Clustering coefficients in large directed graphs

Lasse Leskelä
Aalto University, Finland

joint work with
Mindaugas Bloznelis, U Vilnius

Proc. WAW 2016, https://arxiv.org/abs/1607.02278

20 Dec 2016

Statistical graph models

Uniform random graph with n nodes and m links
- model parameterized by *(n,m)*
- every graph on node set {1,...,*n*} with *m* links is realized with equal probability

Bernoulli random graph with *n* nodes and linkage probability *p*
- model parameterized by (*n*,*p*)
- every node pair is connected with probability *p*, independently of other pairs

Uniform random graph with a given degree list ($d_1,...,d_n$)
- model parameterized by ($n,d_1,...,d_n$)
- aka. configuration model, regular random graph (when d_i = const)

Bernoulli random graph with *n* nodes and node weights a_i
- model parameterized by ($n,a_1,...,a_n$)
- nodes i and j are connected with probability $w(a_i,a_j)$, independently
- special cases: Chung-Lu model, Norros-Reittu model, beta model

Statistical graph models - II

Uniform random graph with *n* nodes and degree distribution *F*
- model parameterized by (*n,F*)
- node degrees are (almost) independent and *F*-distributed
- heavy-tailed degrees when *F* has heavy tails

Bernoulli random graph with n nodes and node weight distribution *F*
- model parameterized by *(n,F)*
- node weights are independent and *F*-distributed
- conditionally on the node weights, each node pair *(i,j)* is linked with probability $w(a_i,a_j)$, independently of other pairs
- special cases: Chung-Lu model, Norros-Reittu model, beta model

These models have heavy-tailed degree distributions when *F* is heavy-tailed.
BUT: no clustering

Clustering in social networks

Friends of your friends
are likely to be friends

Clustering in social networks

Friends of your friends
are likely to be friends

Clustering in social networks

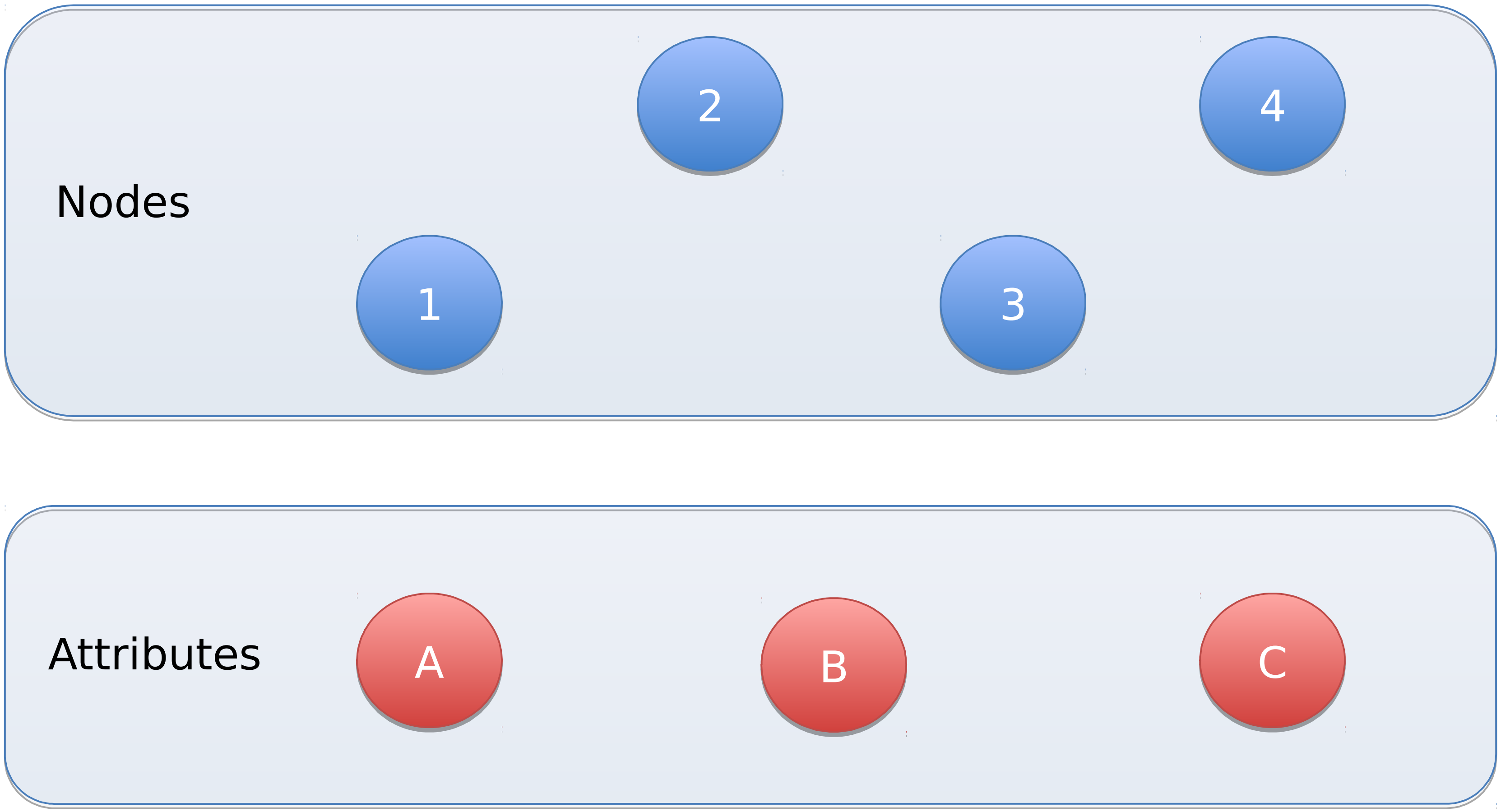
Friends of your friends
are likely to be friends

Shared attributes

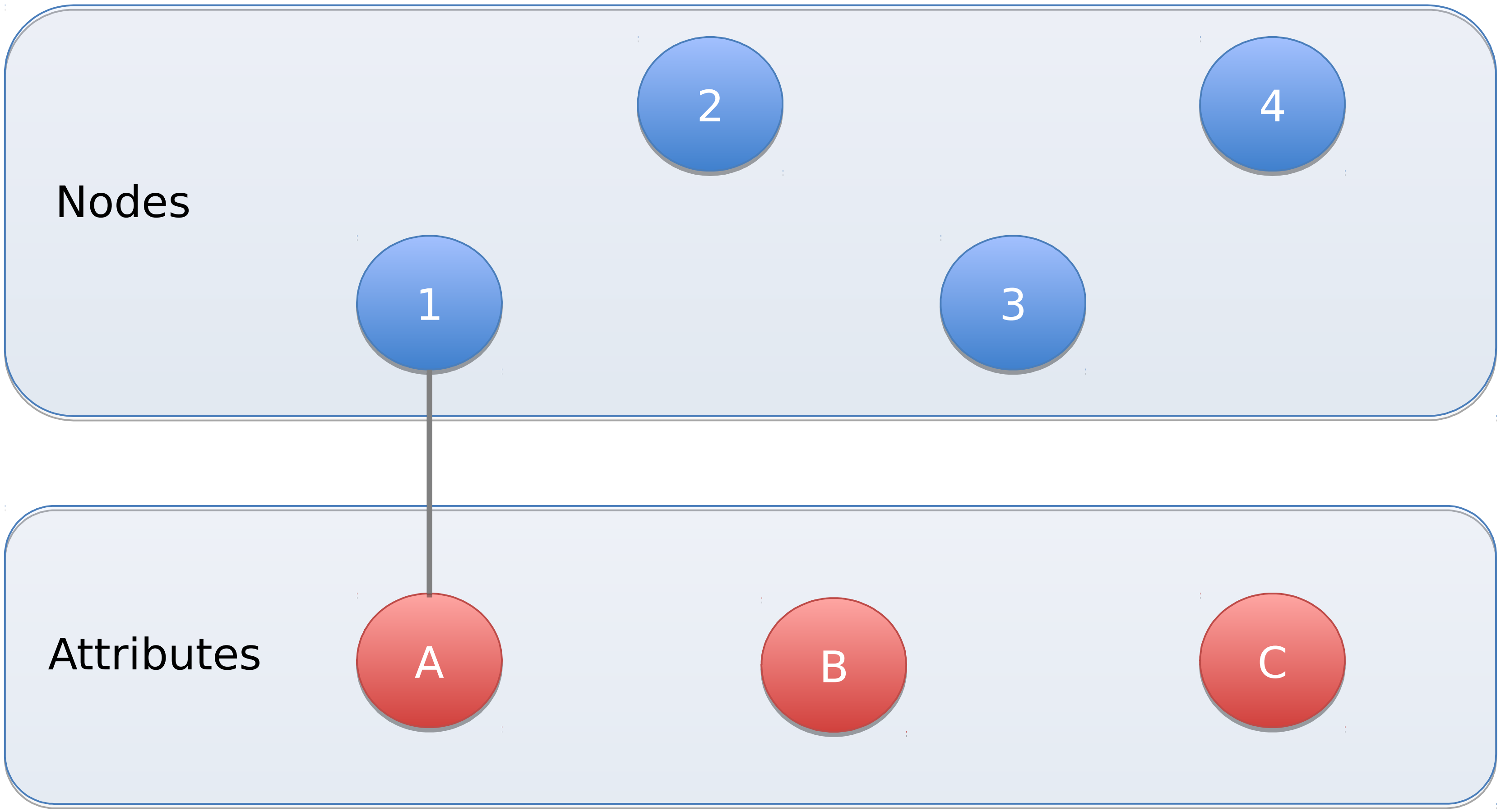
- Common space-time location
- Common relatives
- Common education, jobs, interests
- Common conferences and workshops

