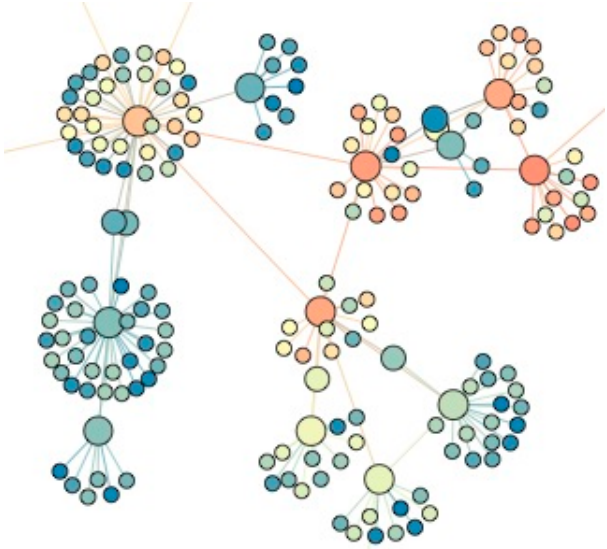


SNA Applications I

CS 224W

Why are you taking SNA?

- Your project will hopefully get you (at least) part way there
  - You want to understand aspects of a real-world network
  - You want to understand underlying mechanisms shaping the networks around us
  - You want to develop new algorithms that take advantage of the network features in the data

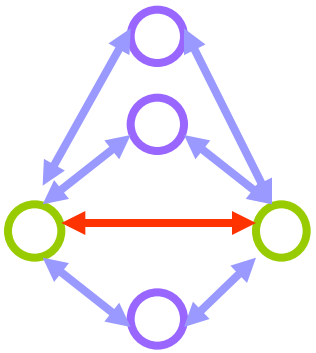
What makes a good SNA project?

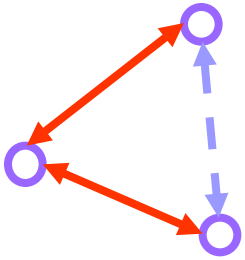
- Correlating network measurement with relevant (non-network) variables
- Assumptions about network formation yield predictions of observed structure
- Network features outperform others in prediction task

Tie strength & information

Are strong ties "local"?

- A strong tie
  - frequent contact
  - affinity
  - many mutual contacts





"forbidden triad": strong ties are likely to "close"

The strength of weak ties

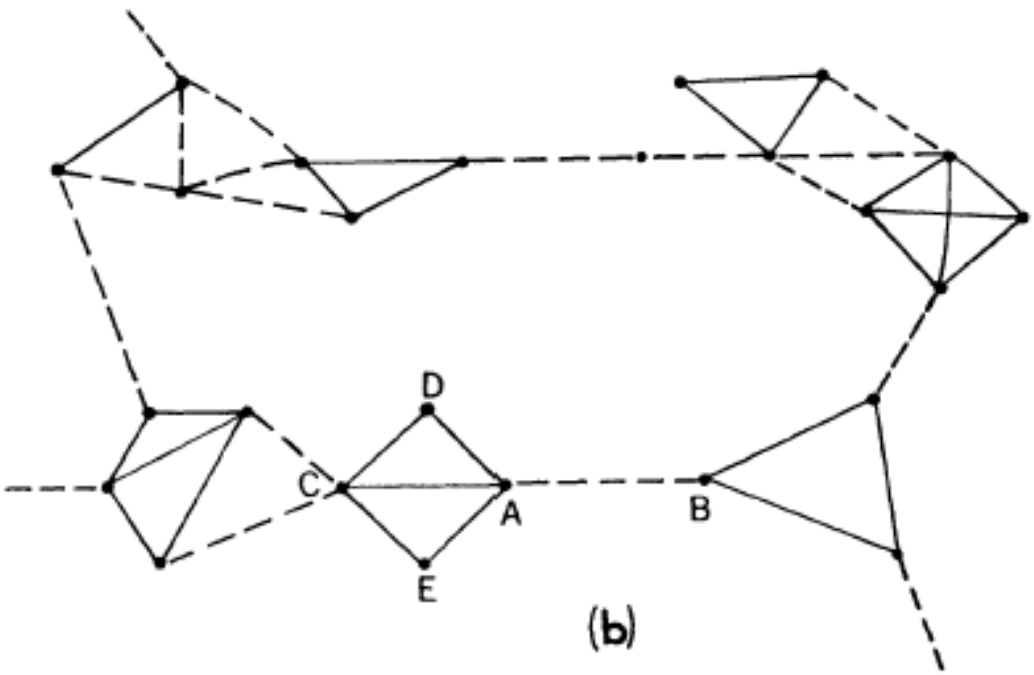
- M. S. Granovetter: *The Strength of Weak Ties*, AJS, 1973:

- finding a job through a contact that one saw
  - frequently (2+ times/week) 16.7%
  - occasionally (more than once a year but < 2x week) 55.6%
  - rarely 27.8%

- but… length of path is short
  - contact directly works for/is the employer
  - or is connected directly to employer

Bridges

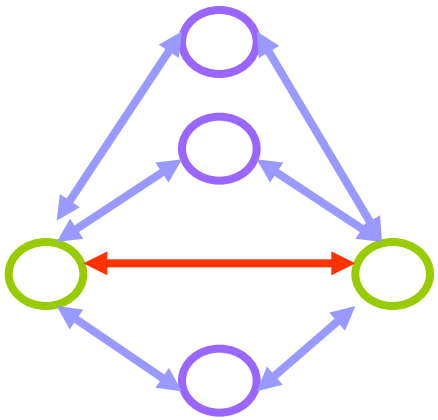
- Strong ties are less likely to be a bridge (or a local bridge)



Source: Granovetter, M. (1973). "The Strength of Weak Ties", American Journal of Sociology, Vol. 78, Issue 6, May 1973, pp. 1360-1380.

edge embeddeness

- embeddeness: number of common neighbors the two endpoints have

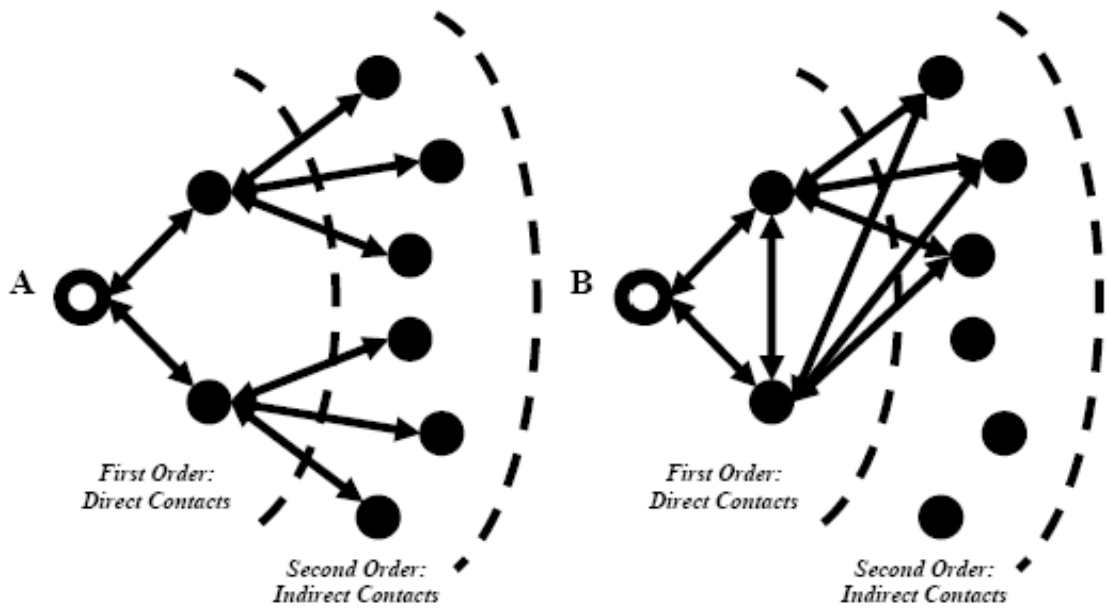


- neighborhood overlap:

$$\frac{\text{number of nodes who are neighbors of } \textit{both } A \text{ and } B}{\text{number of nodes who are neighbors of } \textit{at least one of } A \text{ or } B}$$

school kids and 1st through 8th choices of friends

- snowball sampling:
  - will you reach more different kids by asking each kid to name their 2 best friends, or their 7th & 8th closest friend?

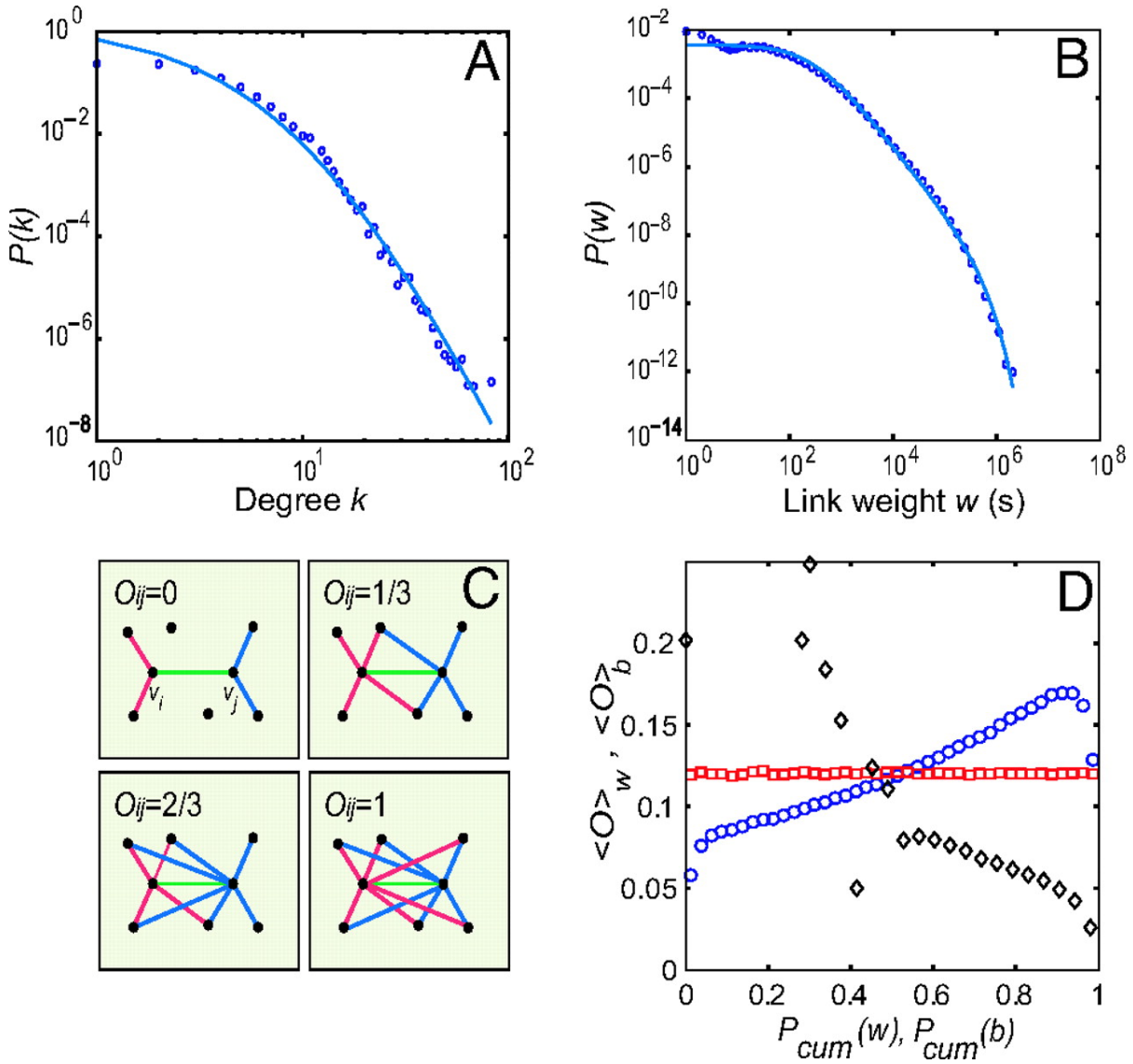


Source: M. van Alstyne, S. Aral. Networks, Information & Social Capital, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=958158

