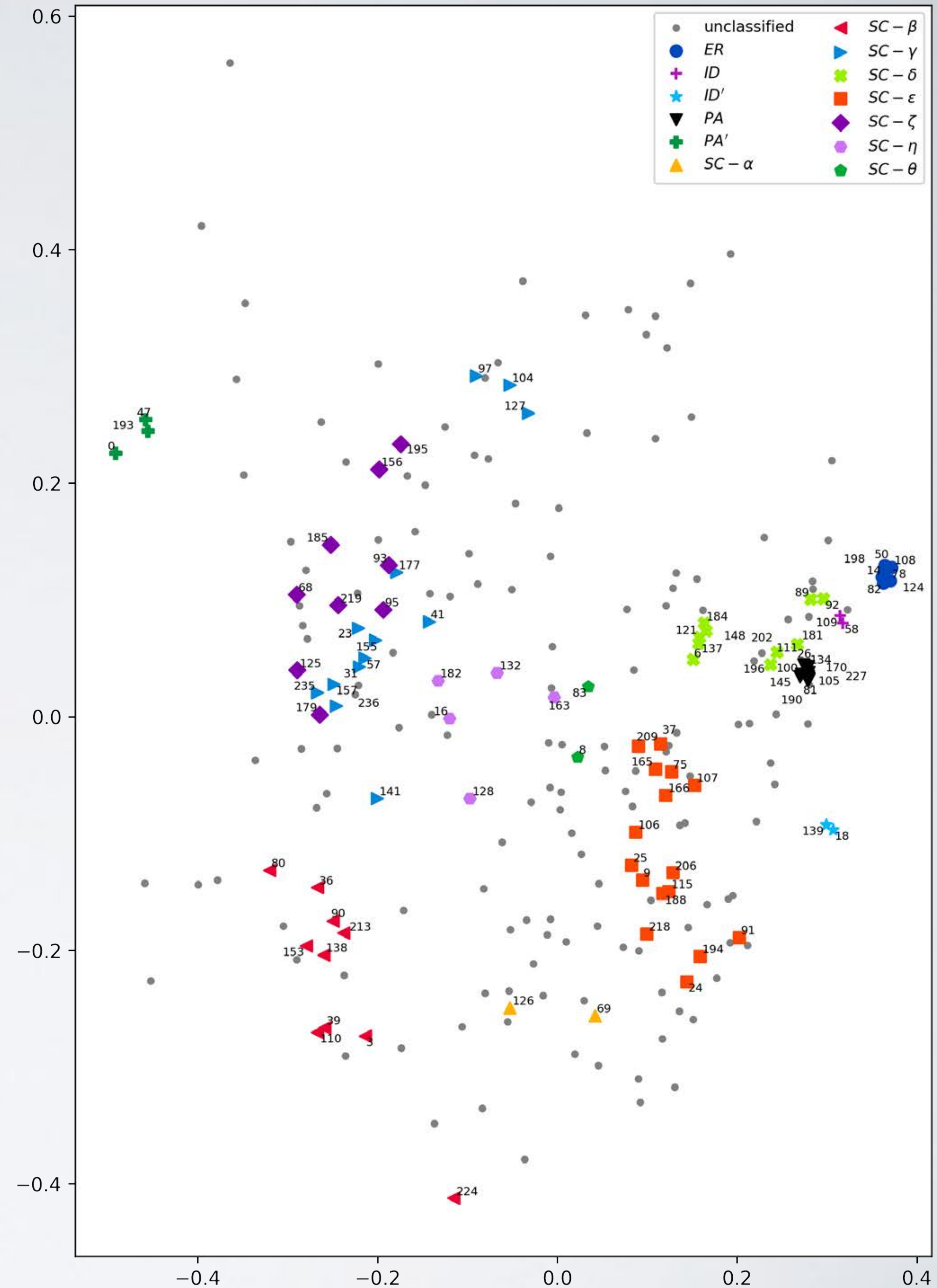


Fig. 3 Network generators mapped into a two-dimensional layout according to their pairwise distances. Different colors and shapes indicate families of generators that were manually identified as semantically similar. The legend shows the pattern that identifies each family.