I Undirected intersection graphs

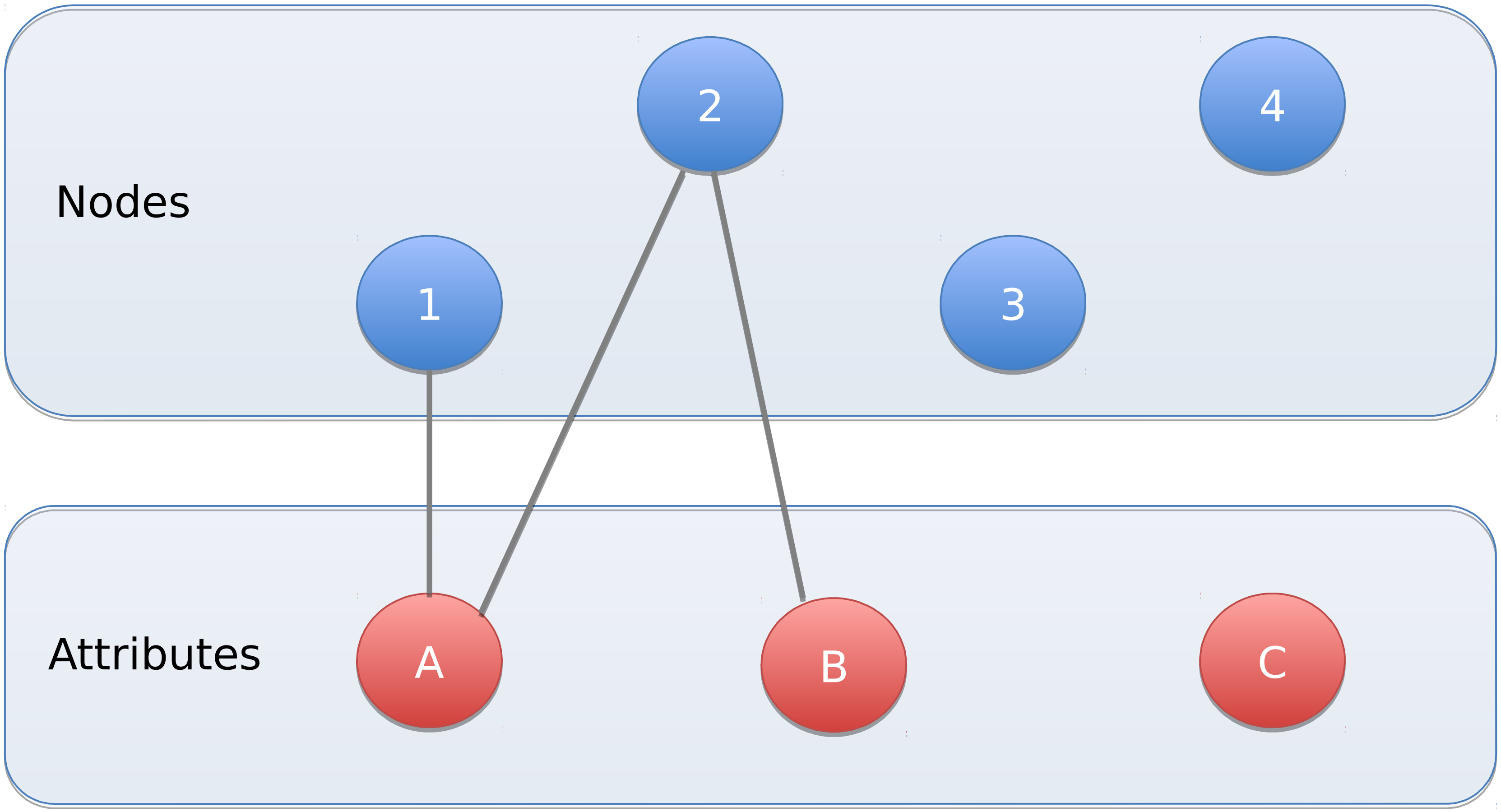
Intersection graph



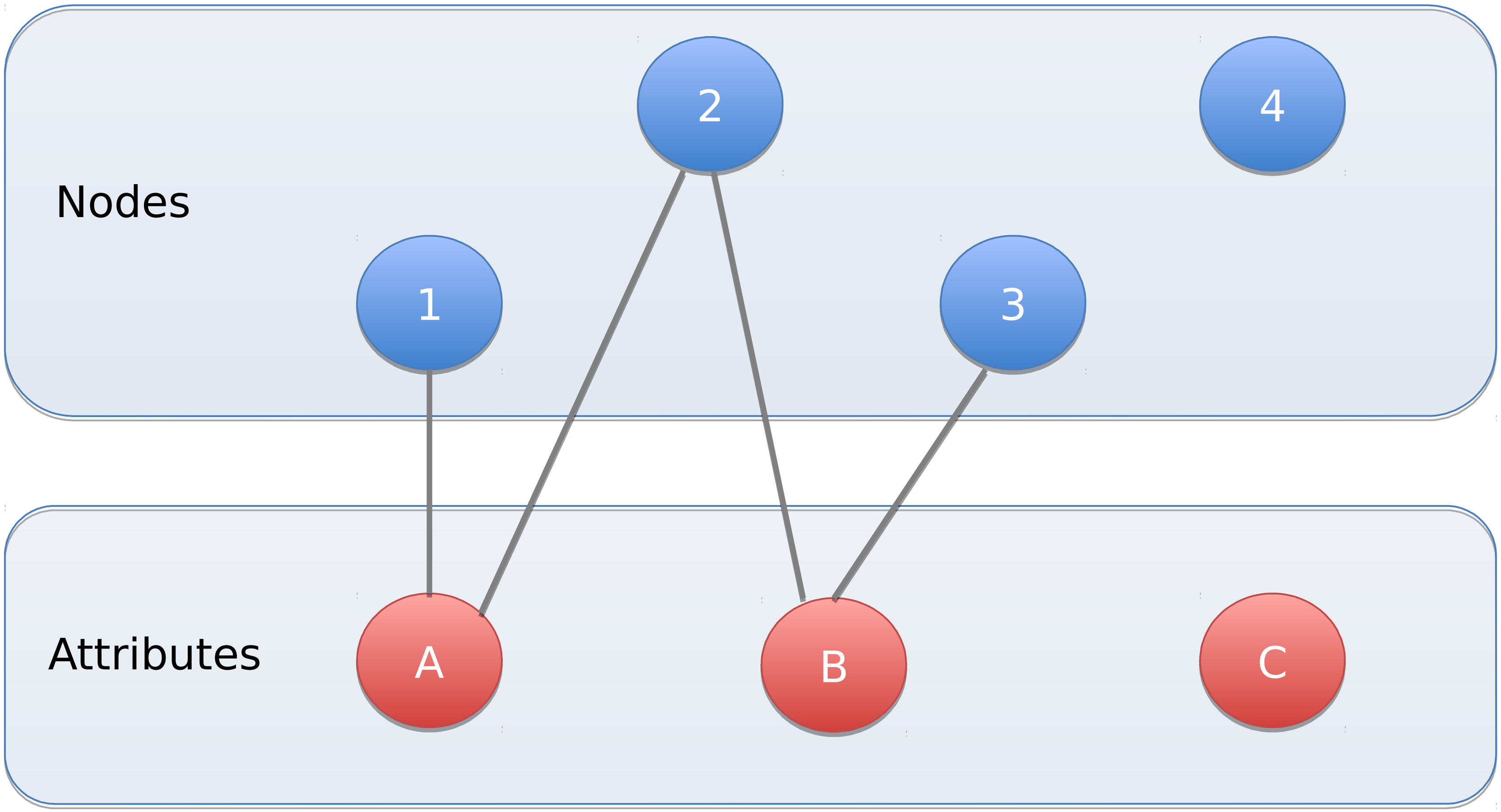
Intersection graph



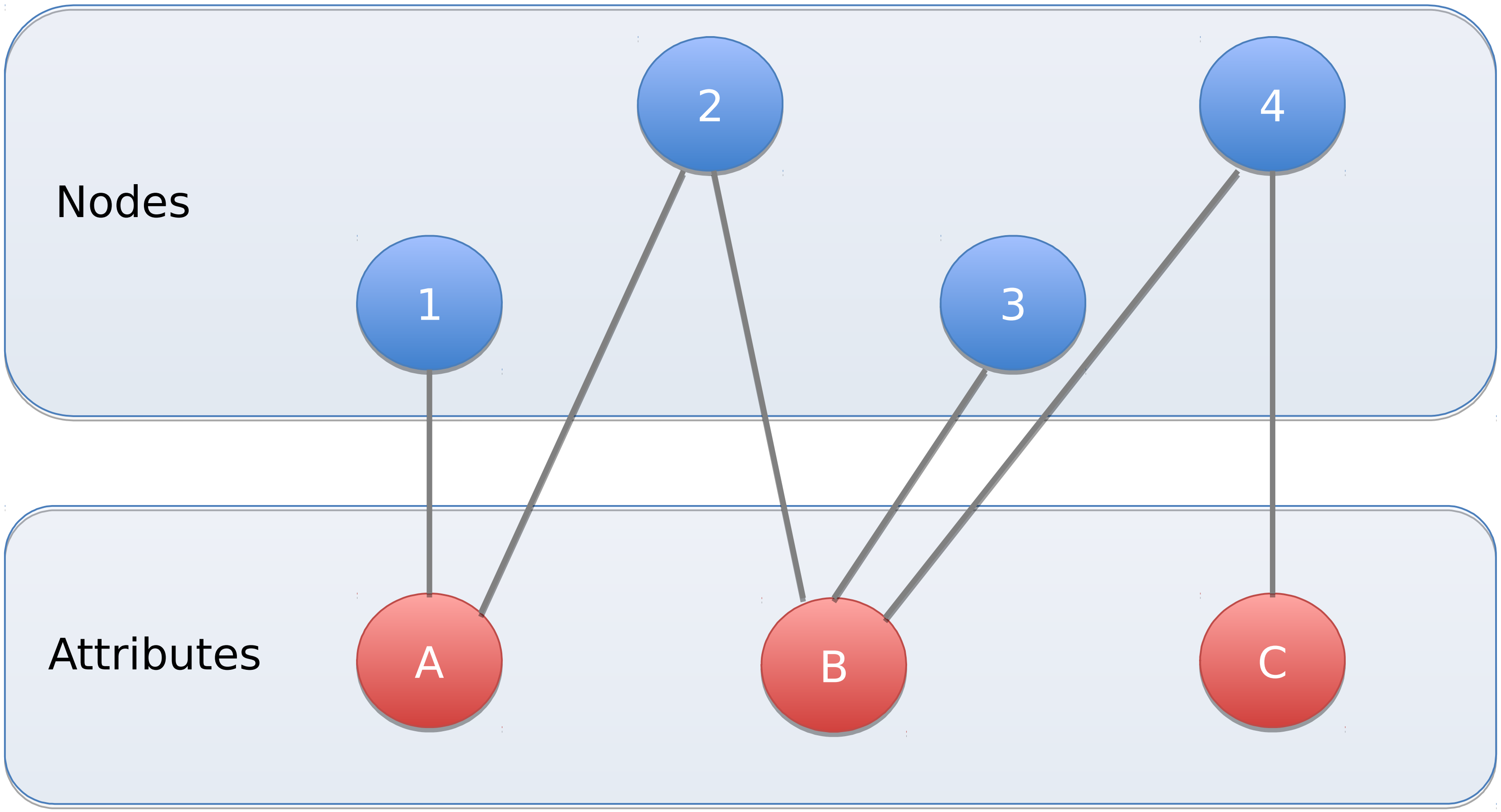
Intersection graph



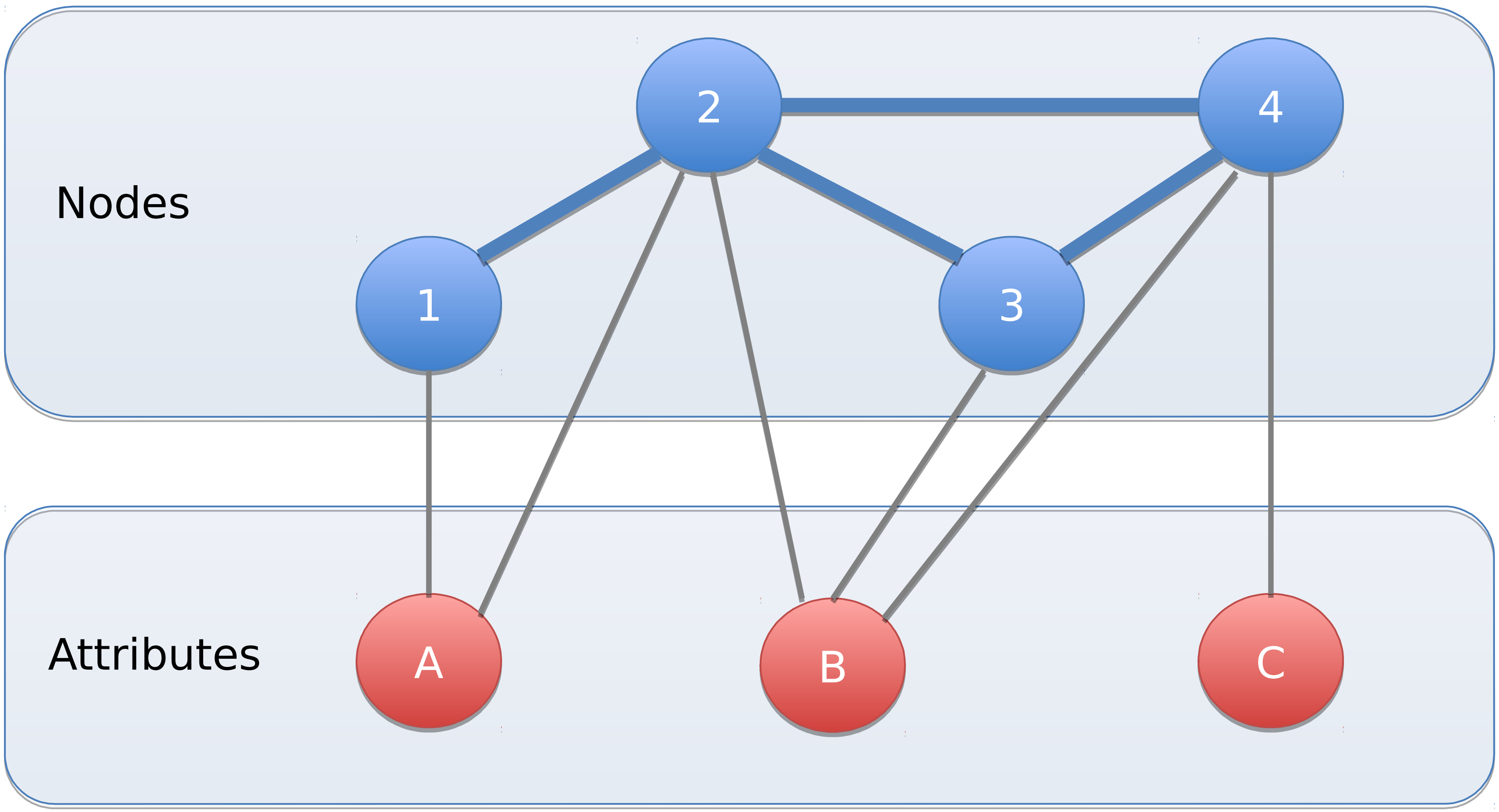
Intersection graph



Intersection graph

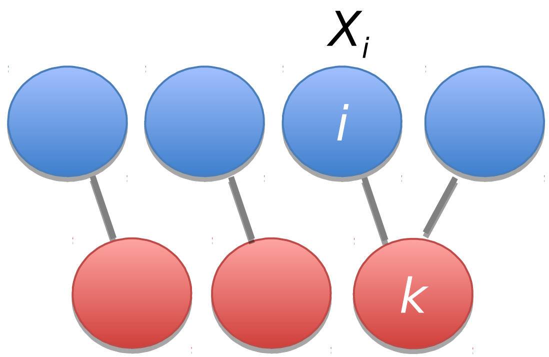


Intersection graph



Two nodes are connected if they share at least one attribute

Random intersection graph



Model parameterized by (n, m, γ, F_{node})

- n nodes
- m attributes
- γ overall attribute density
- node labels X_i distr. as F_{node}, $i=1,\ldots,n$

Given the node labels, node i selects attribute k w.pr. $\min(\gamma X_i, 1)$

$$P_X(i \leftrightarrow j) \sim \begin{cases} X_i X_j m \gamma^2, & \gamma \ll m^{-1/2}, \\ 1 - e^{-X_i X_j m \gamma^2}, & \gamma \sim m^{-1/2}, \\ 1, & \gamma \gg m^{-1/2}. \end{cases}$$

K Singer-Cohen 1995. Stark 2004. M Deijfen, W Kets 2009. M Bloznelis 2013.