the strength of intermediate ties

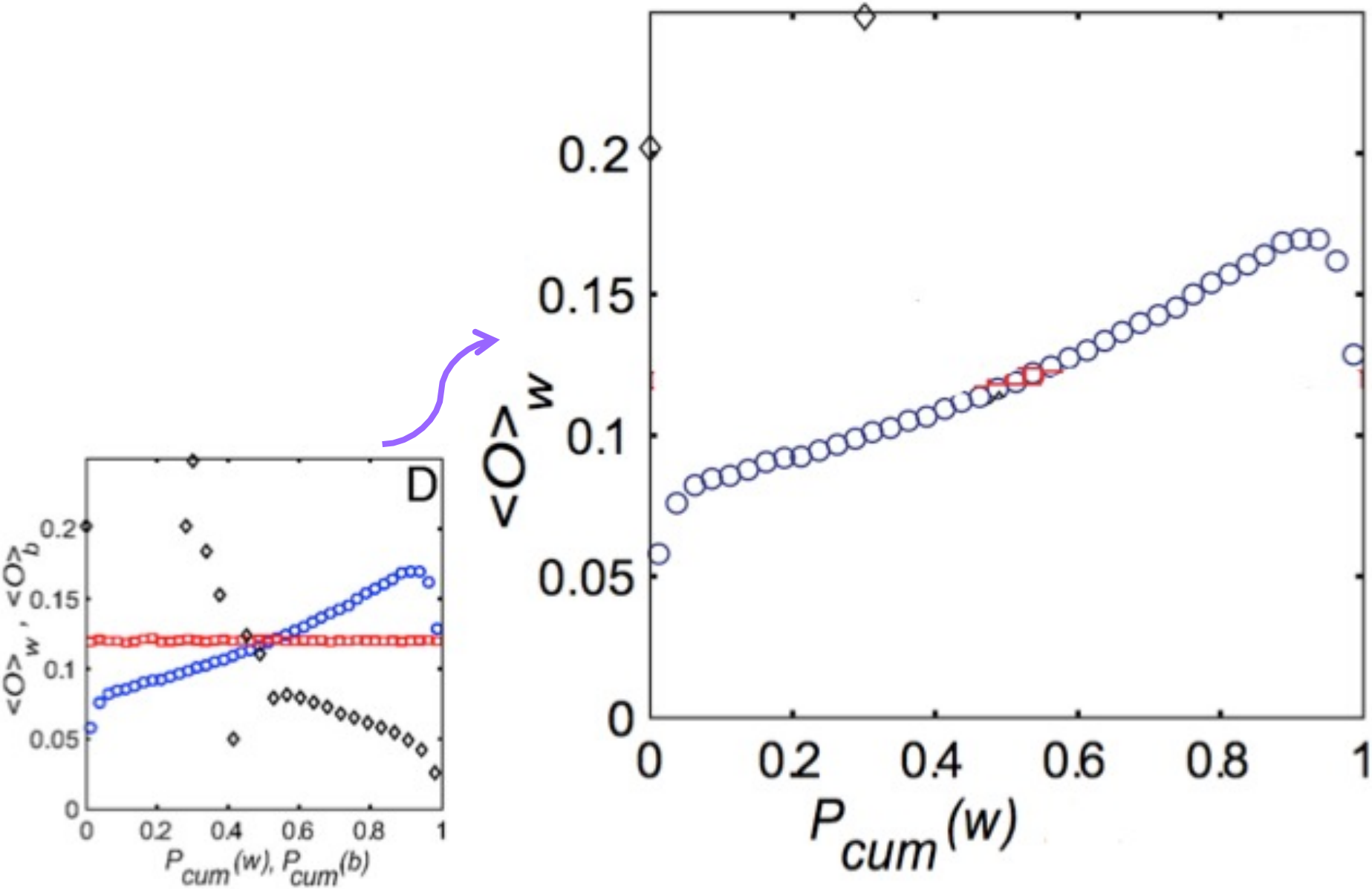
- strong ties
  - frequent communication, but ties are redundant due to high clustering
- weak ties
  - reach far across network, but communication is infrequent…
- Onnela J. et.al. PNAS 2007;104:7332-7336
  - use nation-wide cellphone call records and simulate diffusion using actual call timing
    - in simulation, individuals are most likely to obtain novel information through ties of intermediate strength

Characterizing the large-scale structure and the tie strengths of the mobile call graph

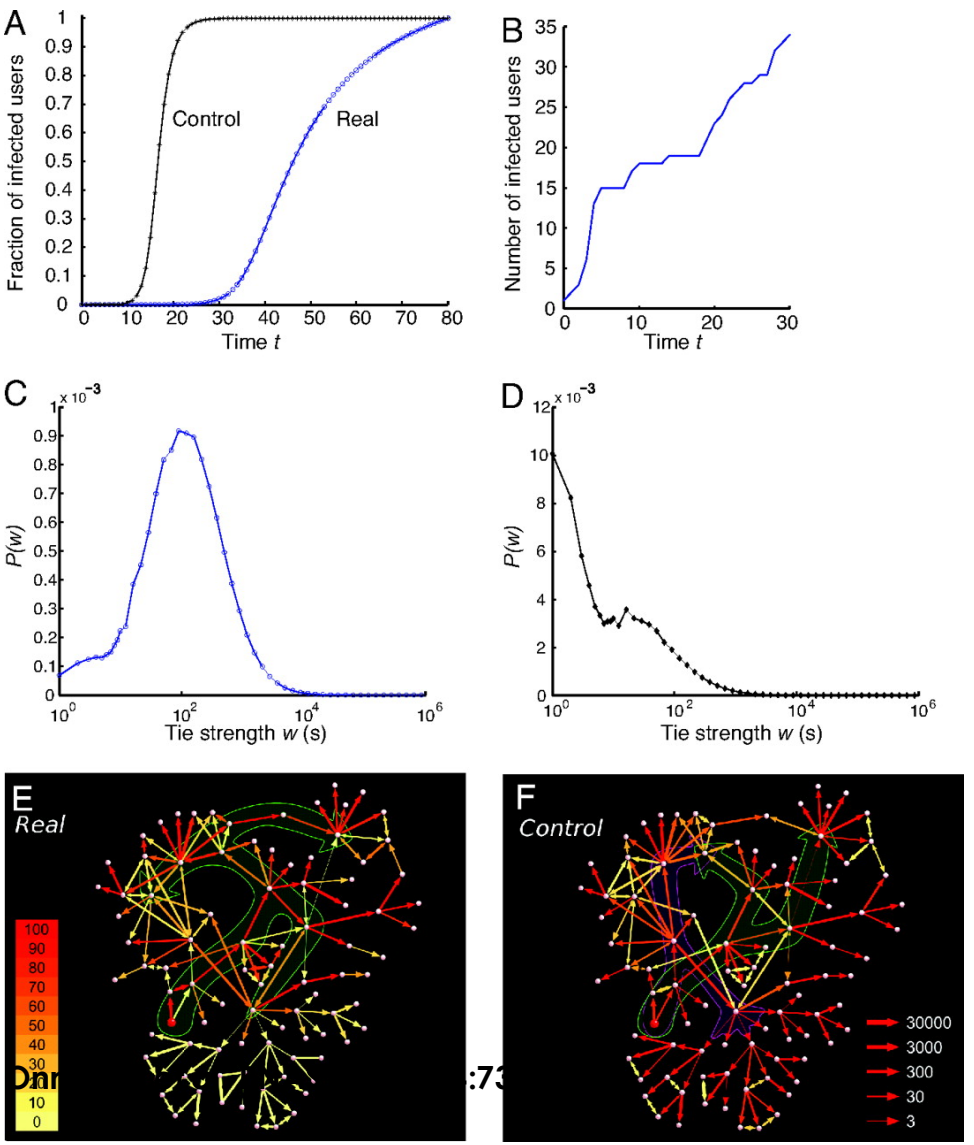


Onnela J et al. PNAS 2007;104:7332-7336

Edge neighborhood overlap as a function of tie strength



The dynamics of spreading on the weighted mobile call graph, assuming that the probability for a node v_i to pass on the information to its neighbor v_j in one time step is given by $P_{ij} = xw_{ij}$, with $x = 2.59 \times 10{-}4$



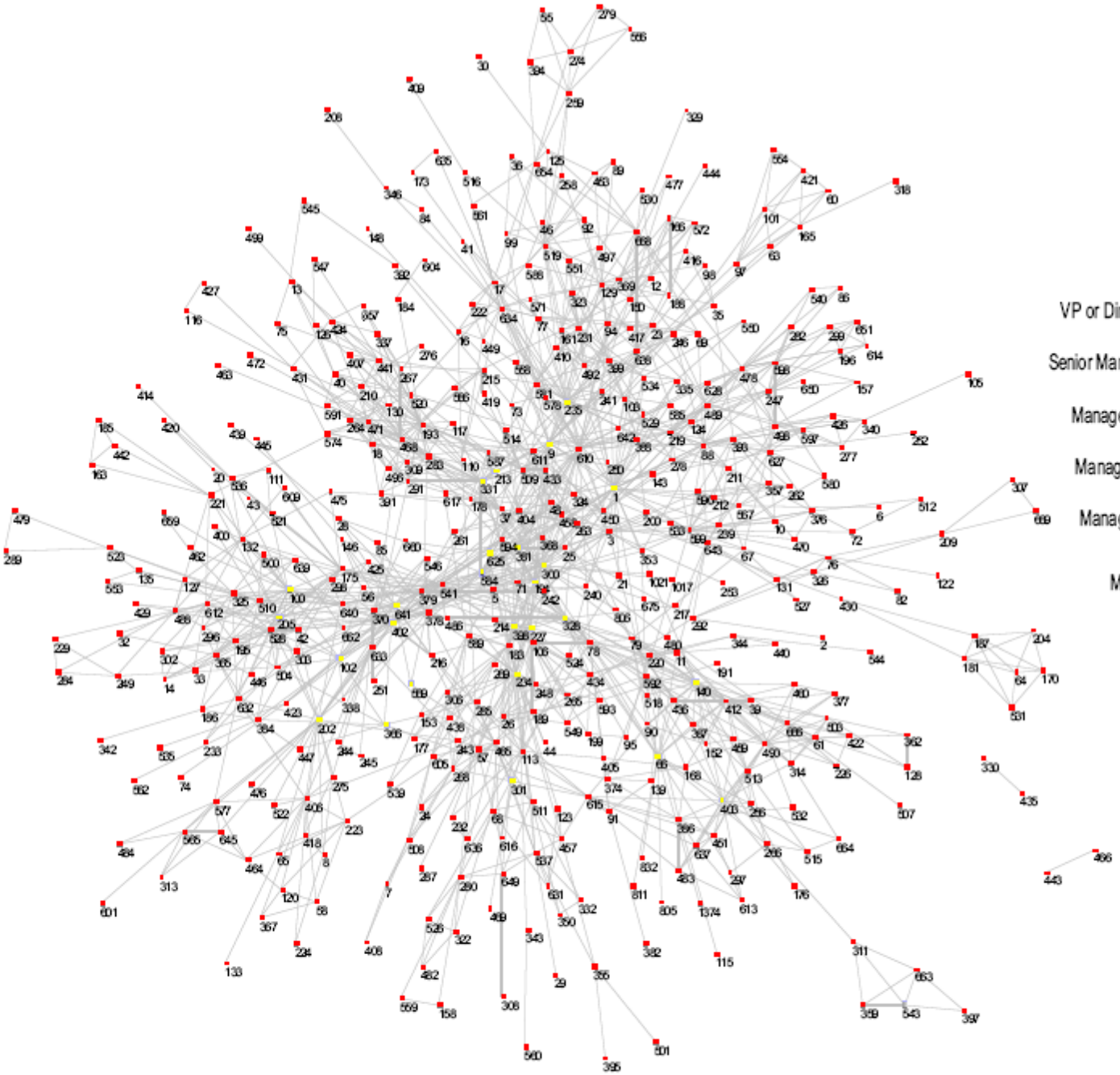
is it good to be embedded?

- What are the advantages of occupying an embedded position in the network?
- What are the disadvantages of being embedded?
- Advantages of being a broker (spanning structural holes)?
- Disadvantages of being a broker?

Burt: structural holes and good ideas

- Managers asked to come up with an idea to improve the supply chain
- Then asked:
  - whom did you discuss the idea with?
  - whom do you discuss supply-chain issues with in general
  - do those contacts discuss ideas with one another?

  - 673 managers (455 (68%) completed the survey)
  - ~ 4000 relationships (edges)



	Percent Social Isolates	Mean Network Size	Mean Network Constraint	Mean Number Cited as Discussion Partners	Mean Network Constraint Cited Discussn. Partners	Mean Path Distance (min-max) for the 476 connected managers in graph
VP or Director (25)	0%	12.6	29.8	4.9	70.2	3.3 (2.7-4.2)
Senior Manager (41)	5%	8.5	37.3	3.8	78.1	3.7 (2.9-6.4)
Manager III (121)	11%	6.4	50.2	3.7	77.9	4.0 (3.0-6.4)
Manager II (199)	27%	4.1	65.0	2.8	83.1	4.3 (2.8-6.4)
Manager I (287)	44%	3.4	73.6	2.4	83.4	4.6 (3.4-7.4)
Mean (673)	29%	5.0	60.5	2.9	81.0	4.2 (2.7-7.4)

Figure 2. Supply-Chain Discussion Network

(excludes 193 social isolates)

Source: Structural Holes and Good Ideas; R. Burt, American Journal of Sociology, 2004