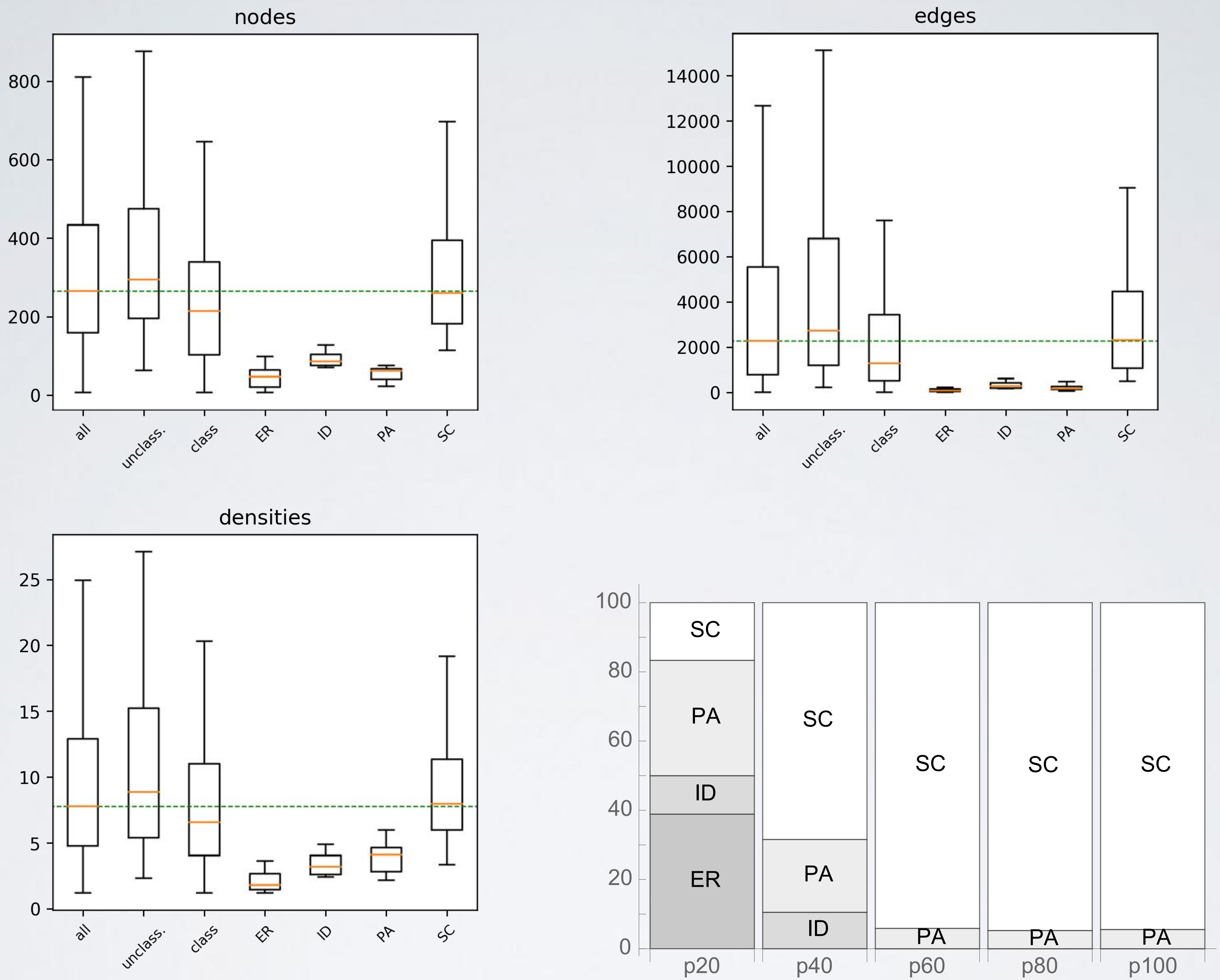


Fig. 4 *Top panel, and bottom-left:* Boxplots of numbers of nodes, edges and densities for the underlying networks of the various families, as well as all, unclassified and classified. Horizontal dashed line indicates overall median. *Bottom-right:* Stacked plot of family ratio per percentile of network density.