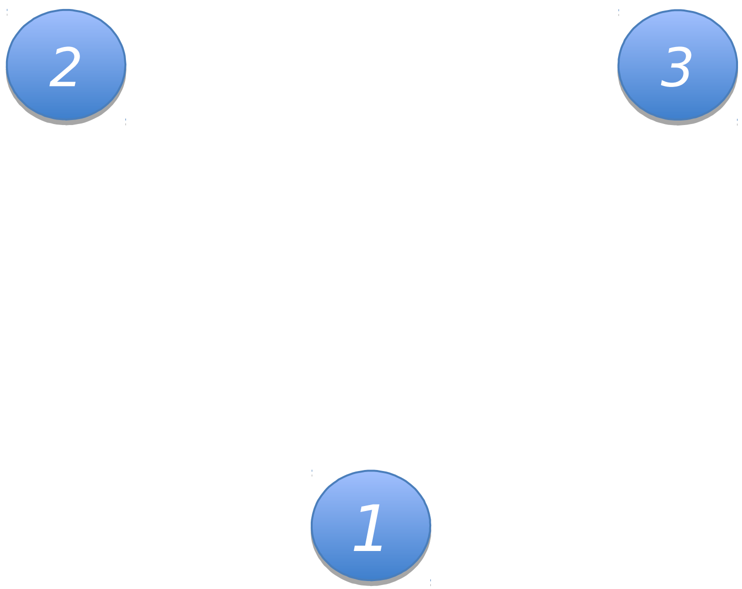
Degree distribution

For $nm\gamma^2 \sim \lambda$, $\gamma \ll m^{-1/2}$

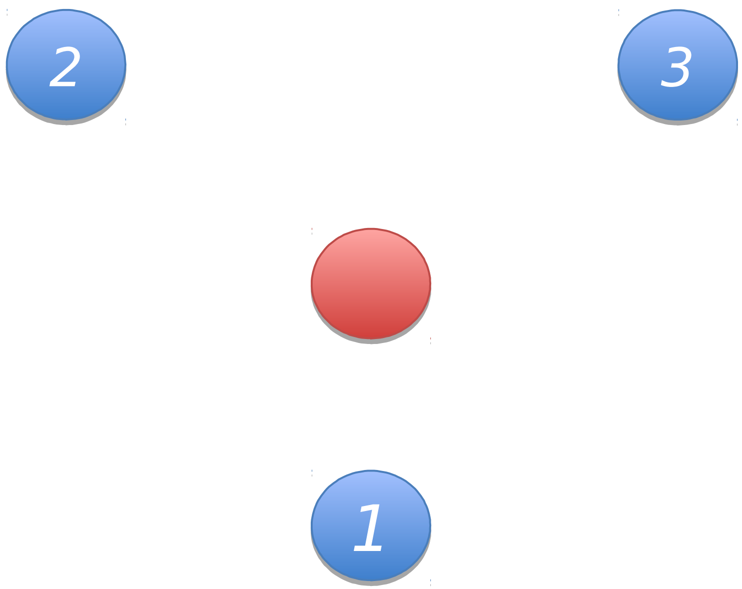
$$\deg(i) \approx \begin{cases} 0, & m \ll n, \\ \mathrm{MPoi}\left((\frac{\lambda}{\beta})^{1/2} E(X_i)\, \mathrm{MPoi}\left((\lambda\beta)^{1/2} X_i\right)\right), & m \sim \beta n, \\ \mathrm{MPoi}(\lambda X_i), & m \gg n. \end{cases}$$

When X_i has a power-law tail, then so does deg(i).