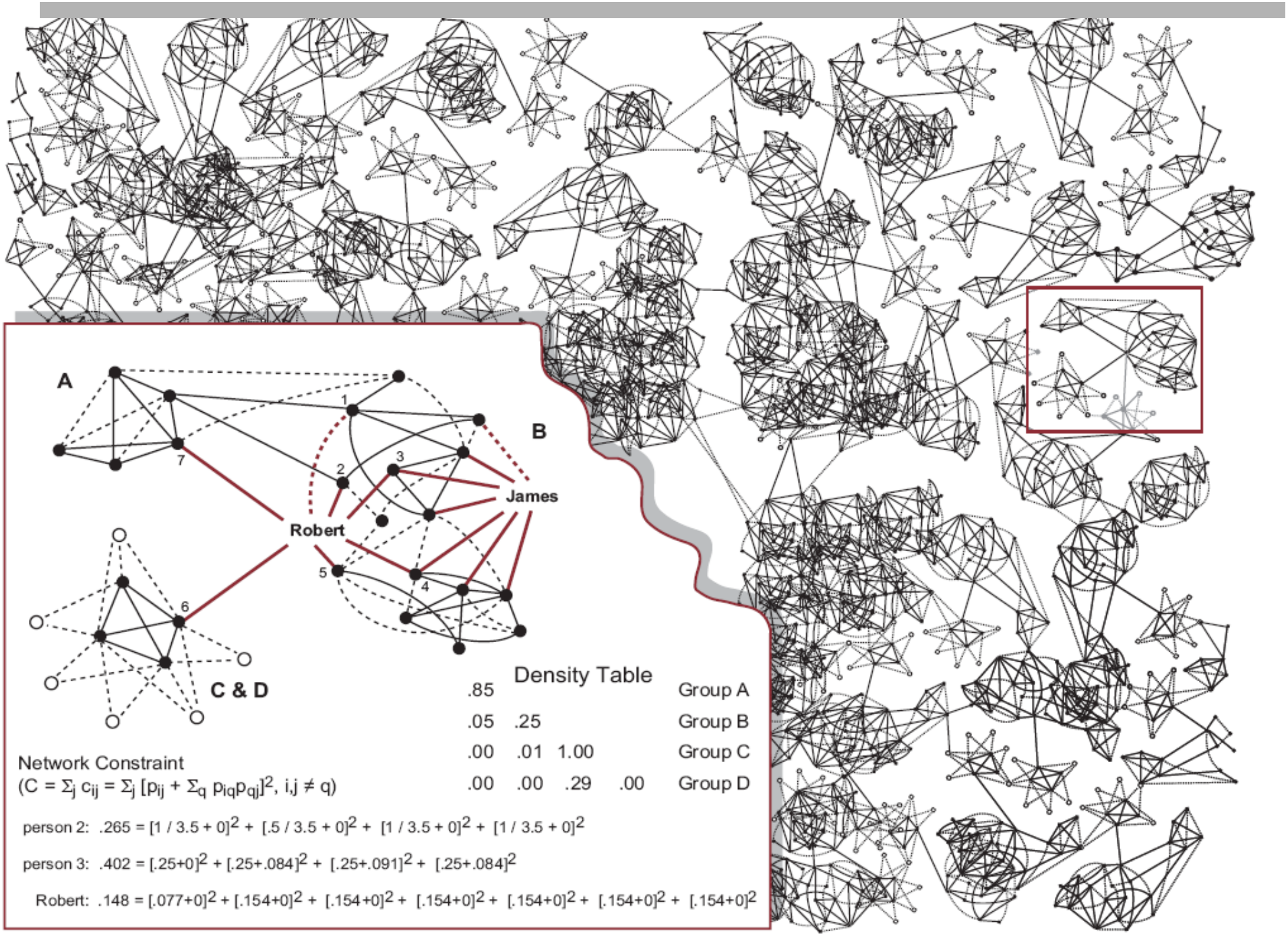


Figure 1. The Small World of Markets and Organizations

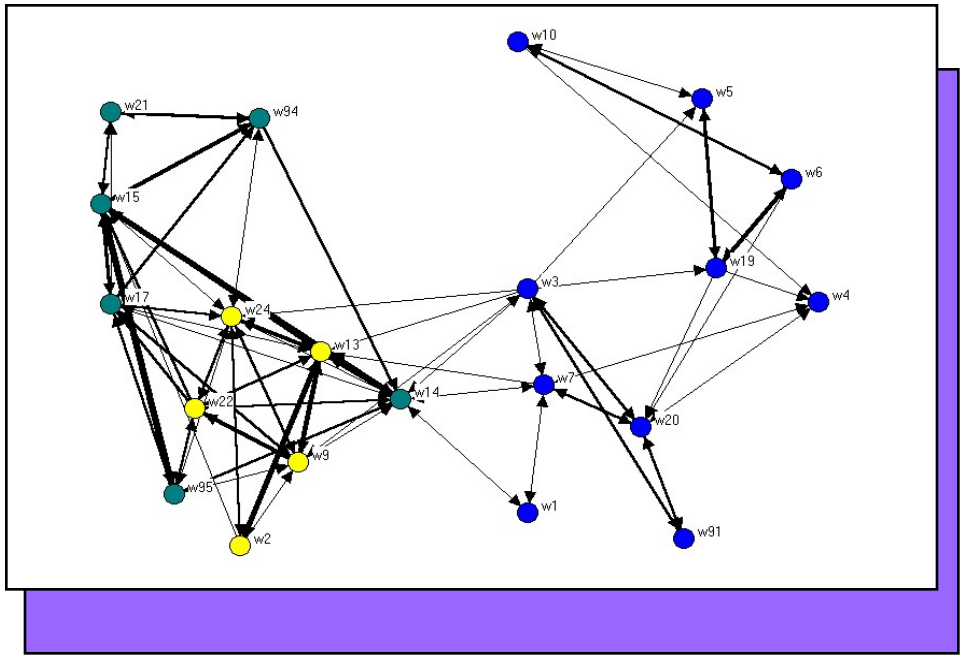
Source: Structural Holes and Good Ideas; R. Burt, American Journal of Sociology, 2004

results

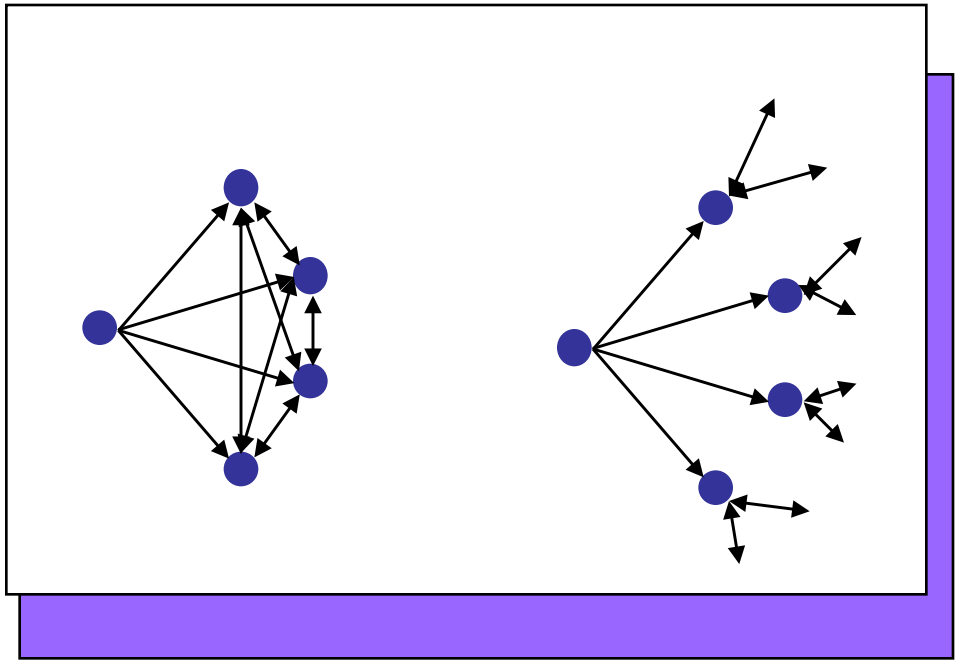
- people whose networks bridge structural holes have
  - higher compensation
  - positive performance evaluations
  - more promotions
  - more good ideas
- these brokers are
  - more likely to express ideas
  - less likely to have their ideas dismissed by judges
  - more likely to have their ideas evaluated as valuable

Position -> information -> $$?

Betweenness



Constrained vs. Unconstrained



Source: M. van Alstyne, S. Aral. Networks, Information & Social Capital (formerly titled 'Network Structure & Information Advantage'), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=958158

slide: Marshall van Alstyne

Study of a head hunter firm

- Three firms initially
- Unusually measurable inputs and outputs
  - 1300 projects over 5 yrs and
  - 125,000 email messages over 10 months (avg 20% of time!)
  - Metrics

    (i) Revenues per person and per project,
    (ii) number of completed projects,
    (iii) duration of projects,
    (iv) number of simultaneous projects,
    (v) compensation per person
- Main firm 71 people in executive search (+2 firms partial data)
  - 27 Partners, 29 Consultants, 13 Research, 2 IT staff
- Four Data Sets per firm
  - 52 Question Survey (86% response rate)
  - E-Mail
  - Accounting
  - 15 Semi-structured interviews

Source: M. van Alstyne, S. Aral. Networks, Information & Social Capital (formerly titled 'Network Structure & Information Advantage'), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=958158

slide: Marshall van Alstyne

Email structure matters

New Contract Revenue — **Coefficients[a]**

	Unstandardized Coefficients			
	B	Std. Error	Adj. R^2	Sig. F Δ
(Base Model)			0.40	
Best structural pred.	12604.0***	4454.0	0.52	.006
Ave. E-Mail Size	-10.7**	4.9	0.56	.042
Colleagues' Ave. Response Time	-198947.0	168968.0	0.56	.248

a. Dependent Variable: **Bookings02**
b. Base Model: YRS_EXP, PARTDUM, %_CEO_SRCH, SECTOR(dummies), %_SOLO.

Contract Execution Revenue — **Coefficients[a]**

	Unstandardized Coefficients			
	B	Std. Error	Adj. R^2	Sig. F Δ
(Base Model)			0.19	
Best structural pred.	1544.0**	639.0	0.30	.021
Ave. E-Mail Size	-9.3*	4.7	0.34	.095
Colleagues' Ave. Response Time	-368924.0**	157789.0	0.42	.026

a. Dependent Variable: **Billings02**
b. N=39. *** p<.01, ** p<.05, * p<.1

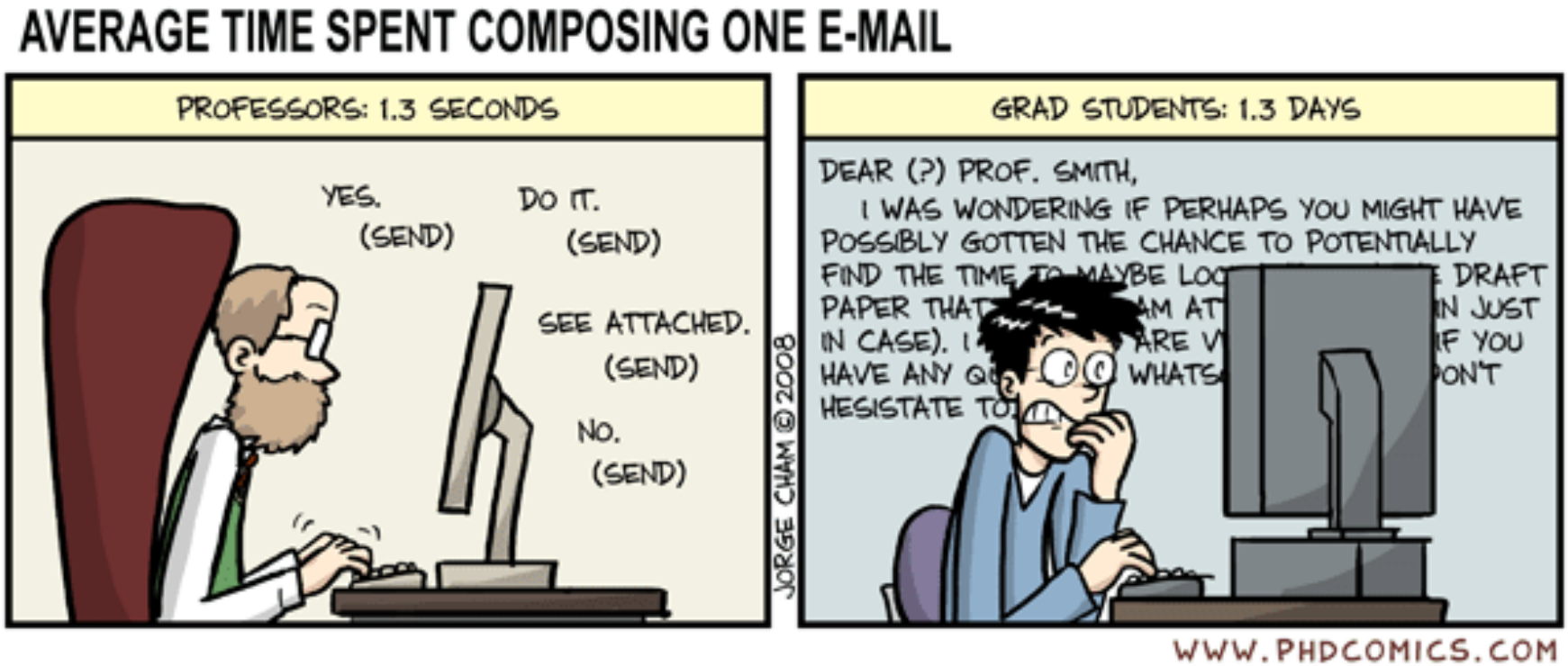
Sending *shorter* e-mail helps get contracts and finish them.

Faster response from colleagues helps finish them.

Source: M. van Alstyne, S. Aral. Networks, Information & Social Capital (formerly titled 'Network Structure & Information Advantage'), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=958158

slide: Marshall van Alstyne

Email structure



H5: Recruiters with larger personal rolodexes generate no more or less output

	Revenue $	$ for completed searches	Completed searches	Multitasking	Duration	Duration controlling for multitasking
Size of rolodex (Q50)	-10.2 (60.3)	-22.9 (32.6)	0.000 (0.001)	0.000 (0.001)	-0.013 (0.021)	-0.013 (0.016)

*** p < 0.10, ** p < 0.05, *** p < 0.01, Standard err in paren.**

Instead, a larger private rolodex is associated with:

- Less information sharing
- Less DB proficiency
- Lower % of e-mail read
- Less learning from others
- Less perceived credit for ideas given to colleagues
- More dissembling on the phone

Source: M. van Alstyne, S. Aral. Networks, Information & Social Capital (formerly titled 'Network Structure & Information Advantage'), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=958158

slide: Marshall van Alstyne

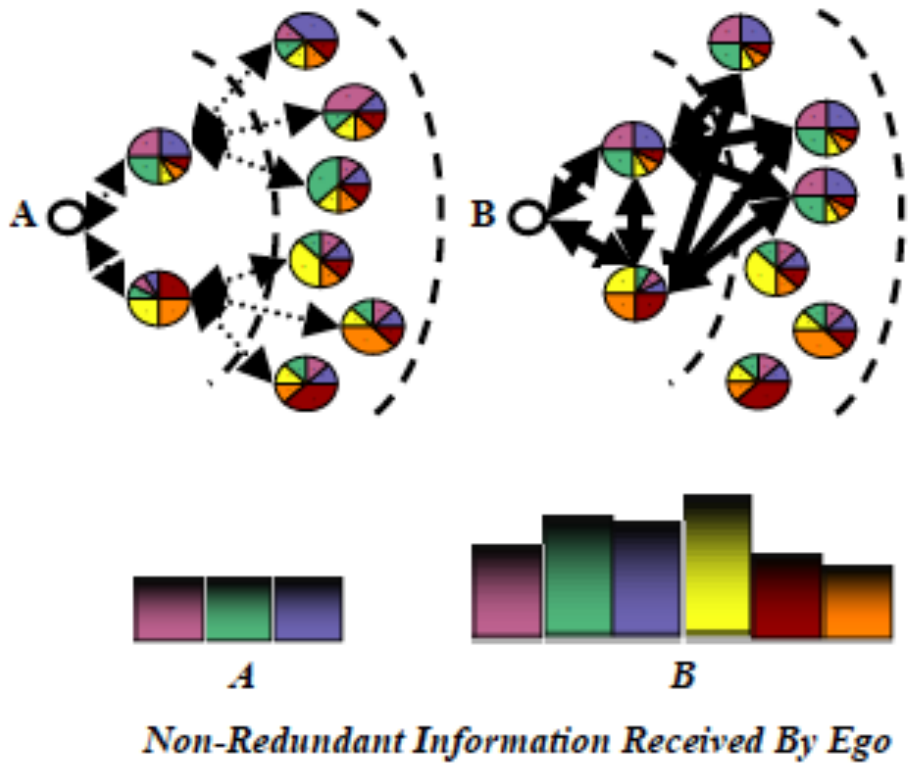
Information access

- network structure (having high degree) correlates with receiving novel information sooner (as deduced from hashed versions of their email)
- getting information sooner correlates with $$ brought in
  - controlling for # of years worked
  - job level
  - ....