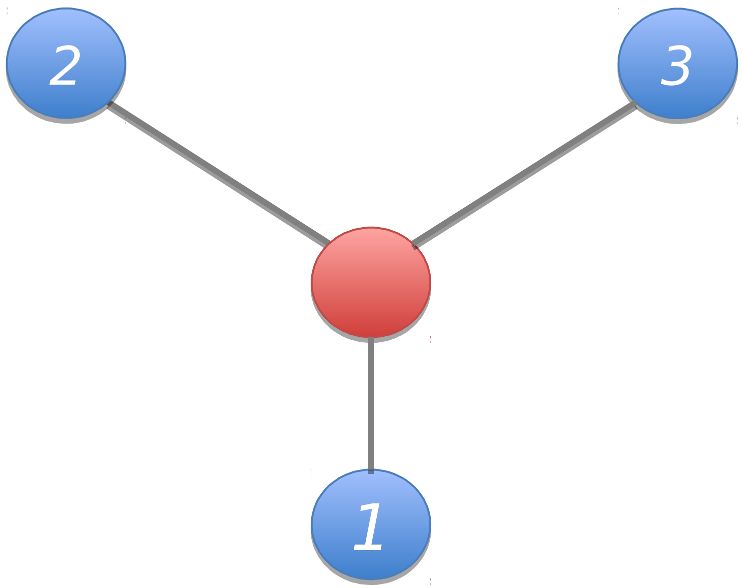
Forming triangles



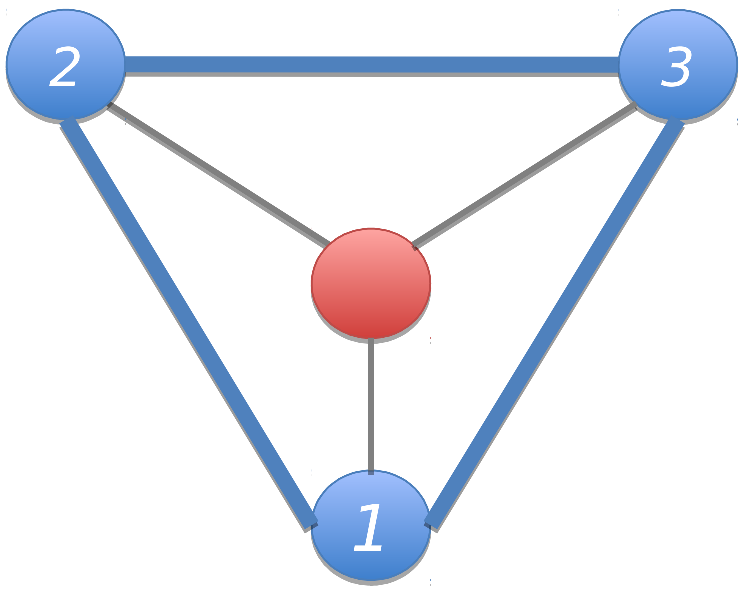
Forming triangles



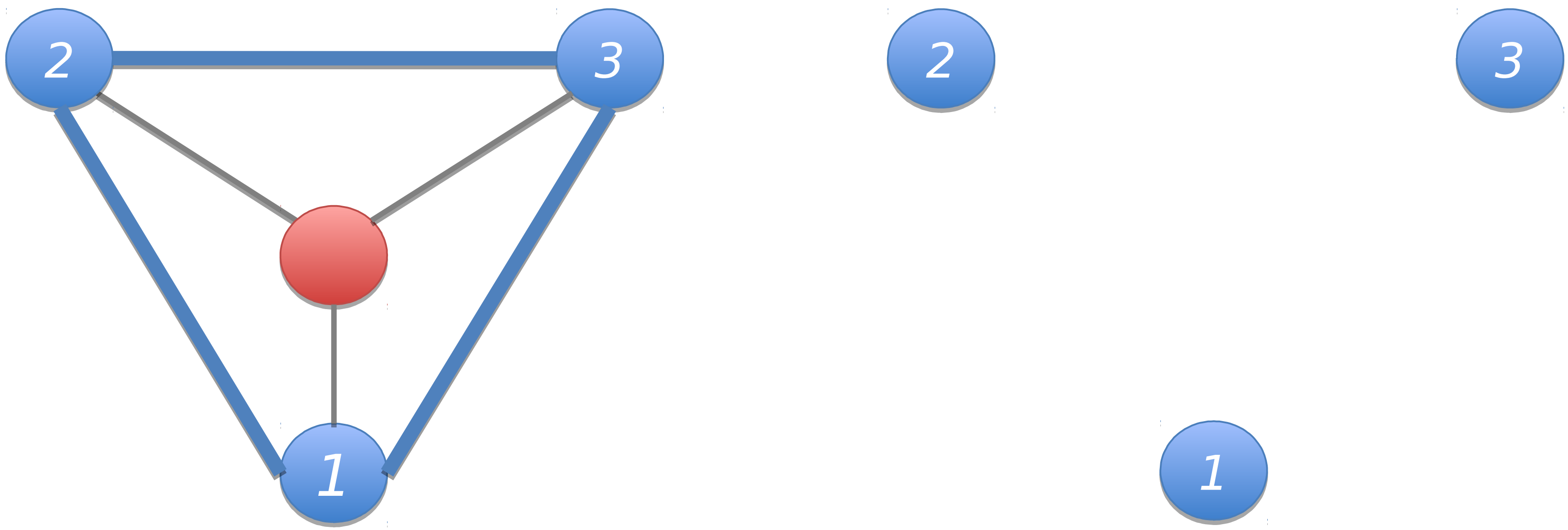
Forming triangles



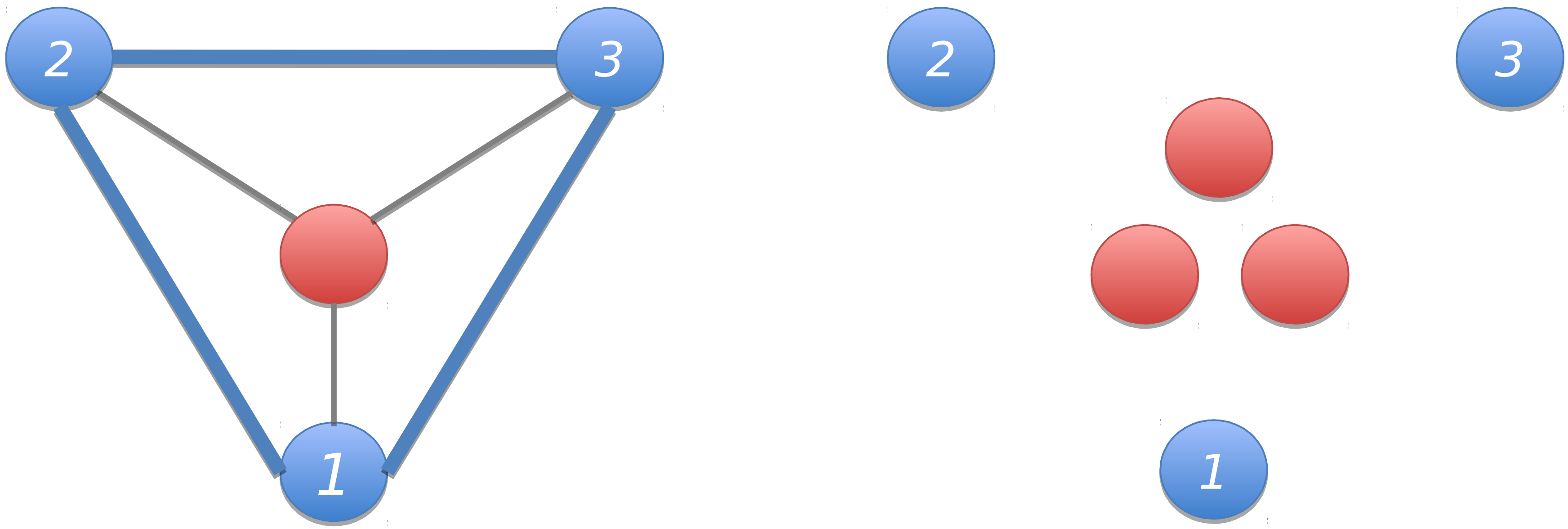
Forming triangles



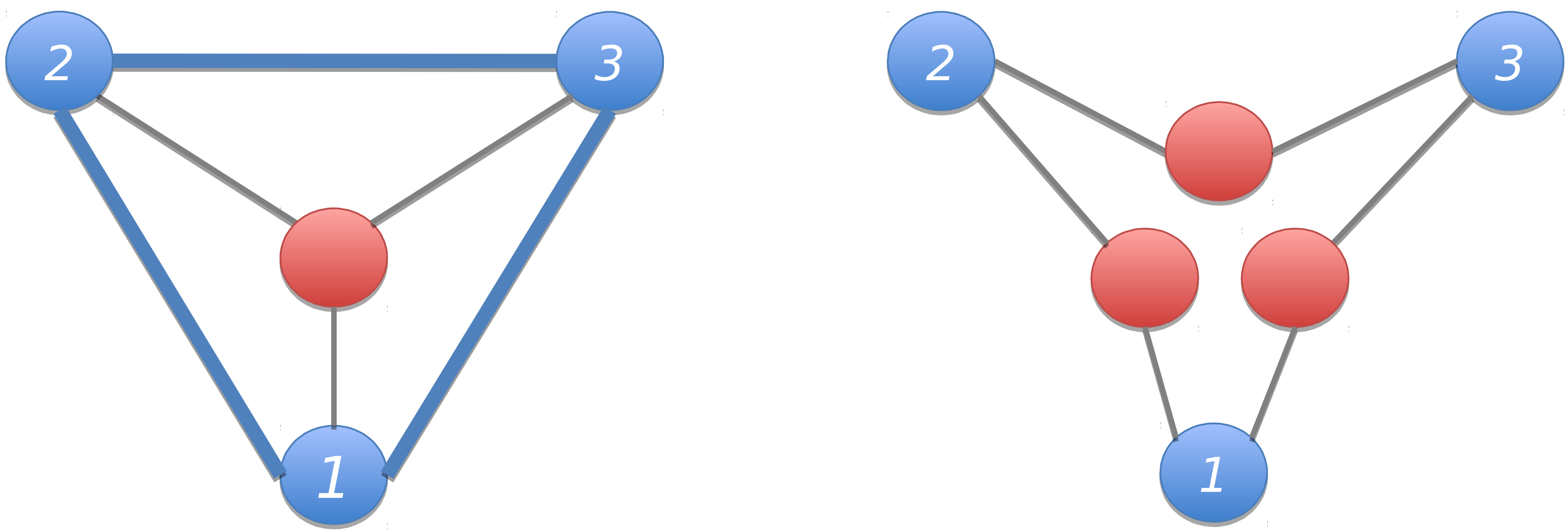
Forming triangles



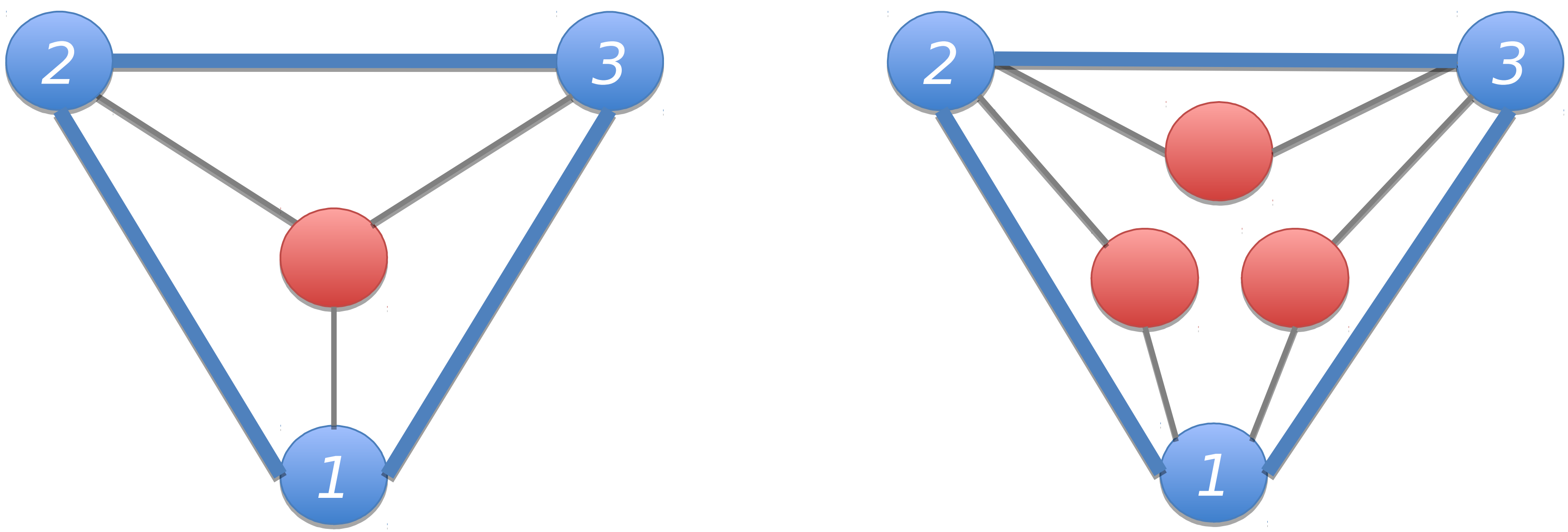
Forming triangles



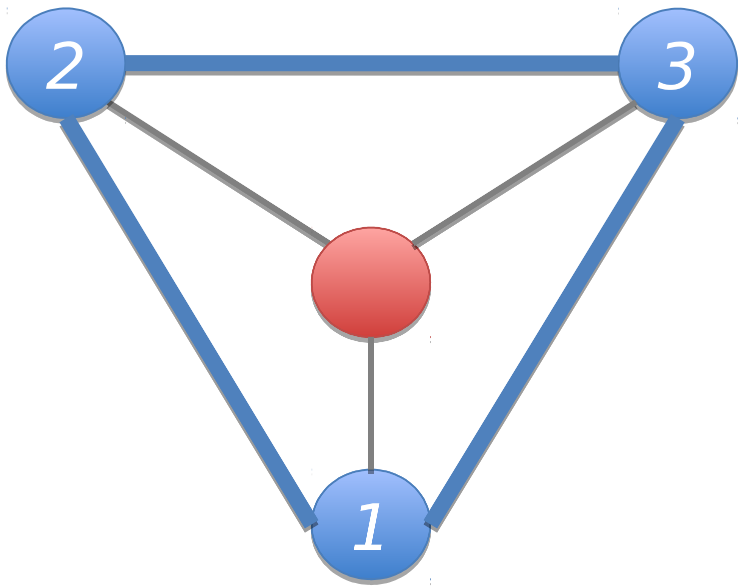
Forming triangles

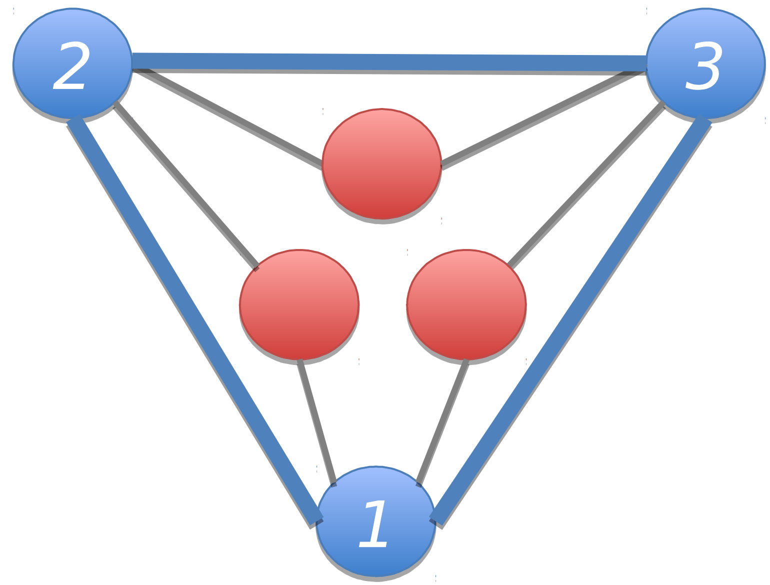


Forming triangles



Forming triangles

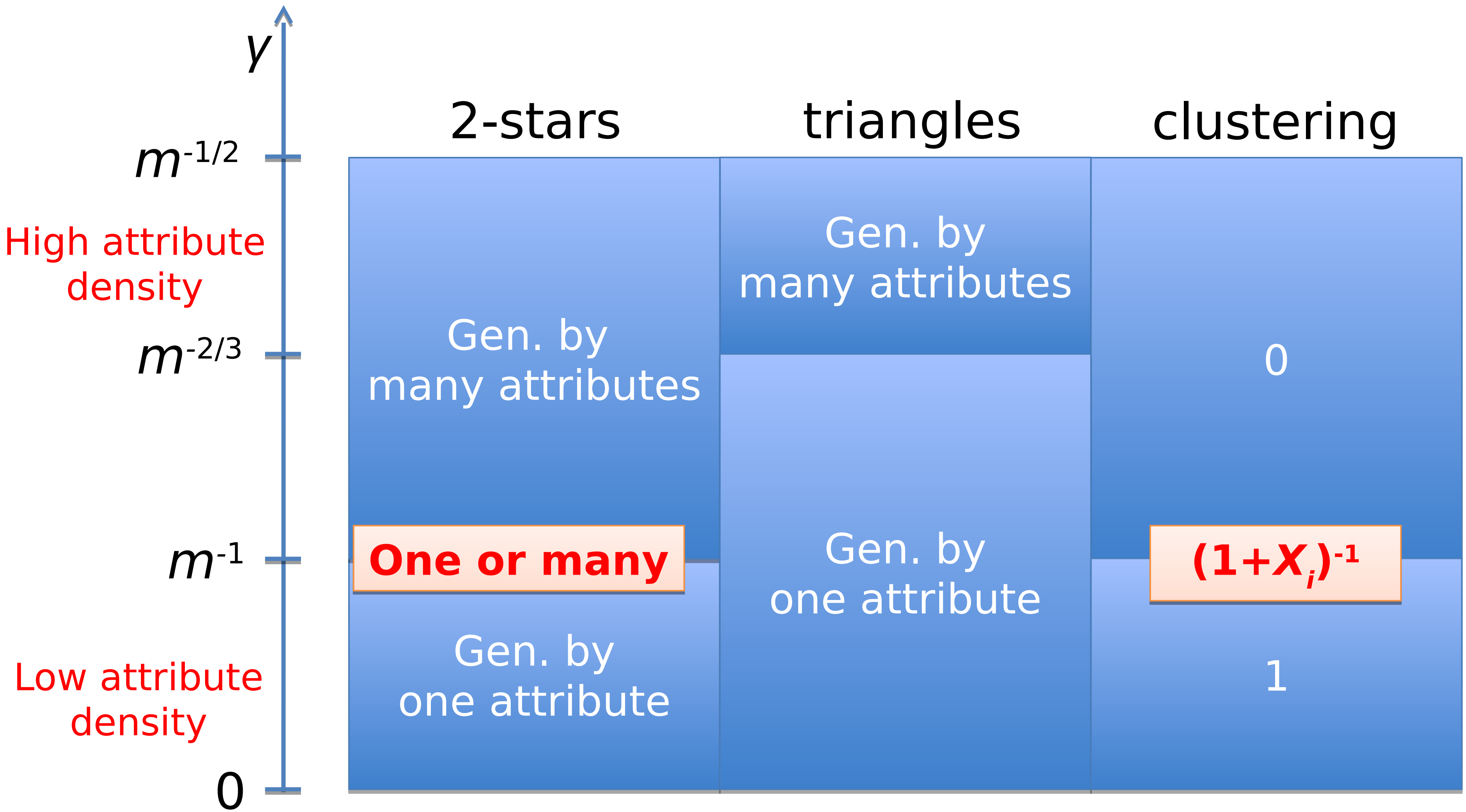




$$X_1 X_2 X_3 m \gamma^3 + O(m^2 \gamma^6)$$

$$X_1^2 X_2^2 X_3^3 m^3 \gamma^6 + O(m^4 \gamma^8)$$

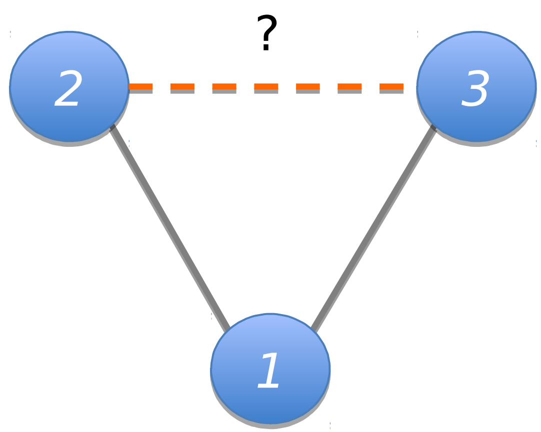
Clustering vs. attribute density



Clustering

In a graph with *m* >> 1 attributes, the leaves of a 2-star centered at *1* are linked with probability

$$P_X(2 \leftrightarrow 3 \,|\, 1 \leftrightarrow 2, 1 \leftrightarrow 3) \sim \begin{cases} 1, & \gamma \ll m^{-1}, \\ \frac{1}{1+\alpha X_1}, & \gamma \sim \alpha m^{-1}, \\ 0, & \gamma \gg m^{-1}. \end{cases}$$

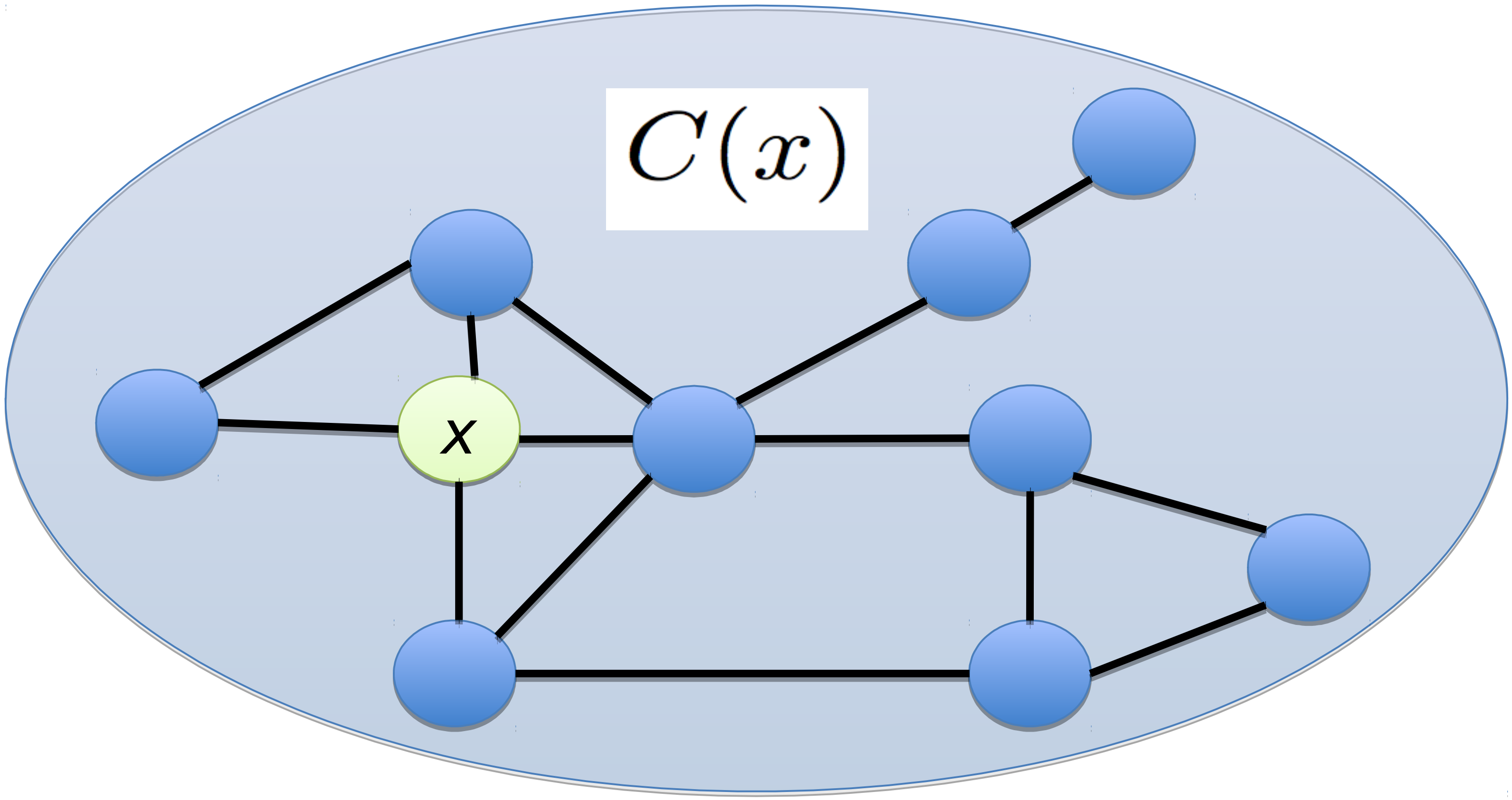


Model parameters:

- number of nodes *n*
- number of attributes *m*
- attribute density *γ*
- node labels X_i

M Deijfen, W Kets 2009.