Source: M. van Alstyne, S. Aral. Networks, Information & Social Capital, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=958158

Diversity/Bandwidth trade-off

- Strong ties may be more redundant, but they also share information more frequently, and are willing to share novel information



Source: M. van Alstyne, S. Aral. Networks, Information & Social Capital, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=958158

Network Structure Matters

New Contract Revenue — **Coefficients[a]**

	Unstandardized Coefficients			
	B	Std. Error	Adj. R^2	Sig. F Δ
(Base Model)			0.40	
Size Struct. Holes	13770***	4647	0.52	.006
Betweenness	1297*	773	0.47	.040

a. Dependent Variable: **Bookings02**
b. Base Model: YRS_EXP, PARTDUM, %_CEO_SRCH, SECTOR(dummies), %_SOLO.

Contract Execution Revenue — **Coefficients[a]**

	Unstandardized Coefficients			
	B	Std. Error	Adj. R^2	Sig. F Δ
(Base Model)			0.19	
Size Struct. Holes	7890*	4656	0.24	.100
Betweenness	1696**	697	0.30	.021

a. Dependent Variable: **Billings02**
b. N=39. *** p<.01, ** p<.05, * p<.1

Bridging diverse communities is significant.

Being in the thick of information flows is significant.

Source: M. van Alstyne, S. Aral. Networks, Information & Social Capital (formerly titled 'Network Structure & Information Advantage'), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=958158

slide: Marshall van Alstyne

Information backbone

- Kossinets, Watts, Kleinberg, KDD 2008:
  - which paths yield the most up to date info?
  - how many of the edges form the "backbone"?

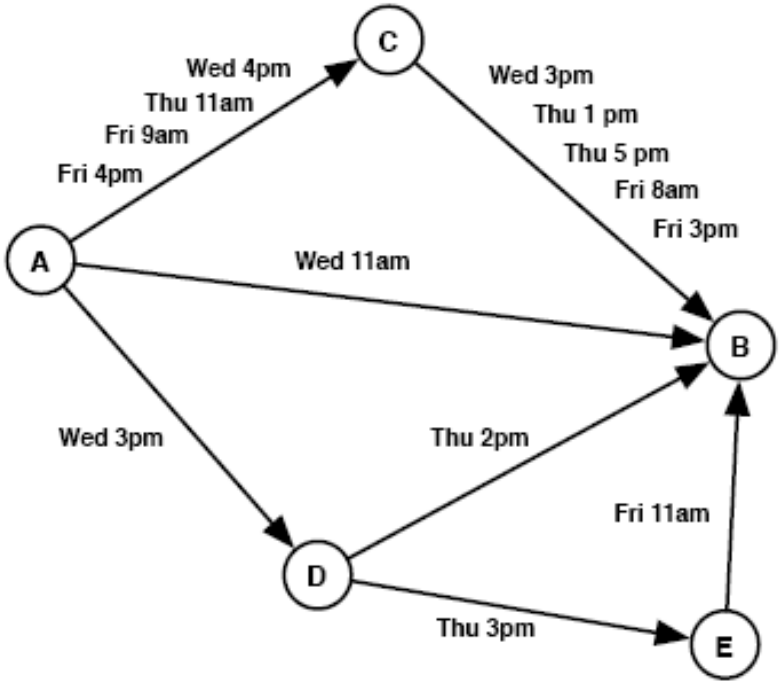
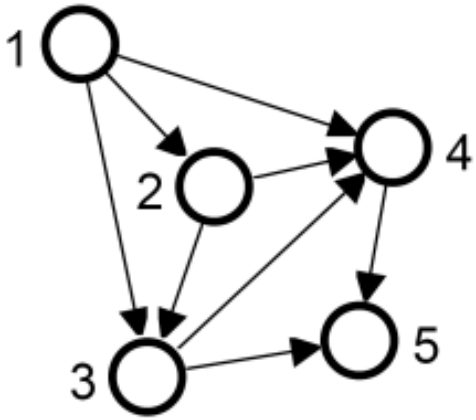


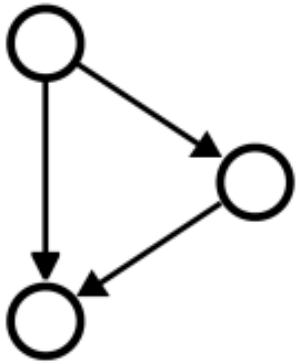
image source: Kossinets et al. "The structure of information pathways in a social communication network", KDD 2008

Motifs

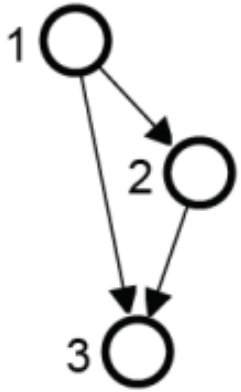
Resolving local structure: network motifs



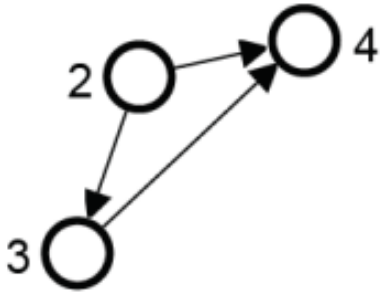
graph



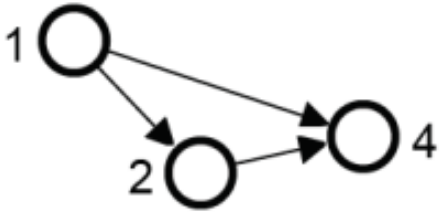
motif to be found



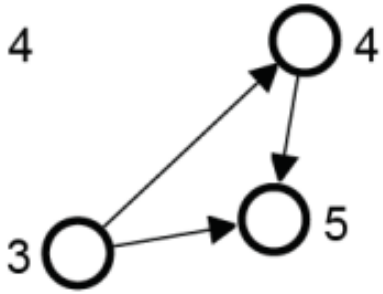
M1



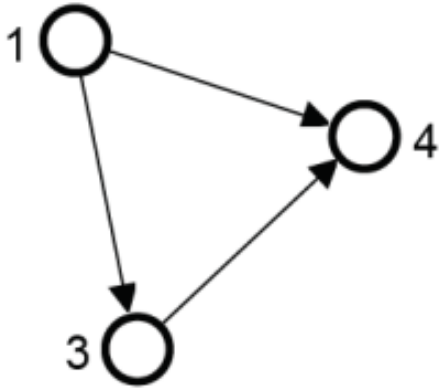
M2



M3



M4

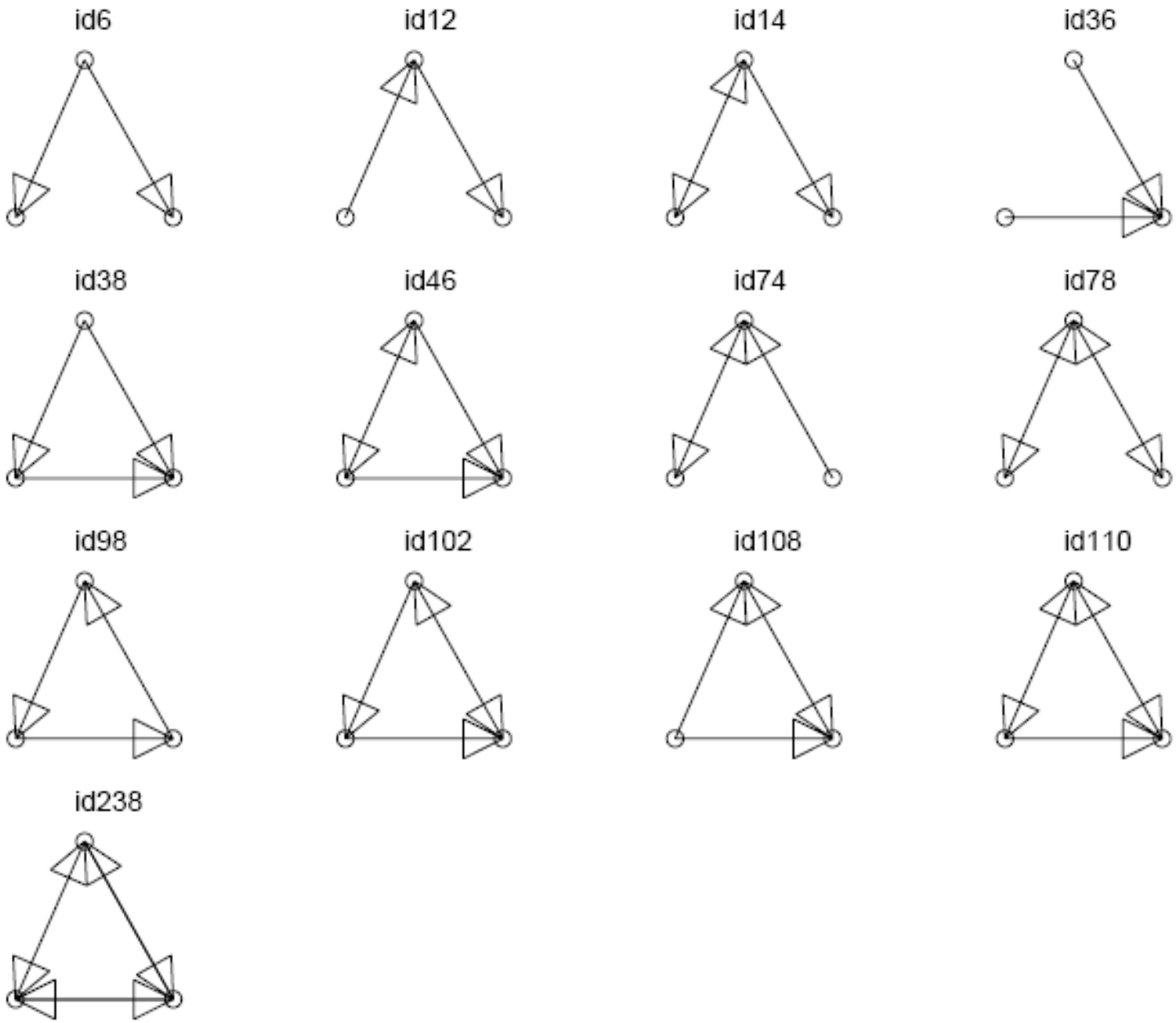


M5

motif matches in the target graph

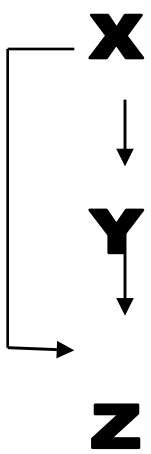
http://mavisto.ipk-gatersleben.de/frequency_concepts.html

All 3 node motifs

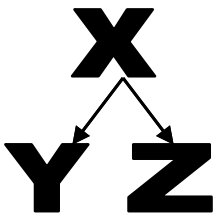


Examples of network motifs (3 nodes)

- Feed forward loop
  - Found in neural networks
  - Seems to be used to neutralize "biological noise"

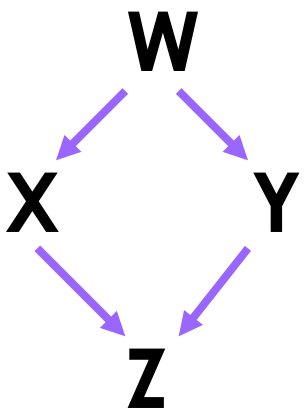


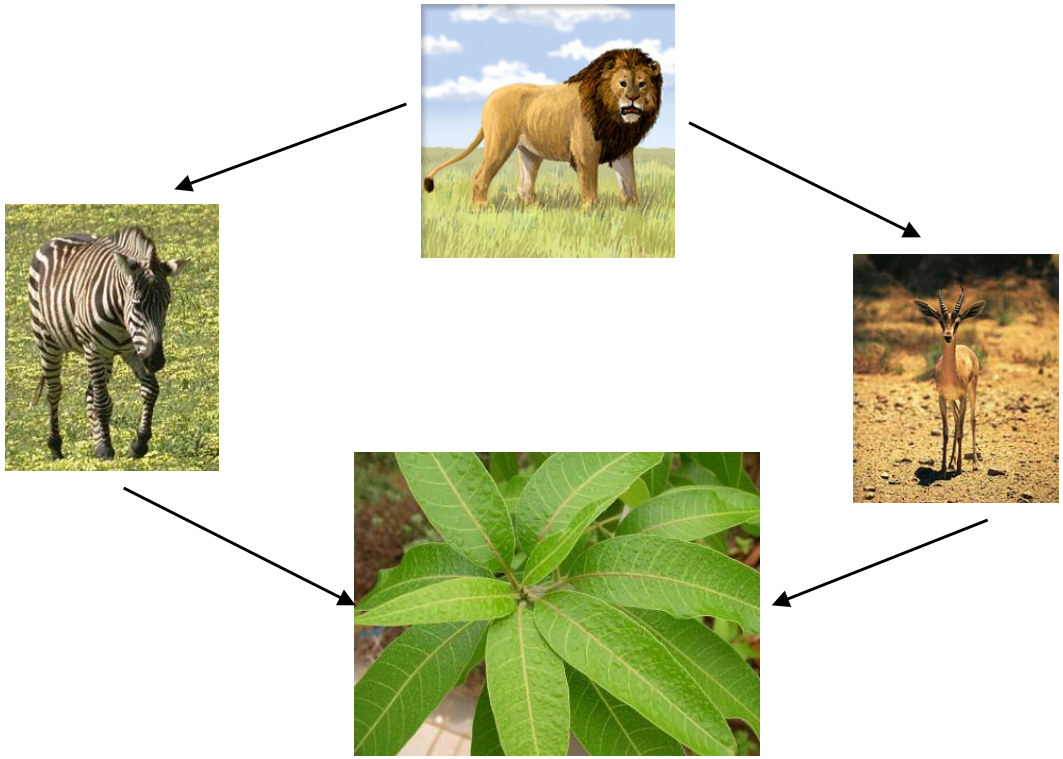
- Single-Input Module
  - e.g. gene control networks



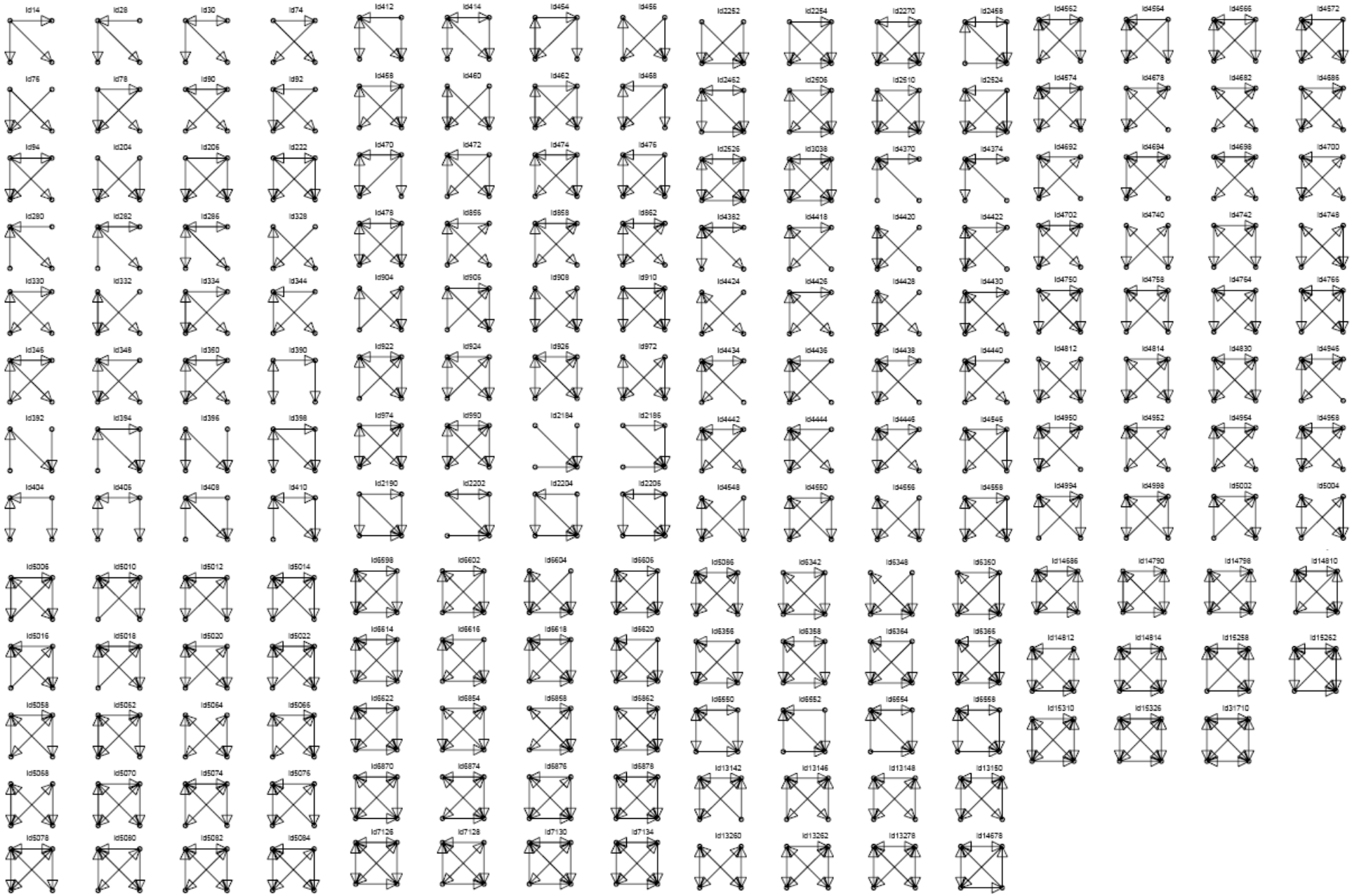
Examples of network motifs (4 nodes)

- Parallel paths
  - Found in neural networks
  - Food webs

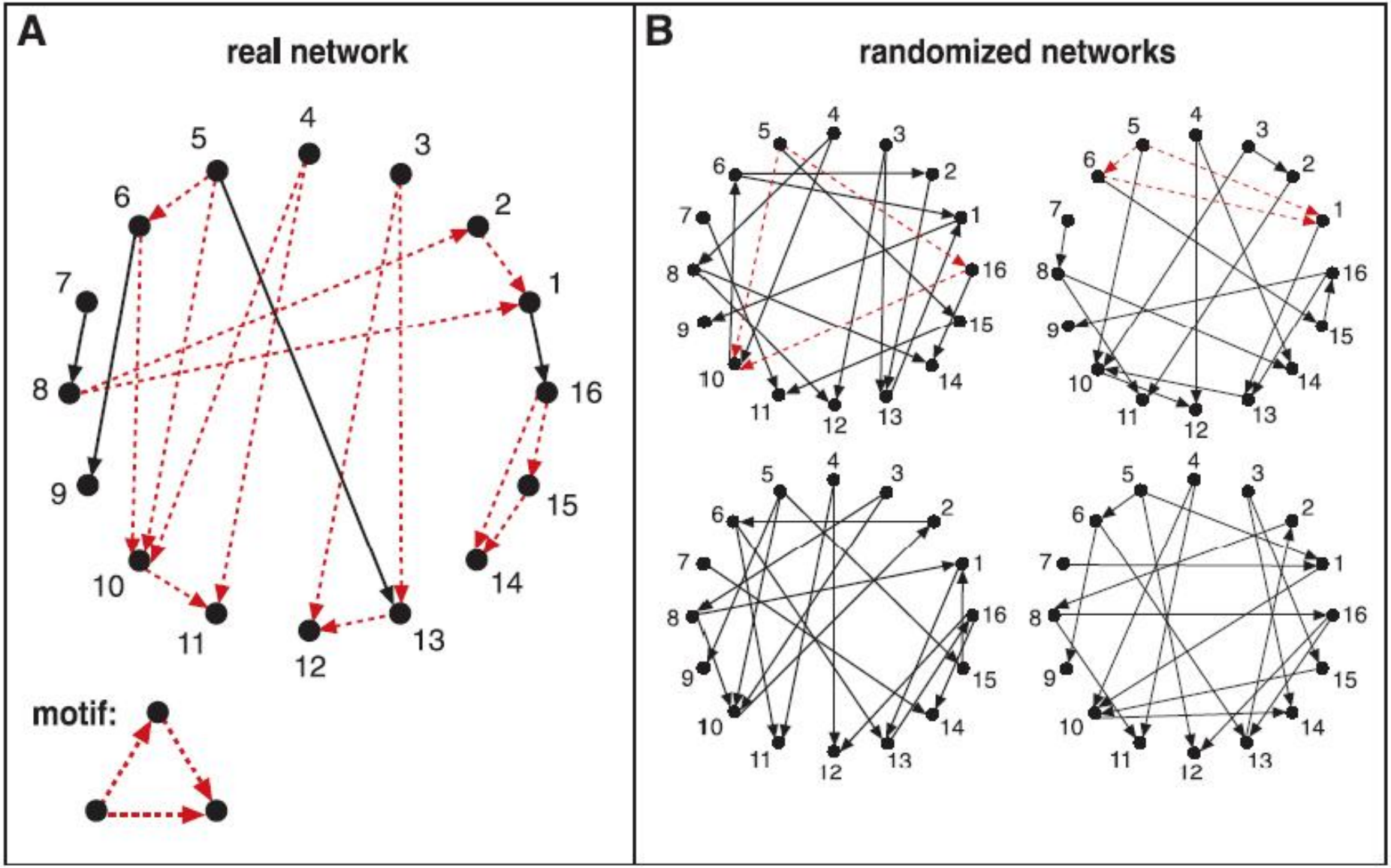




4 node subgraphs (computational expense increases with the size of the graph!)



Compare to “equivalent” random graph



Milo et al., Network motifs: Simple building blocks of complex networks, Science 298:824-827, 2002

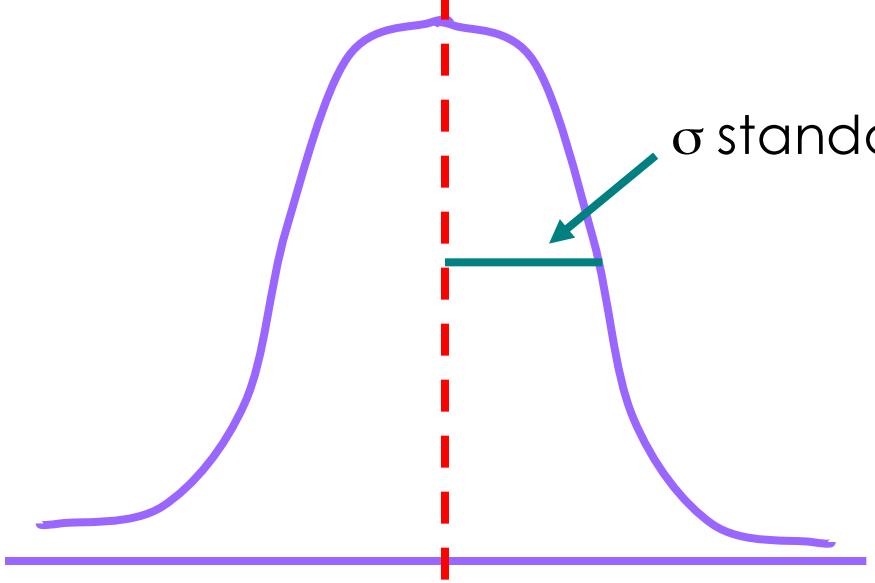
Network motif detection

- Some motifs will occur more often in real world networks than random networks
- Technique:
  - construct many random graphs with the same number of nodes and edges (same node degree distribution?)
  - count the number of motifs in those graphs
  - calculate the Z score: the probability that the given number of motifs in the real world network could have occurred by chance
- Software available:
  - http://www.weizmann.ac.il/mcb/UriAlon/ (the original)
  - FanMod http://theinf1.informatik.uni-jena.de/~wernicke/motifs/index.html

  (faster and more user friendly)