Connected components



Component of node x

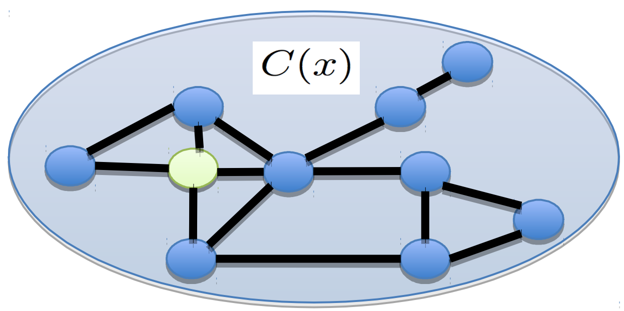
Connected components

In a homogeneous (X_i=1) large graph (n >> 1) with many attributes (m >> 1) and attribute density $\gamma \approx \lambda^{1/2}(mn)^{-1/2}$,

$$\frac{c_{\max}}{n} \xrightarrow{\mathbb{P}} \zeta$$

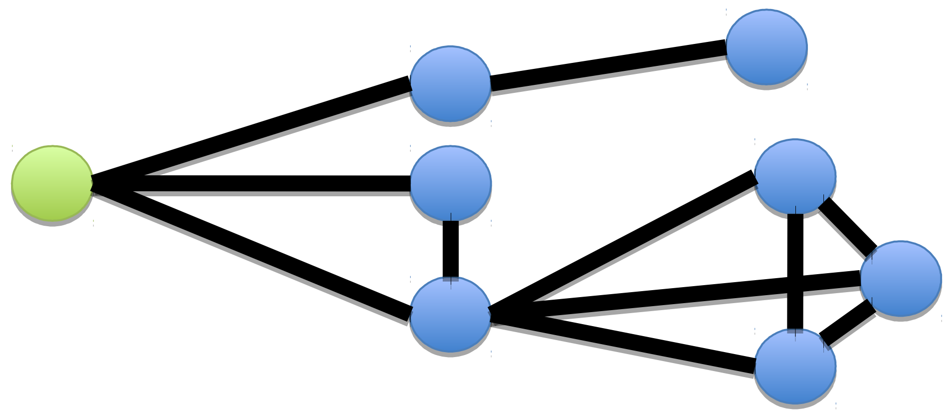
where ς is the survival probability of a Galton-Watson branching process with offspring

$$\deg(x) \approx \begin{cases} \mathrm{Poi}\,(\lambda)\,, & m \gg n \\ \mathrm{Poi}\,\left((\lambda/\beta)^{1/2}\,\mathrm{Poi}\,\left((\lambda\beta)^{1/2}\right)\right), & m \sim \beta n \end{cases}$$



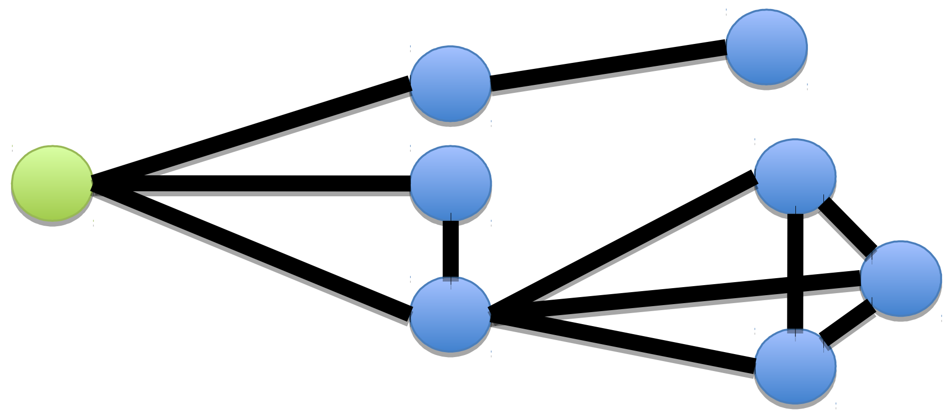
M Behrisch 2007. A Lagerås, M Lindholm 2008. M Bloznelis 2010.
F Ball, D Sirl, P Trapman 2014.

Why branching analysis works?

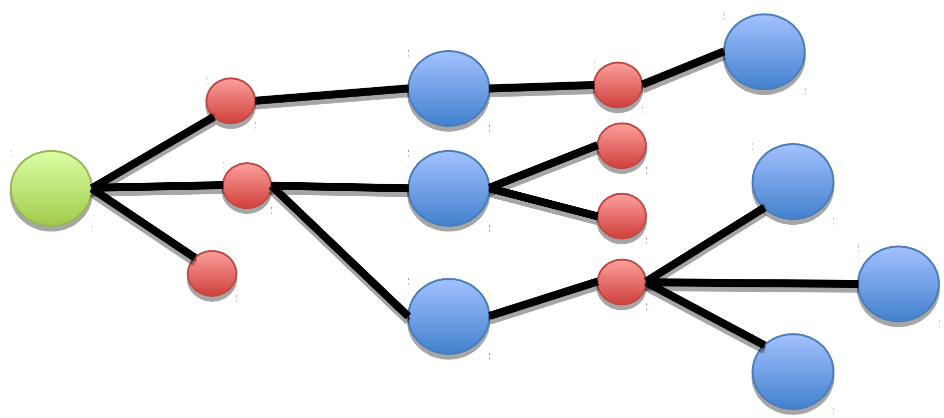


The random intersection graph is not locally treelike

Why branching analysis works?

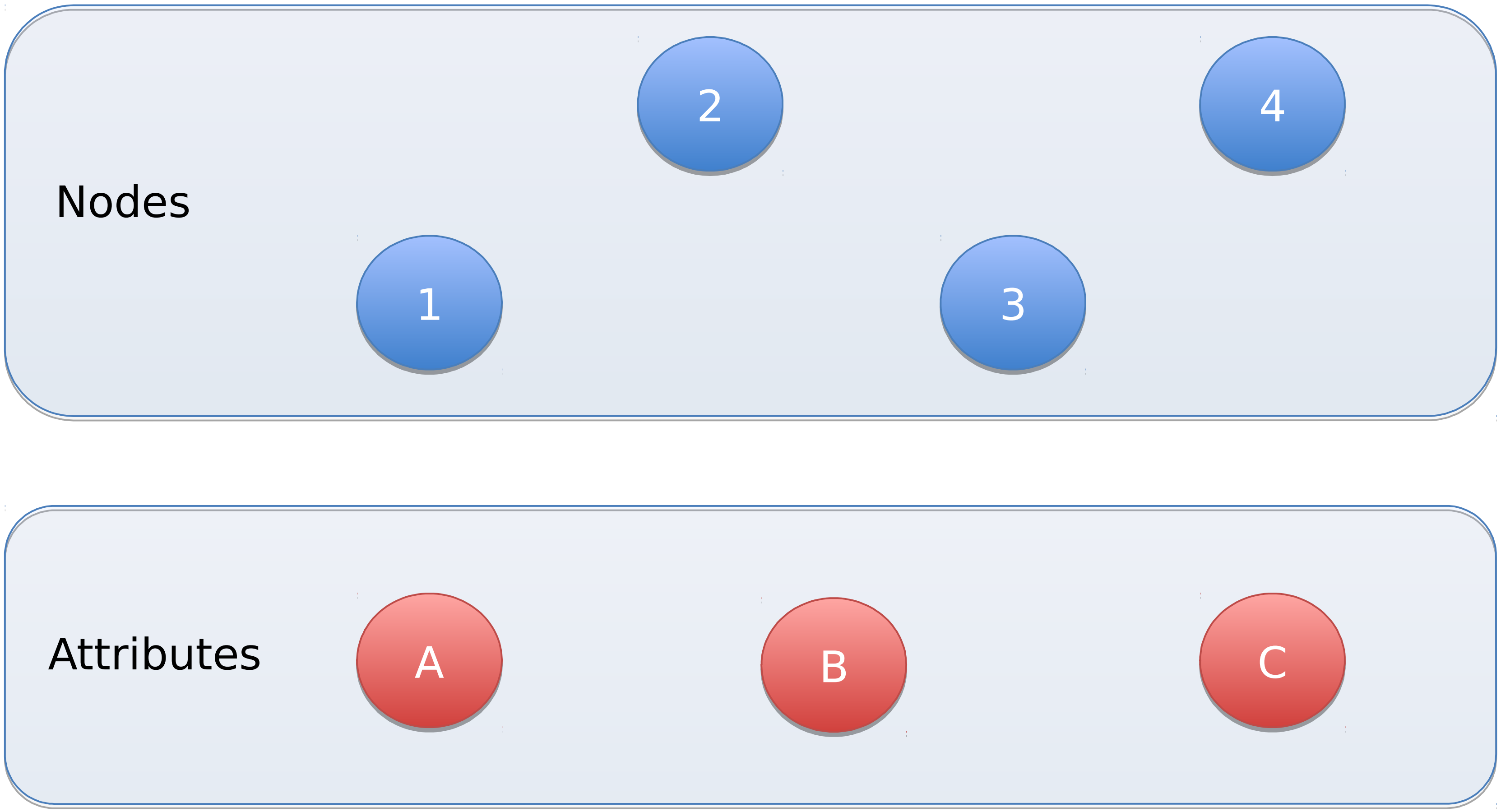


The random intersection graph is not locally treelike but the underlying random bipartite graph is (whp).

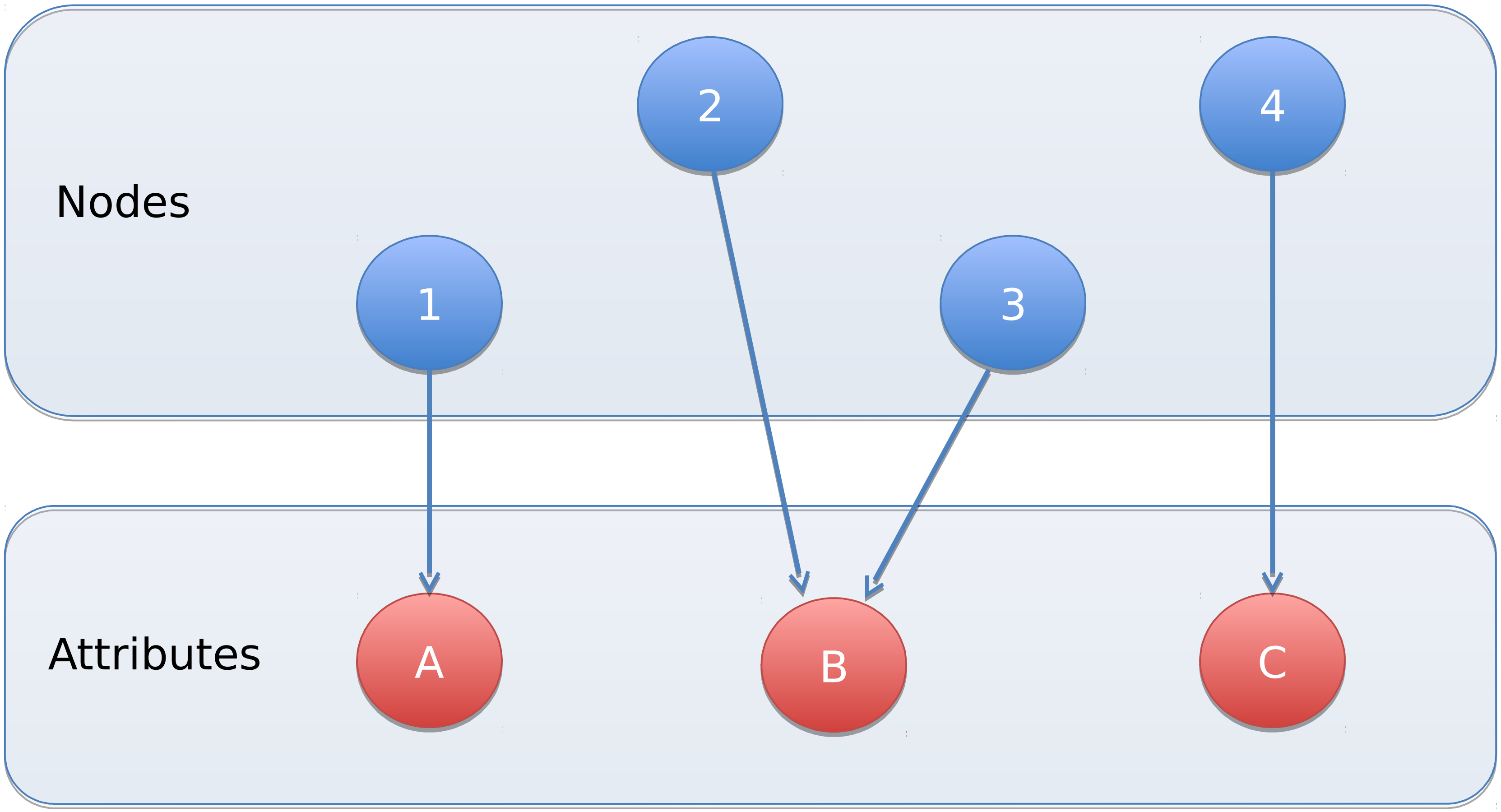


II Directed intersection graphs

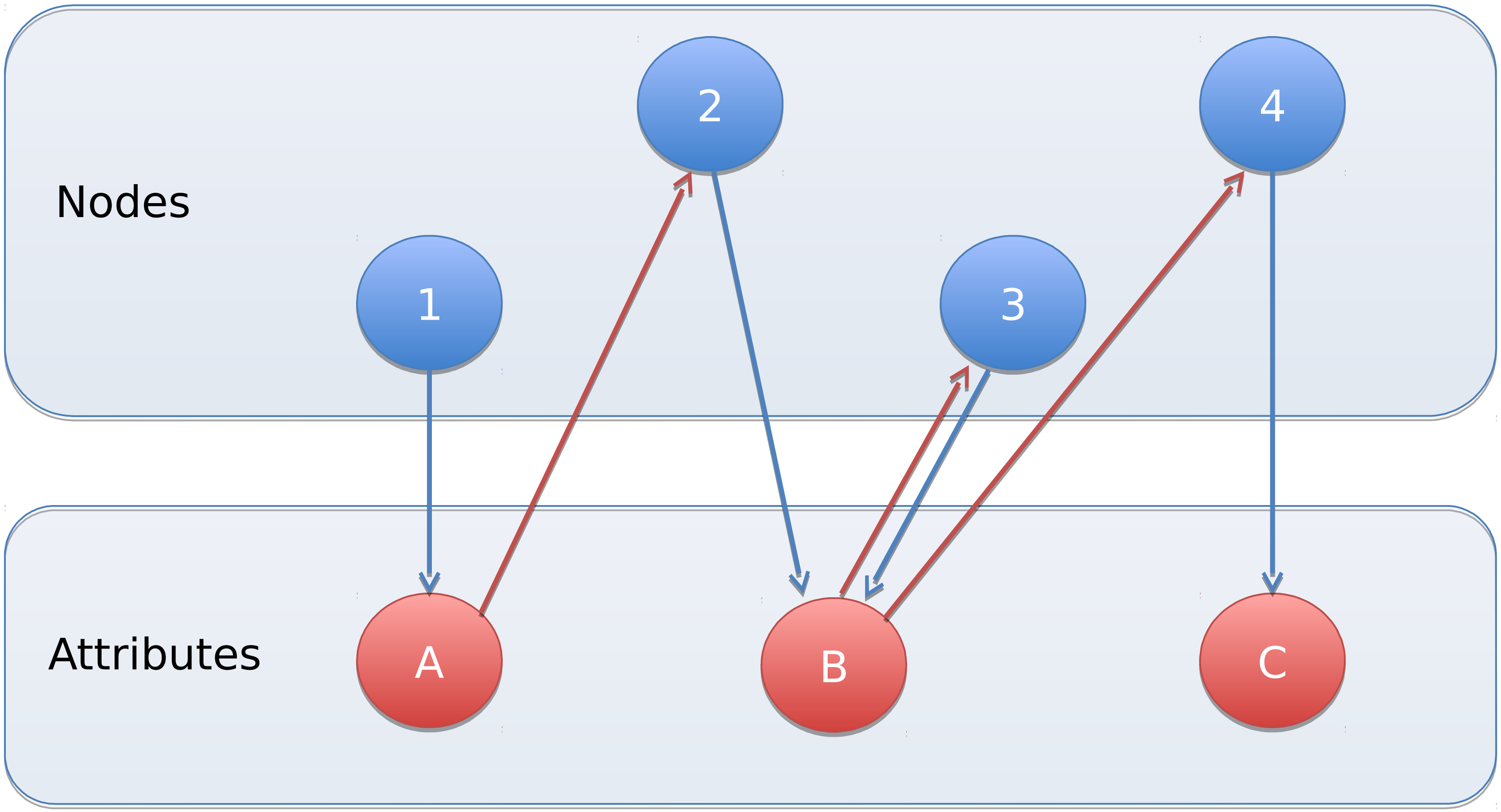
Directed intersection graph



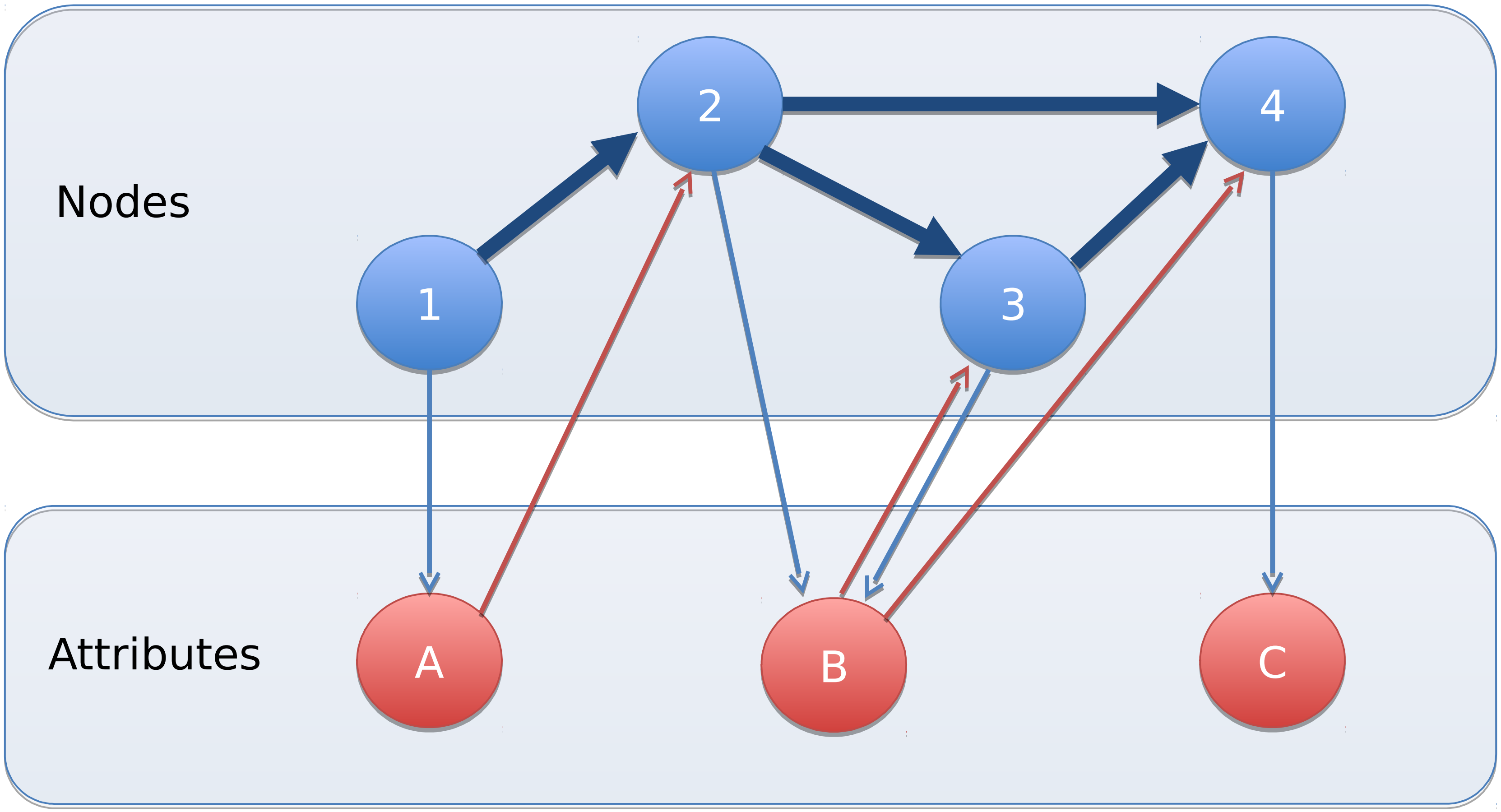
Directed intersection graph



Directed intersection graph

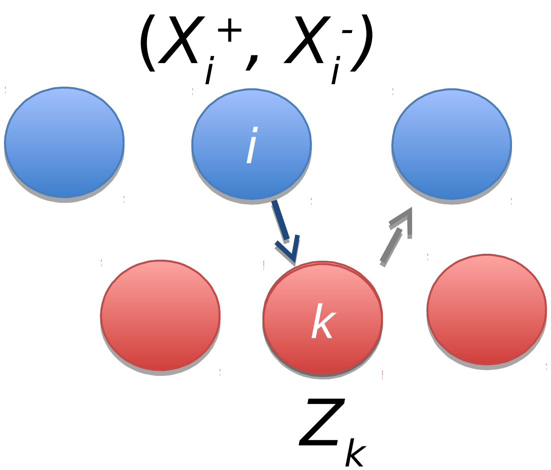


Directed intersection graph



i follows *j*, if *i* demands something that *j* supplies

Directed random intersection graph



Model parameterized by $(n, m, \gamma, F_{node}, F_{attr})$

Node labels $(X_i^+ X_i^-)$ distr. as $F_{node,}$ $i=1,\ldots,n$

Attribute labels Z_k distr. as F_{attr} , $k=1,\ldots,m$

Given the node and attribute labels, node i

- demands attribute k w.pr. $\min(\gamma X_i^+ Z_k, 1)$
- supplies attribute k w.pr. $\min(\gamma X_i^- Z_k, 1)$

$$P_X(i \to j) \sim \begin{cases} X_i^+ X_j^- m\gamma^2, & \gamma \ll m^{-1/2}, \\ 1 - e^{-X_i^+ X_j^- m\gamma^2}, & \gamma \sim m^{-1/2}, \\ 1, & \gamma \gg m^{-1/2}. \end{cases}$$

Outdegree and indegree

For $\gamma \sim \alpha m^{-1}$ and $m \sim \beta n$ with $m, n \gg 1$,

$$\deg^+(i) \;\approx\; \sum_{k=1}^{M_i^+} N_k^- \qquad \deg^-(i) \;\approx\; \sum_{k=1}^{M_i^-} N_k^+$$

with $M_i^{\pm} =_{\mathrm{st}} \mathrm{MPoi}(\alpha X_i^{\pm} E(Z_1))$ and $N_k^{\pm} =_{\mathrm{st}} \mathrm{MPoi}((\alpha/\beta) E(X_i^{\pm}) Z_k^*)$.

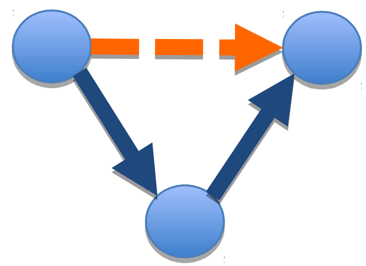
Z^*_k is a size-biased version of Z_k

When X_i^+ has a power-law tail, so does $\deg^+(i)$.

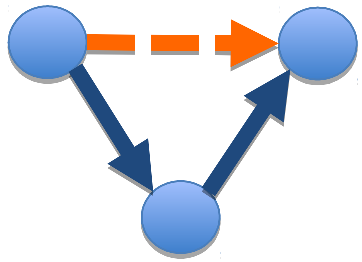
When X_i^- has a power-law tail, so does $\deg^-(i)$.

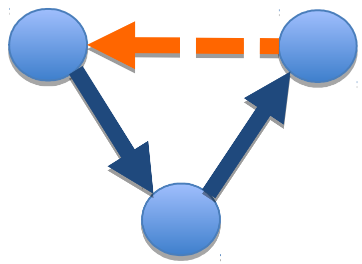
M Bloznelis 2010.