What the Z score means

μ = mean number of times the motif appeared in the random graph



# of times motif appeared in random graph

the probability observing a Z score of 2 is 0.02275

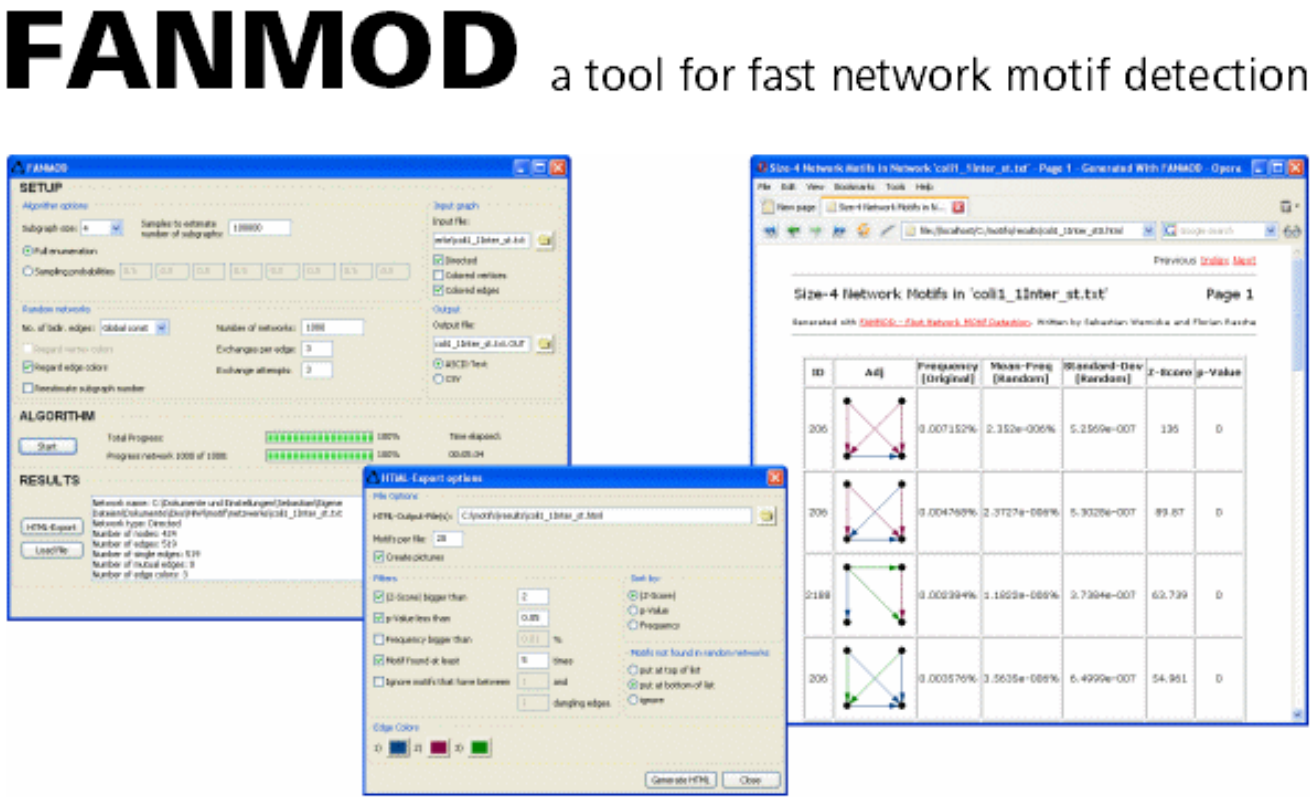
In the context of motifs:
Z > 0, motif occurs more often than for random graphs
Z < 0, motif occurs less often than in random graphs

$$Z_x = \frac{x - \mu_x}{\sigma_x}$$

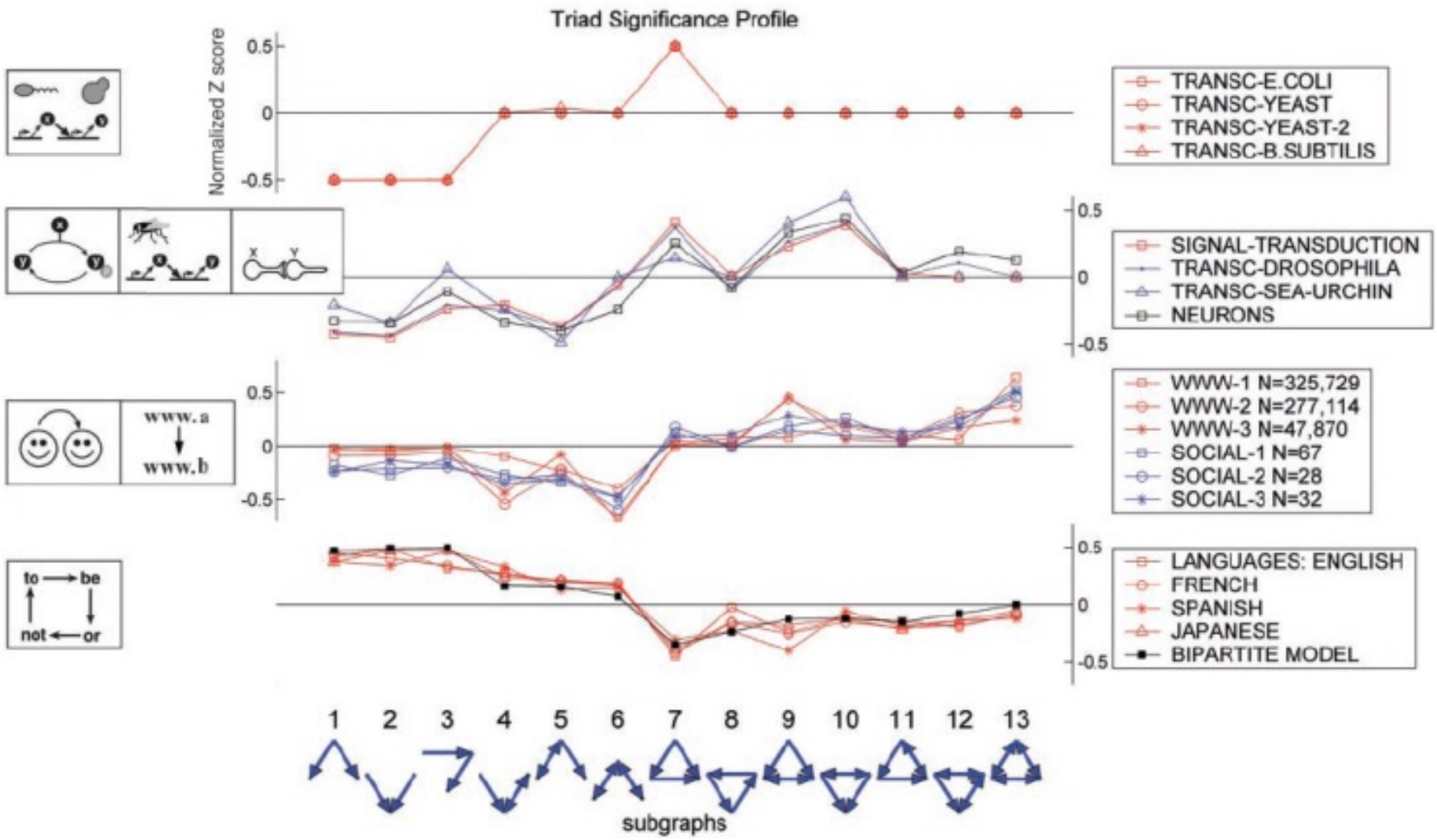
|Z| > 1.65, only a 5% chance of random occurence

software: FANMOD (also igraph)

- http://theinf1.informatik.uni-jena.de/~wernicke/motifs/index.html



Superfamilies of networks

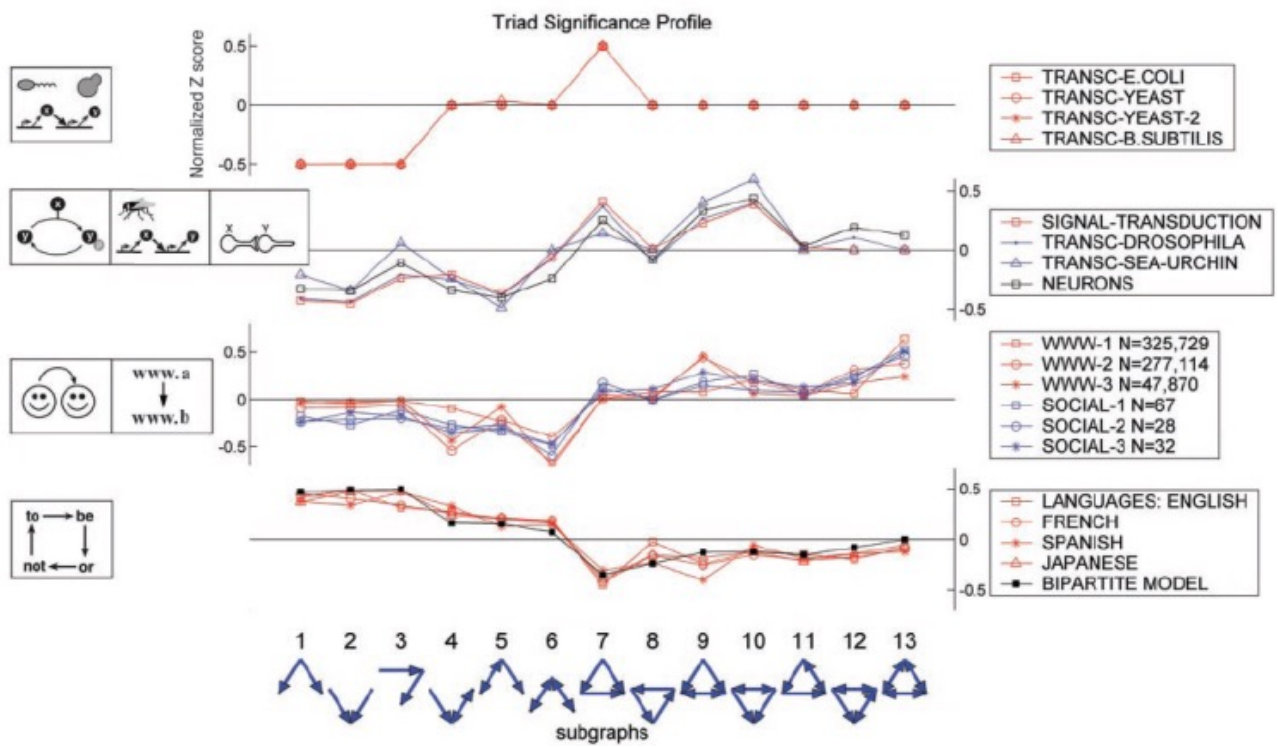


source: Milo et al., Superfamilies of Evolved and Designed Networks, Science 303:1538-1542, 2004

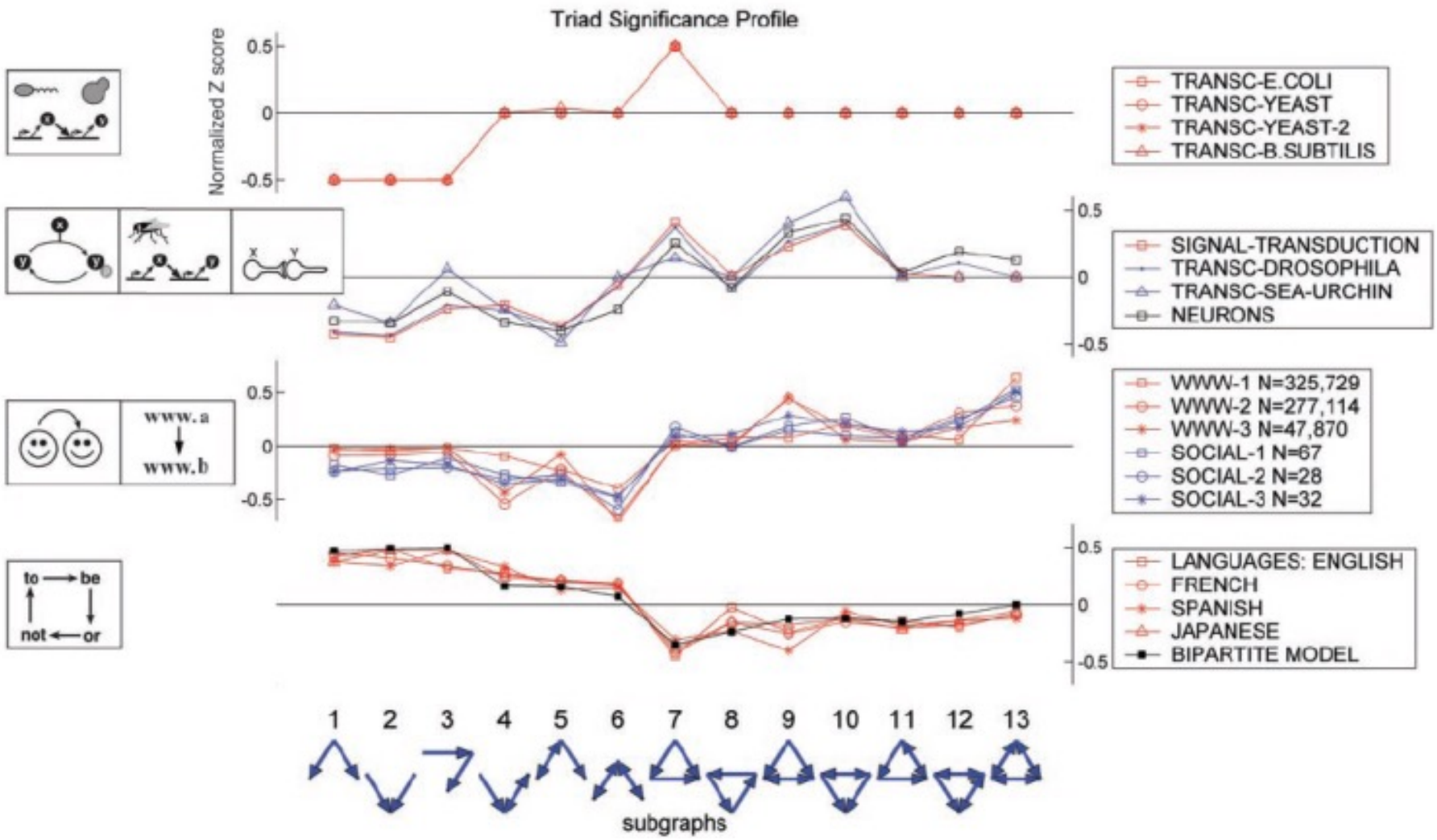
Quiz Q:

- Based on their triad census profiles, which two kinds of networks exhibit similar structure?