Clustering among triplets

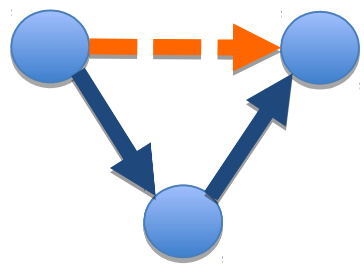


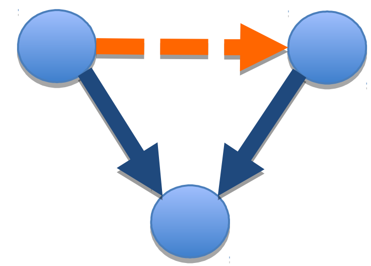
Clustering among triplets

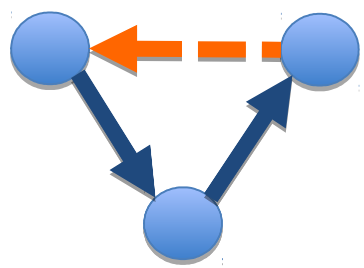




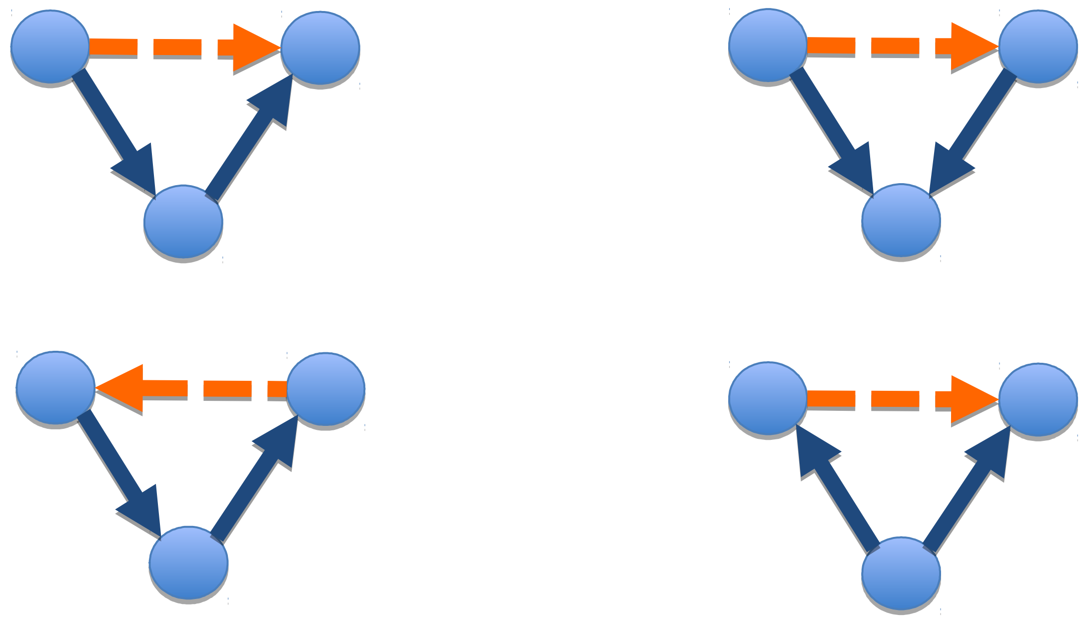
Clustering among triplets



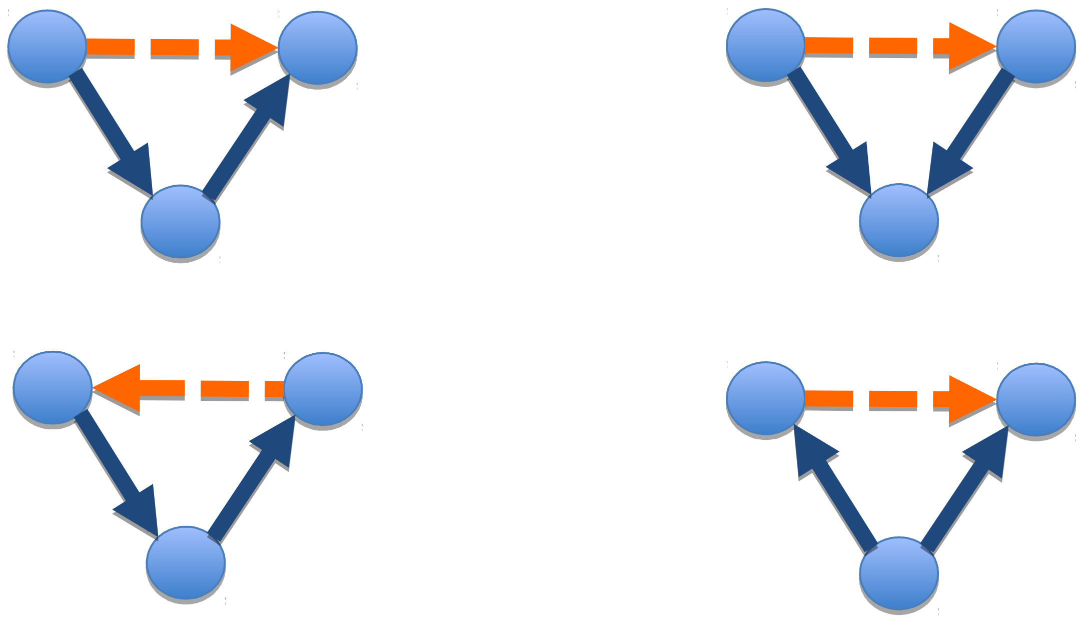




Clustering among triplets

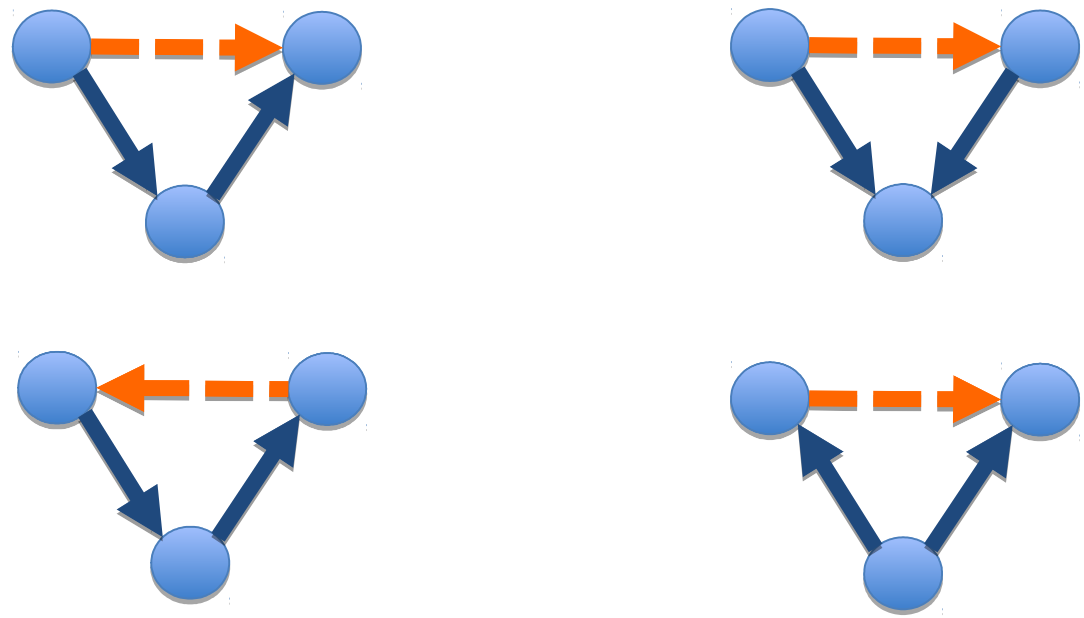


Clustering among triplets



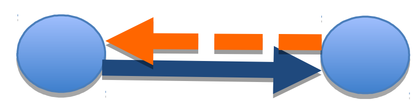
The above completions are unlikely in the sparse regime with $\gamma << m^{-1/2}$.

Clustering among triplets



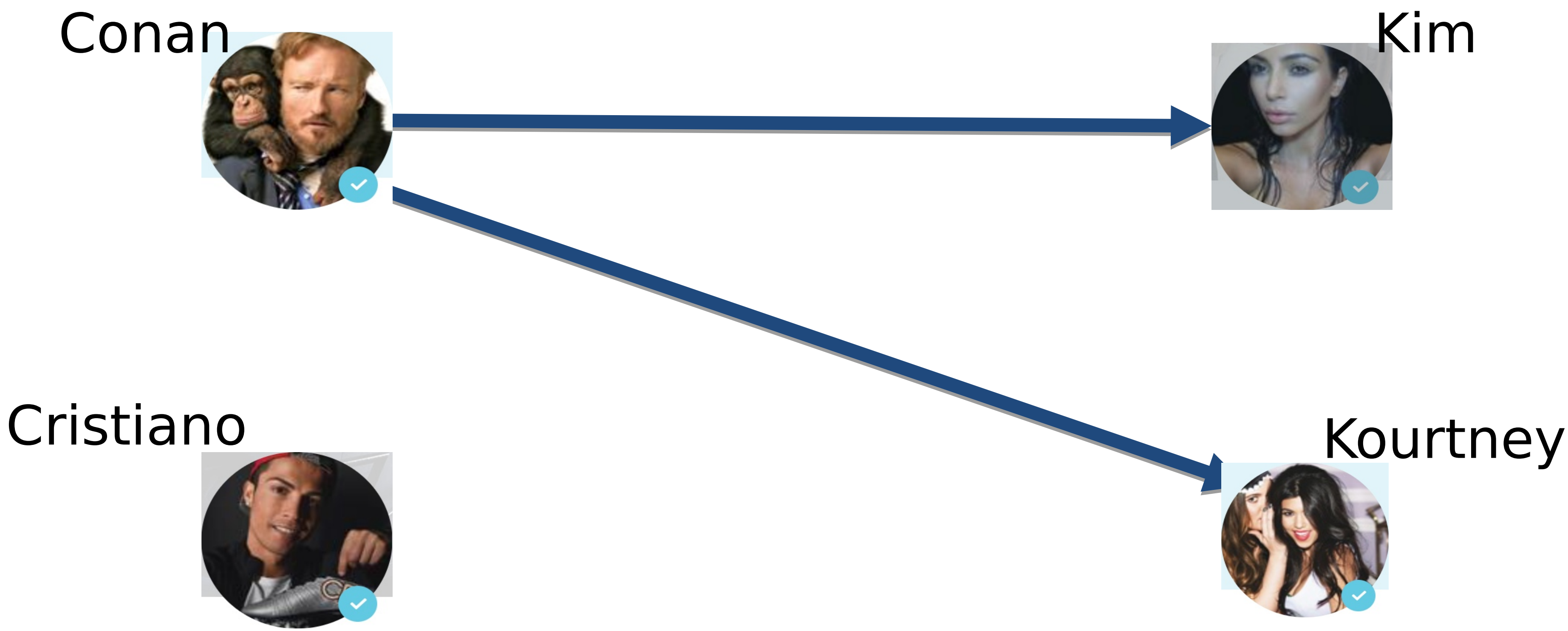
The above completions are unlikely in the sparse regime with $\gamma << m^{-1/2}$.

Link reversals are unlikely as well.



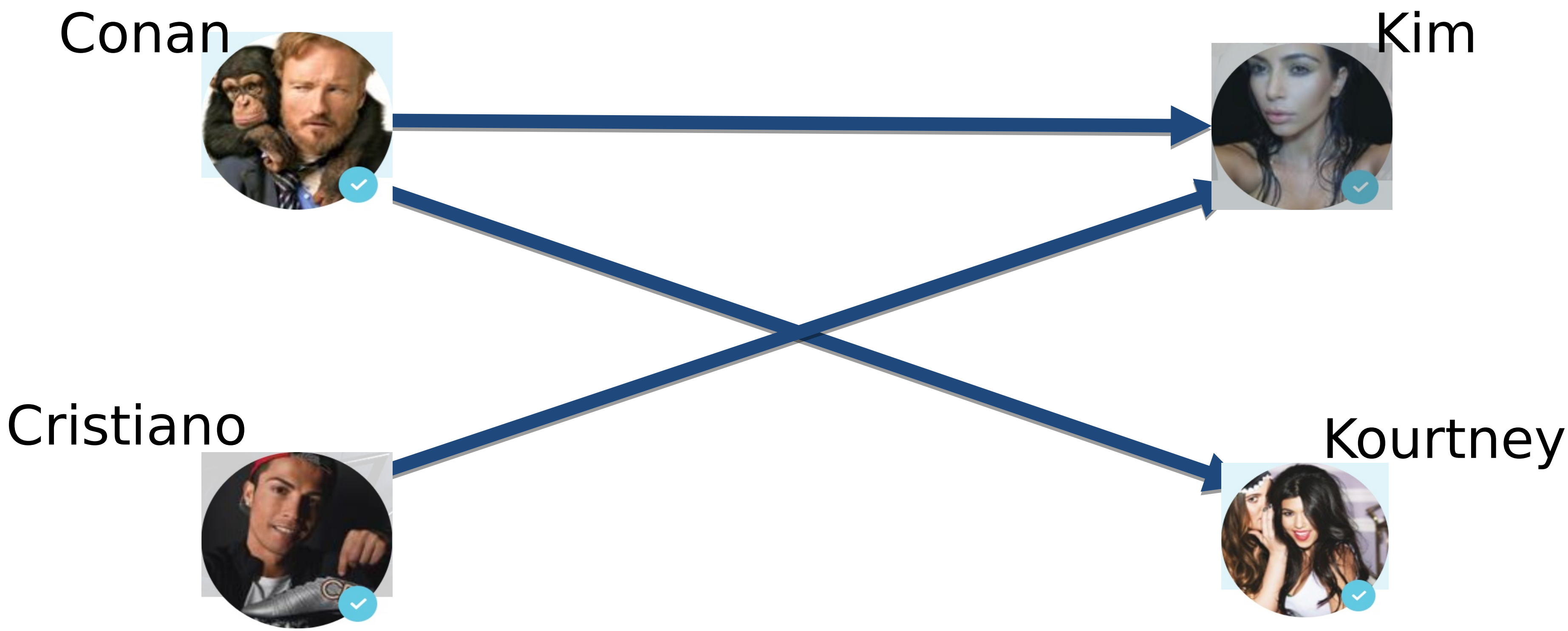
Diclique clustering

Diclique clustering



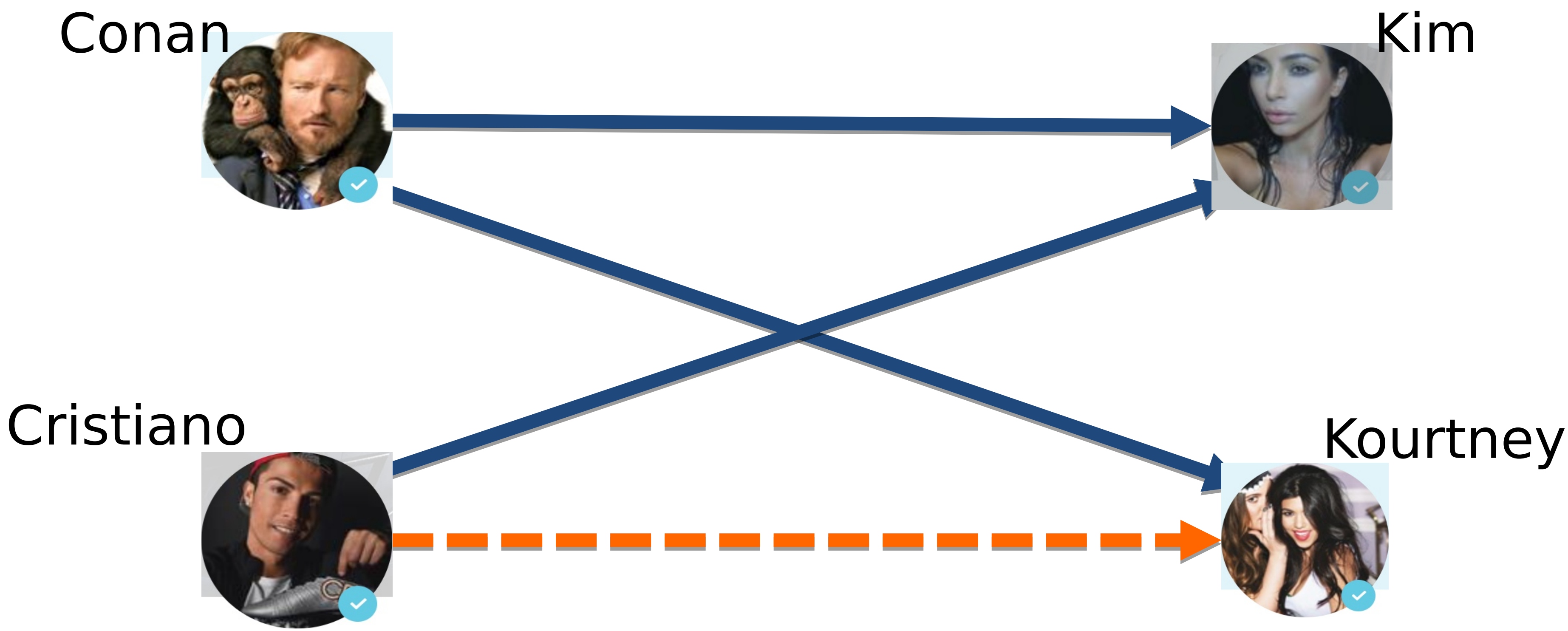
If Conan follows Kim and Kourtney,

Diclique clustering



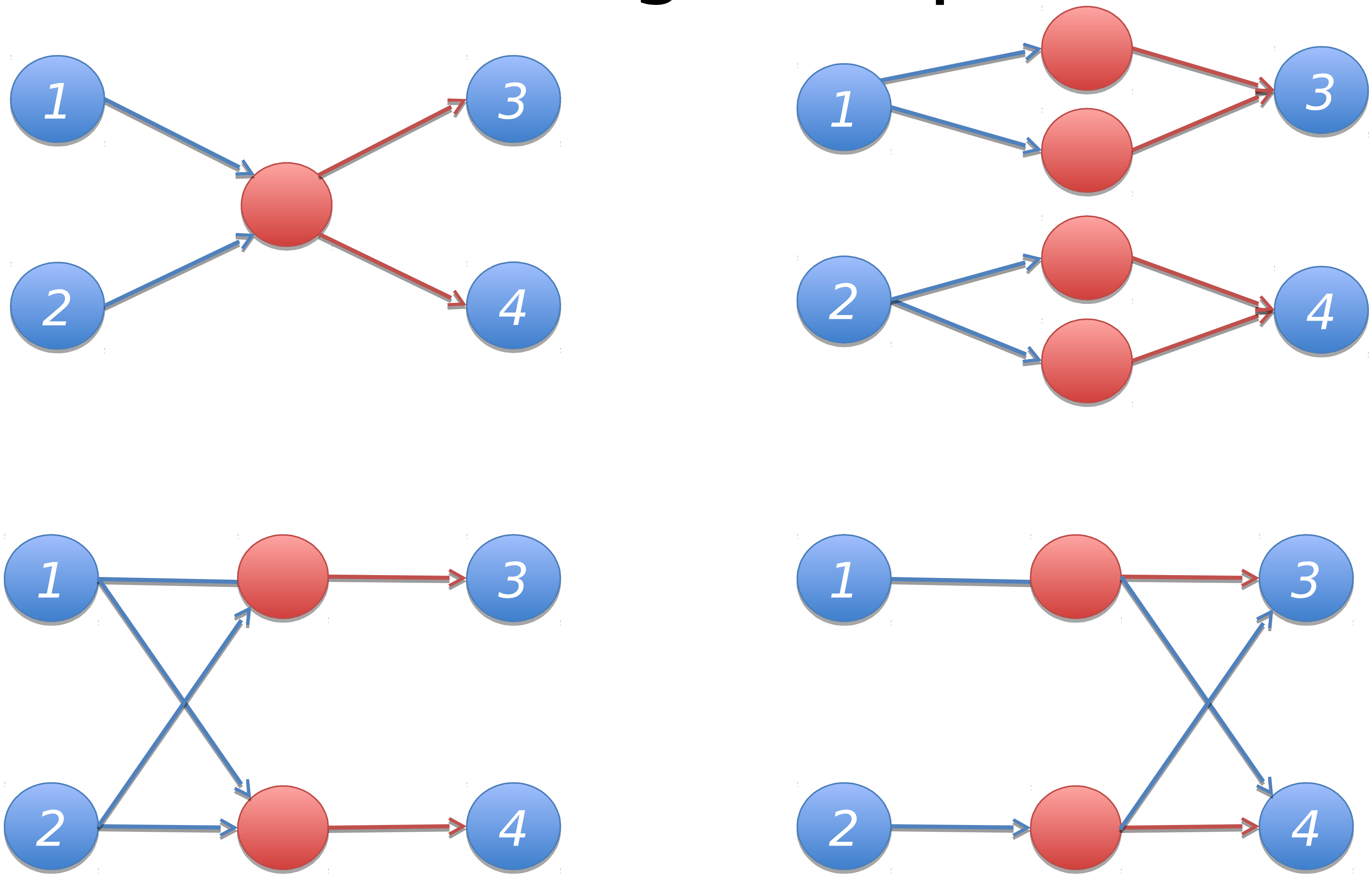
If Conan follows Kim and Kourtney, and Cristiano follows Kim,

Diclique clustering

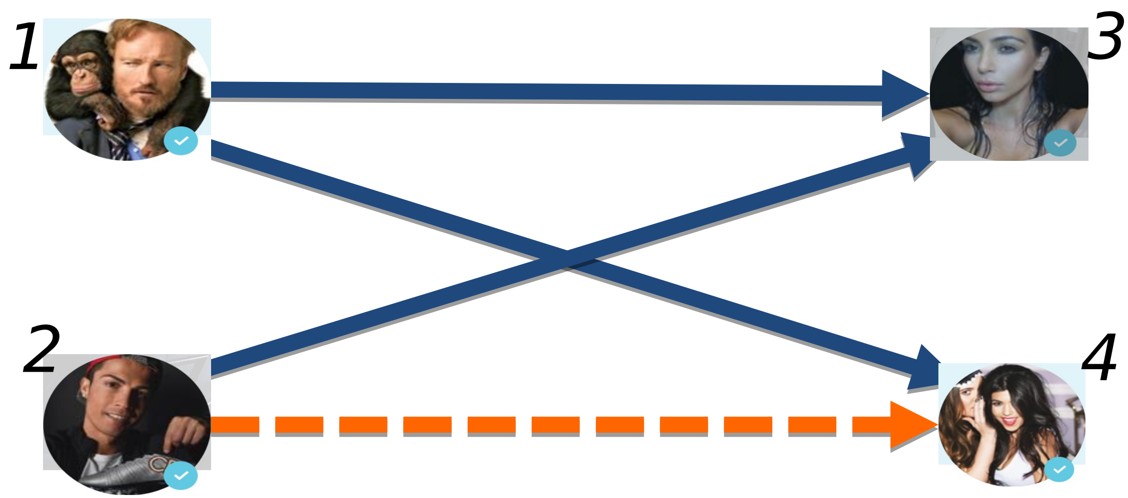


If Conan follows Kim and Kourtney,
and Cristiano follows Kim,
is Cristiano likely to follow Kourtney as well?

Forming dicliques



Diclique clustering



Theorem. *For $\gamma \sim \alpha m^{-1}$, and homogeneous attributes $(Z_k = 1)$,*

$$P_X(2 \to 4 \,|\, 1 \to 3, 1 \to 4, 2 \to 3) \;\sim\; \frac{1}{(1+\alpha X_1^+)(1+\alpha X_3^-)}.$$

Correlation is big if Conan demands and Kim supplies few attributes.

Diclique clustering – main result

Theorem. *For $m \gg 1$ and $\gamma \sim \alpha m^{-1}$,*

- $P_X\big(2 \to 4 \,\big|\, 1 \to 3,\, 1 \to 4,\, 2 \to 3\big)$

$$\approx \left(1 + \alpha(X_1^+ + X_3^-)\frac{(EZ_1^3)(EZ_1^2)}{EZ_1^4} + \alpha^2 X_1^+ X_3^- \frac{(EZ_1^2)^3}{EZ_1^4}\right)^{-1}.$$

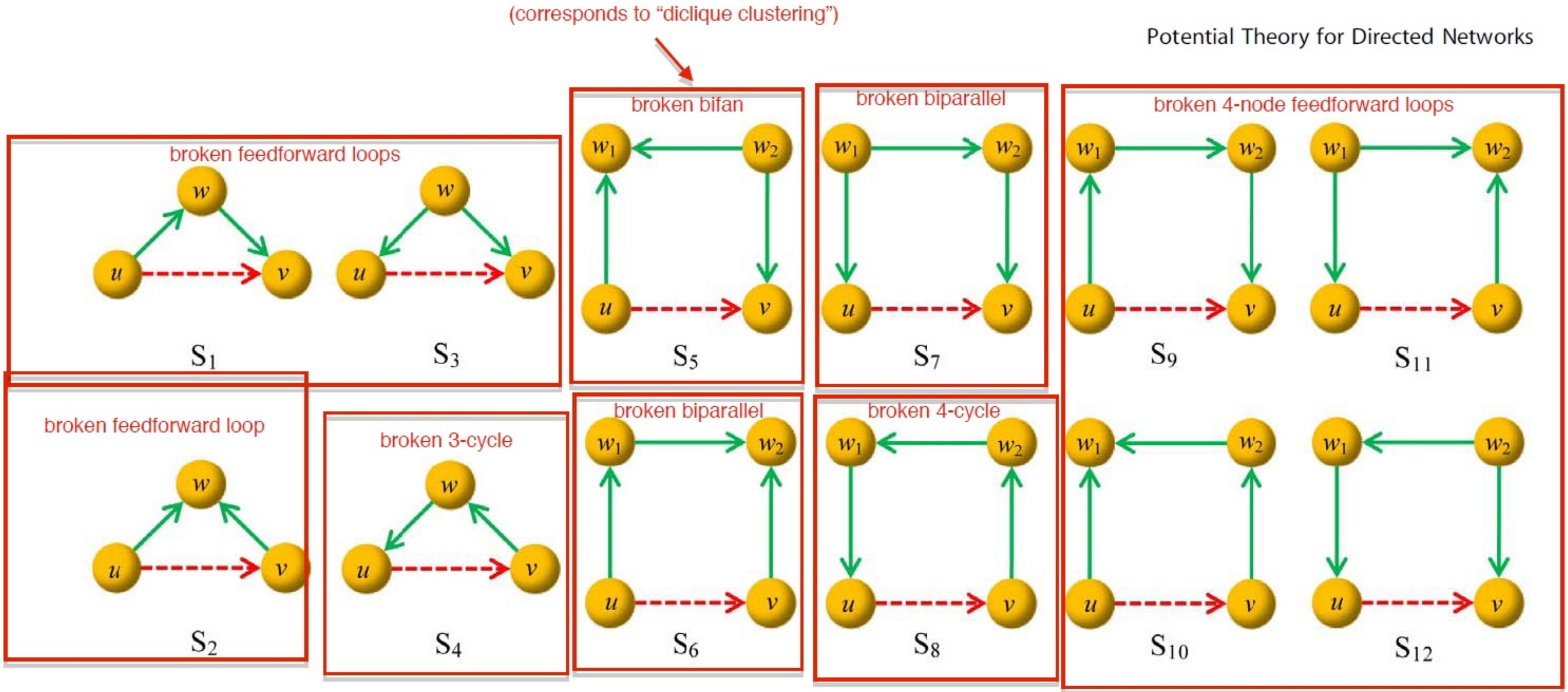
- $P_{X_3}\big(2 \to 4 \,\big|\, 1 \to 3,\, 1 \to 4,\, 2 \to 3\big)$

$$\approx \left(1 + \alpha\left(\frac{E(X_1^+)^2}{EX_1^+} + X_3^-\right)\frac{(EZ_1^3)(EZ_1^2)}{EZ_1^4} + \alpha^2 X_3^- \frac{E(X_1^+)^2}{EX_1^+}\frac{(EZ_1^2)^3}{EZ_1^4}\right)^{-1}.$$

- $P\big(2 \to 4 \,\big|\, 1 \to 3,\, 1 \to 4,\, 2 \to 3\big)$

$$\approx \left(1 + \alpha\left(\frac{E(X_1^+)^2}{EX_1^+} + \frac{E(X_1^-)^2}{EX_1^-}\right)\frac{(EZ_1^2)(EZ_1^3)}{EZ_1^4} + \alpha^2 \frac{E(X_1^+)^2}{EX_1^+}\frac{E(X_1^-)^2}{EX_1^-}\frac{(EZ_1^2)^3}{EZ_1^4}\right)^{-1}$$

12 directed clustering coefficients



QM Zhang, L Lü WQ Wang, YXi Zhu, T Zhou, PLoS ONE 2013.

Clustering in real directed networks

Table 1. AUC values of the 12 predictors shown in figure 5.

diclique clustering

Datasets	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}
FW1	0.7400	0.4634	0.6156	0.4903	**0.9066**	0.6147	0.7811	0.4172	0.7848	0.4254	0.3236	0.5697
FW2	0.7629	0.5507	0.6367	0.4809	**0.8964**	0.6965	0.7838	0.4972	0.6822	0.4255	0.3818	0.5456
FW3	0.7333	0.5364	0.5675	0.3997	**0.9105**	0.7282	0.7757	0.4303	0.6683	0.3517	0.3210	0.4532
C.elegans	0.7886	0.7127	0.7569	0.5671	**0.8679**	0.7686	0.7991	0.5755	0.7990	0.6528	0.6667	0.7591
SmaGri	0.7074	0.6517	0.6905	0.4922	**0.8852**	0.7108	0.7476	0.4851	0.6677	0.6242	0.5982	0.5761
Kohonen	0.6693	0.6124	0.6642	0.4991	**0.8605**	0.6333	0.7335	0.4985	0.6148	0.5614	0.5778	0.5946
SciMet	0.6462	0.6192	0.6371	0.4980	**0.8371**	0.6672	0.7045	0.4968	0.5977	0.5794	0.5753	0.5895
PB	0.9025	0.8181	0.8243	0.6948	**0.9595**	0.8659	0.8679	0.7518	0.9479	0.8349	0.7616	0.8584
Delicious	0.7298	0.7077	0.7192	0.6577	**0.7839**	0.7141	0.7344	0.6739	0.7378	0.7081	0.7046	0.7273
Youtube	0.7518	0.7453	0.7522	0.7456	0.8517	0.8422	0.8576	0.8442	0.8505	0.8430	0.8507	**0.8624**
FriendFeed	0.8801	0.7503	0.7382	0.5895	**0.9766**	0.7863	0.8100	0.7150	0.9690	0.8324	0.7318	0.8027
Epinions	0.8273	0.8326	0.8081	0.7460	**0.9101**	0.8969	0.8843	0.8584	0.8995	0.8956	0.8804	0.8831
Slashdot	0.7164	0.7133	0.7124	0.7072	**0.9035**	0.8984	0.8982	0.8925	0.9009	0.8982	0.8926	0.8985
Wikivote	0.9073	0.7448	0.7470	0.5962	**0.9699**	0.7679	0.7451	0.6209	0.9583	0.7562	0.6096	0.7468
Twitter	0.8937	0.7226	0.8289	0.7586	**0.9734**	0.7856	0.9444	0.7545	0.9582	0.8108	0.7557	0.9527
Average	0.7771	0.6787	0.7133	0.5949	**0.8995**	0.7584	0.8045	0.6341	0.8024	0.6800	0.6421	0.7213

QM Zhang, L Lü WQ Wang, YXi Zhu, T Zhou, PLoS ONE 2013.

Summary and conclusions

In undirected graphs

- Clustering coefficients measure transitivity (triplet closure):

“*Your friends are likely to be friends*”

- Most sparse random graphs have negligible transitivity
- Random intersection graphs form an exception

In directed graphs

- There are 4 different ways to define a triplet closure
- Most sparse random graphs have negligible triplet closure rates
- A prominent type of clustering in real graphs is *diclique clustering:*

“*Your followers are likely to follow common targets*”

- Directed random intersection graphs capture this phenomenon