(a) Transcription and language
(b) WWW and language
(c) Social and WWW
(d) Language and social



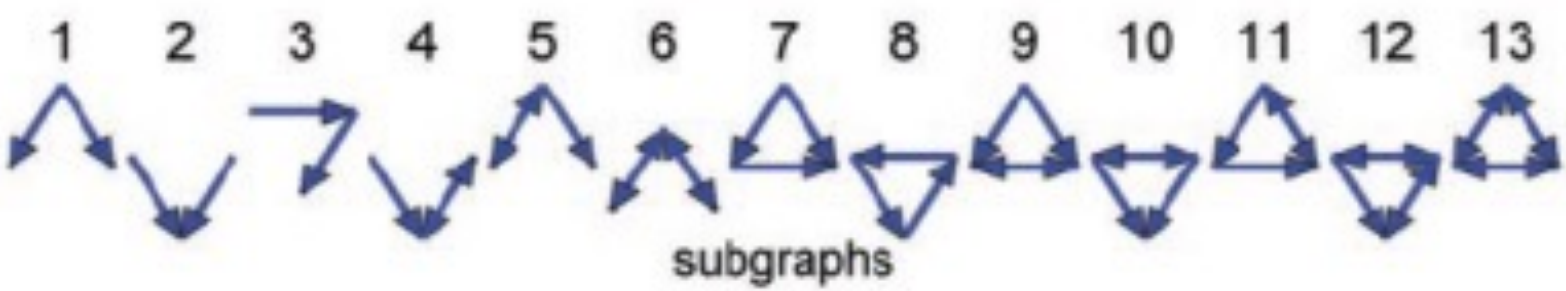
Superfamilies of networks



source: Milo et al., Superfamilies of Evolved and Designed Networks, Science 303:1538-1542, 2004

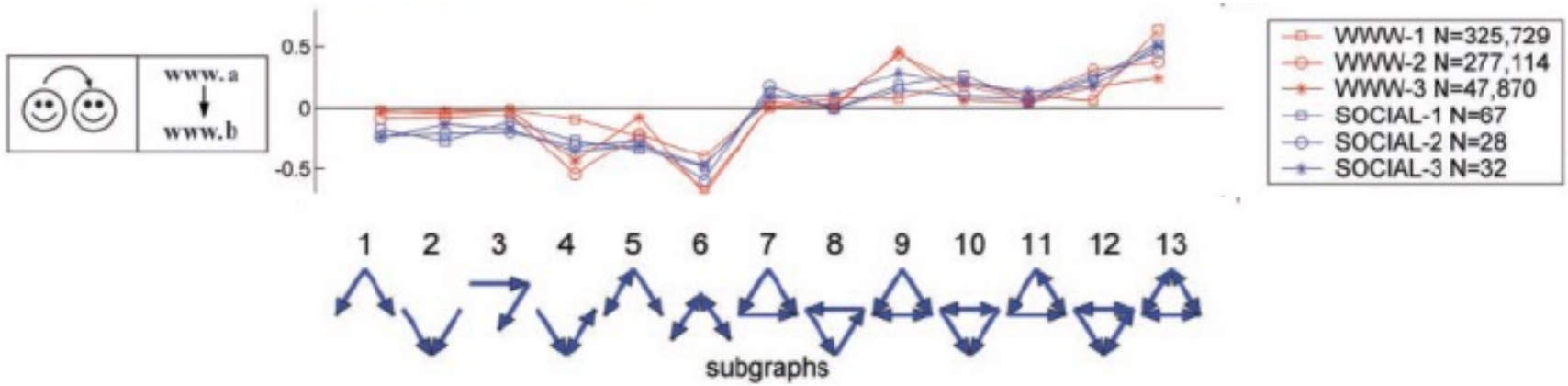
Quiz Q:

- Which of the following triads is underrepresented in social networks?



(a) 6 (b) 9 (c) 12 (d) 13

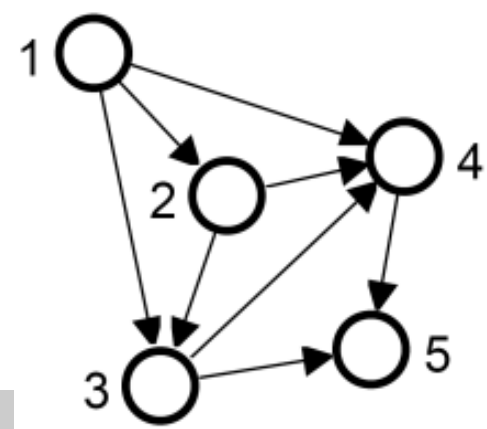
Superfamilies of networks



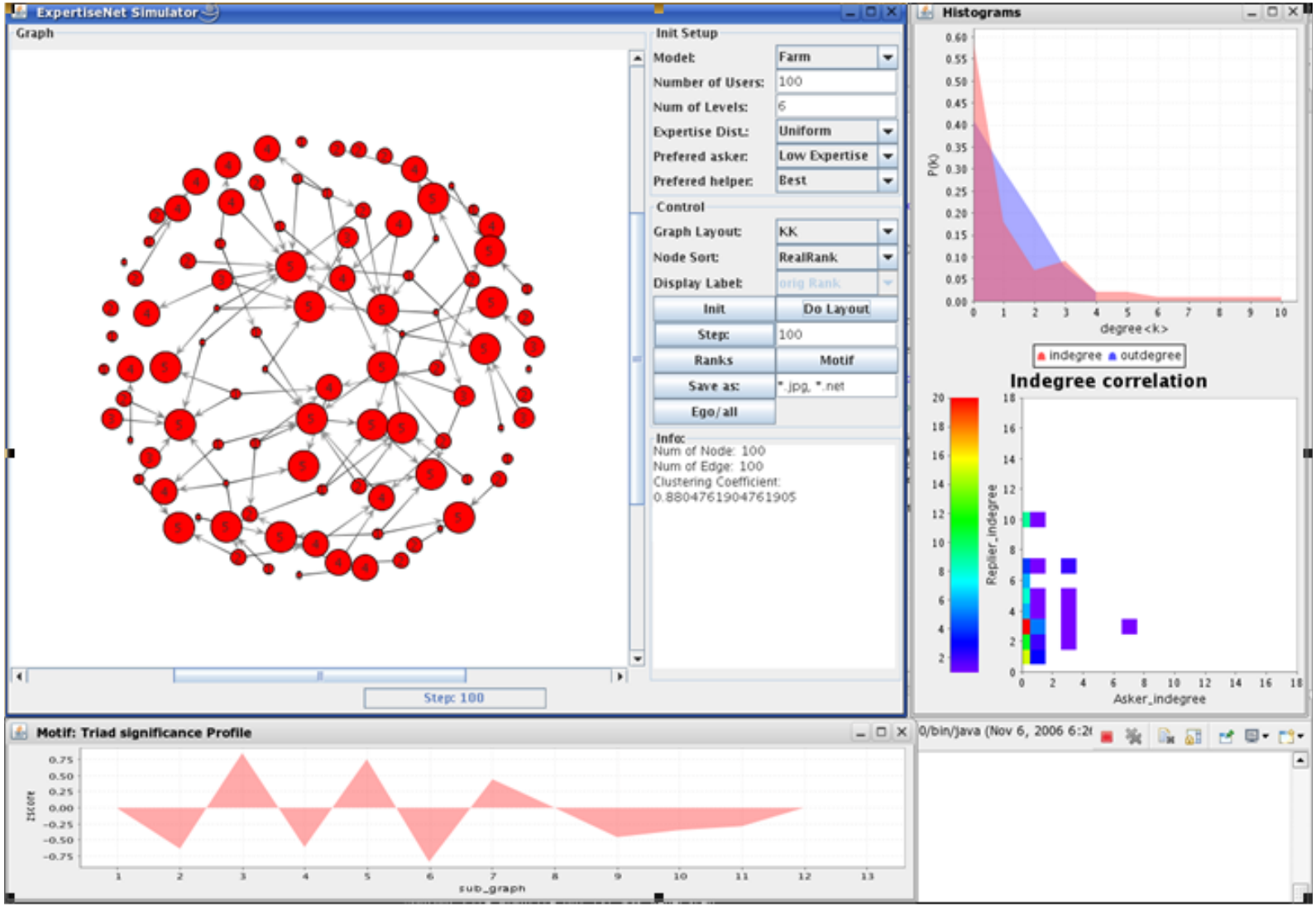
source: Milo et al., Superfamilies of Evolved and Designed Networks, Science 303:1538-1542, 2004

Motifs: recap

- Given a particular structure, search for it in the network, e.g. complete triads
- advantage: motifs can correspond to particular functions, e.g. in biological networks
- disadvantage: don't know if motif is part of a larger cohesive community



Modeling community structure in a Q&A forum



Zhang, Ackerman, Adamic, WWW'07

simulating probability of expertise pairing

suppose:

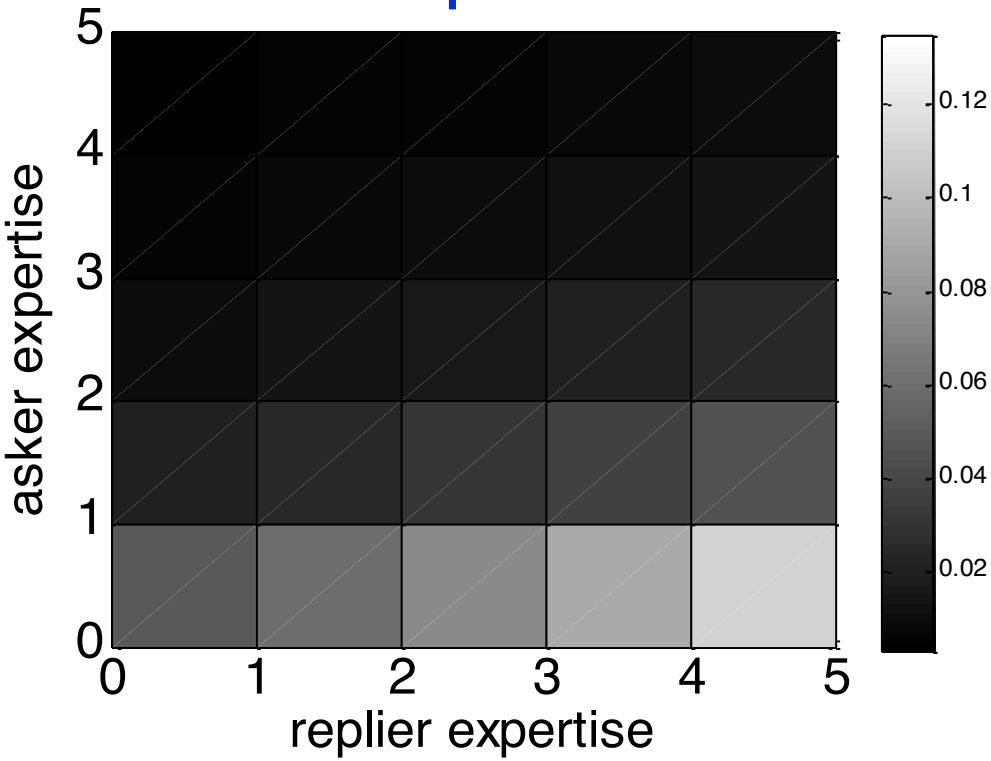
expertise is uniformly distributed

probability of posing a question is inversely proportional to expertise

p_{ij} = probability a user with expertise *j* replies to a user with expertise

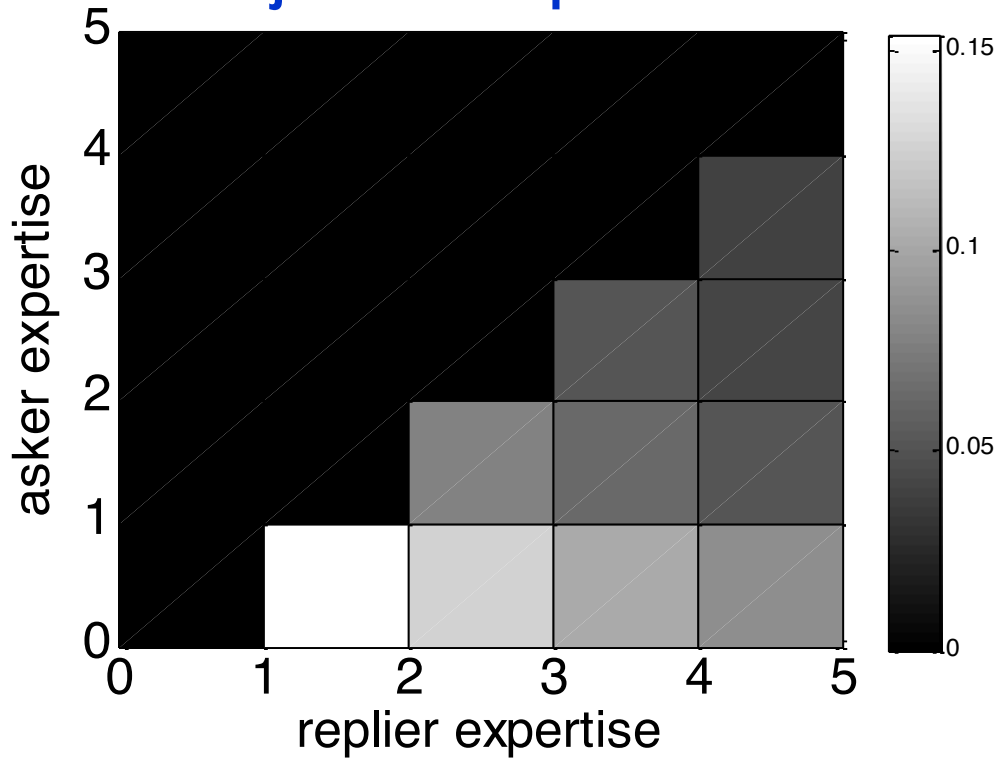
2 models:

'best' preferred



$$p_{ij} \sim i^{-1} e^{\beta(j-i)}$$

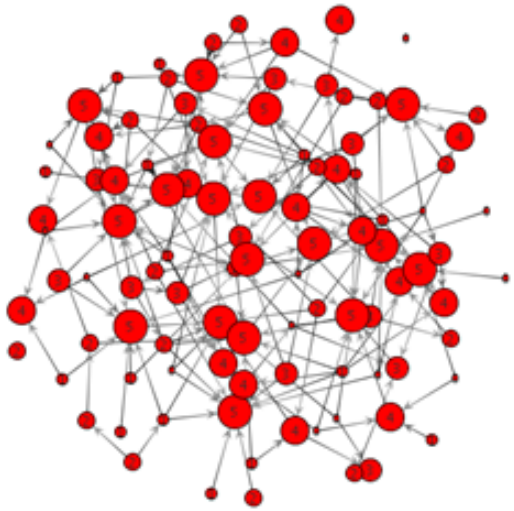
'just better' preferred

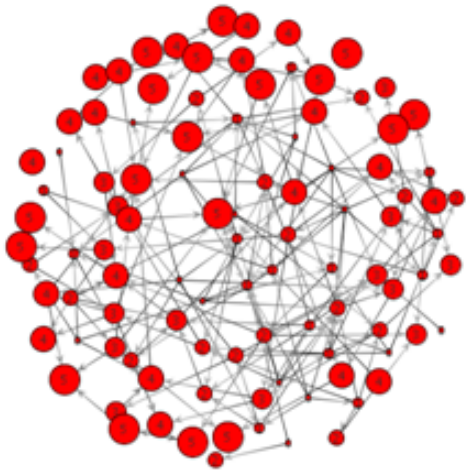


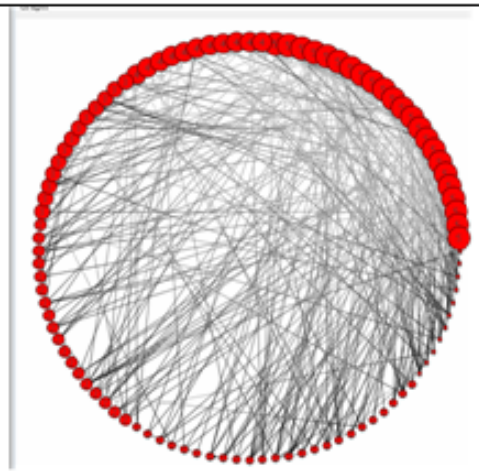
$$p_{ij} \sim i^{-1} e^{\gamma(i-j)}$$

j>i

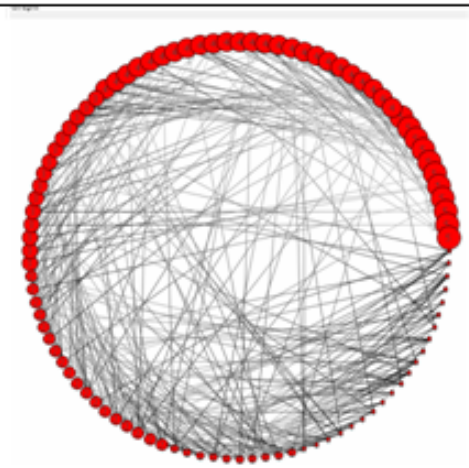
visualization







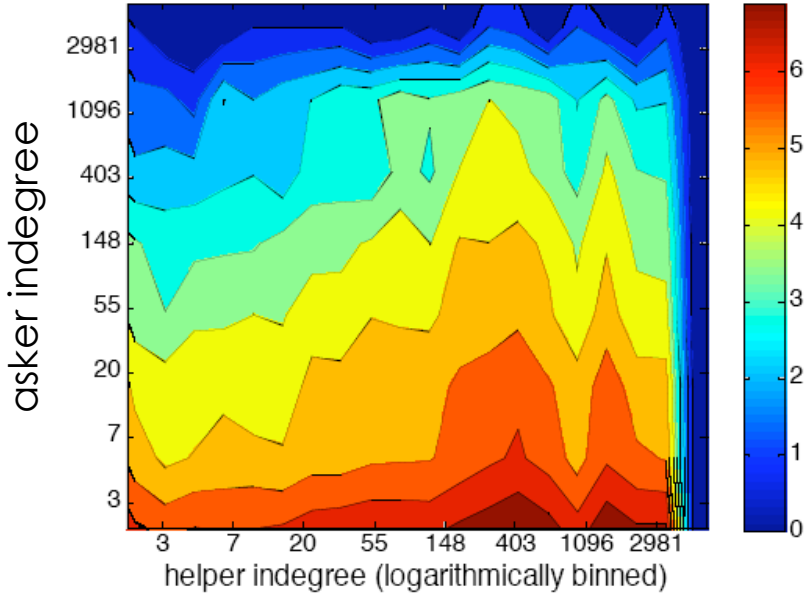
Best “preferred”

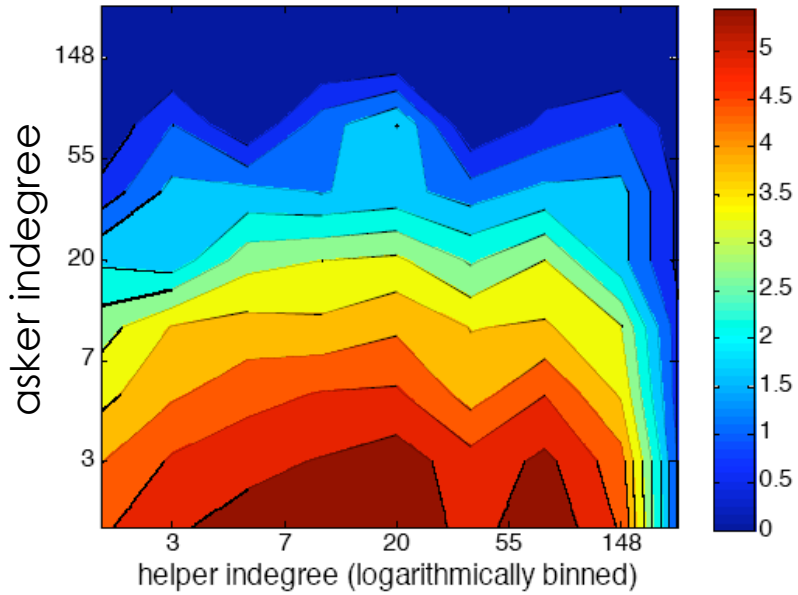


just better

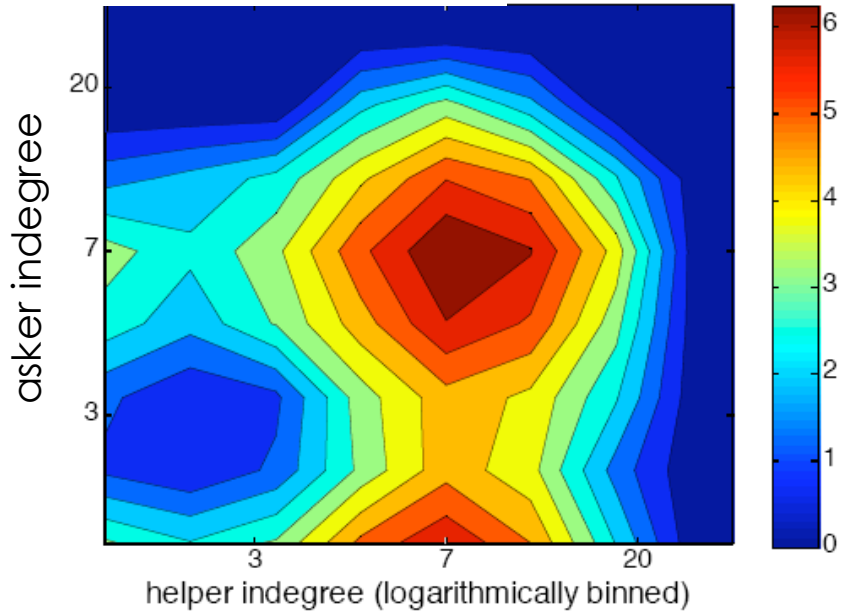
degree correlation profiles

degree-degree correlations between asker and helper (empirically observed)



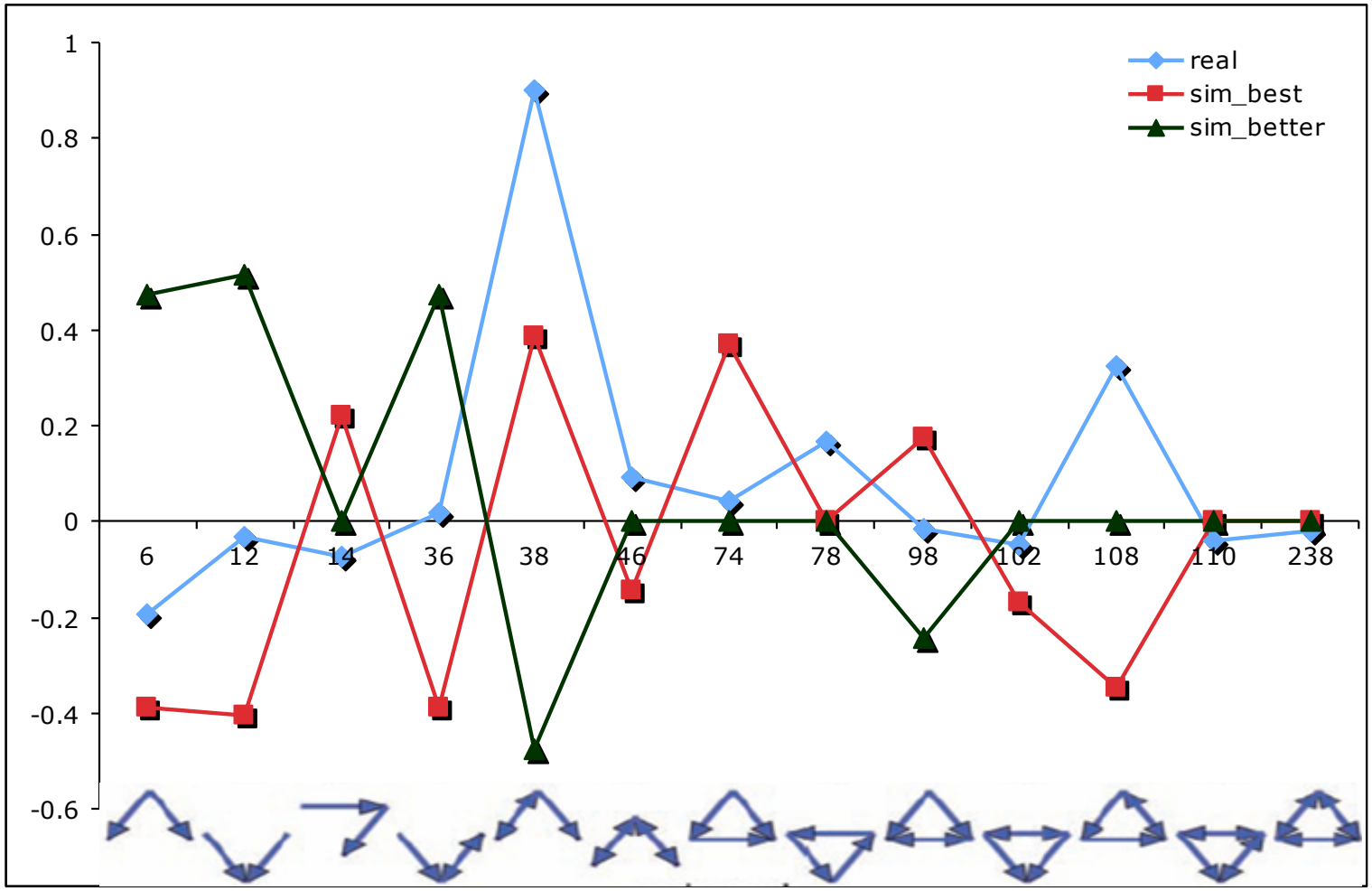


best preferred (simulation)

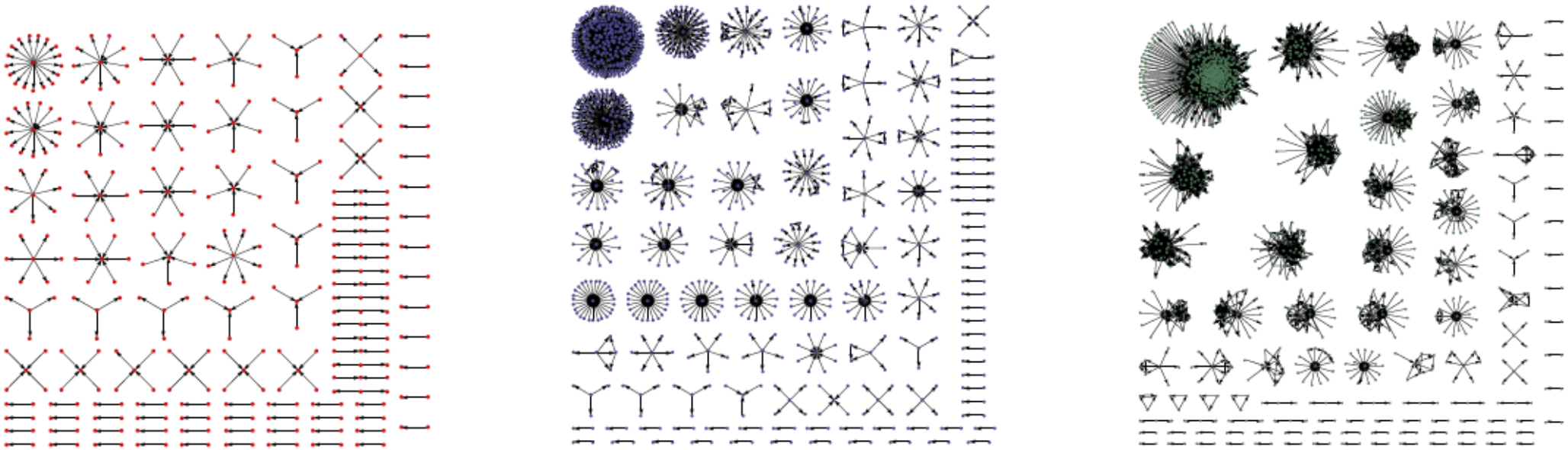


just better (simulation)

But motifs reveal that we are not capturing all the local structure!



Egonetworks: another way of capturing local structure



(a) Programming (b) Marriage (c) Wrestling

Adamic, Zhang, Bakshy, Ackerman WWW'08

Recipe recommendation using ingredient networks

Chun-Yuen Teng[1], Yu-Ru Lin[2,3], Lada A. Adamic[1]

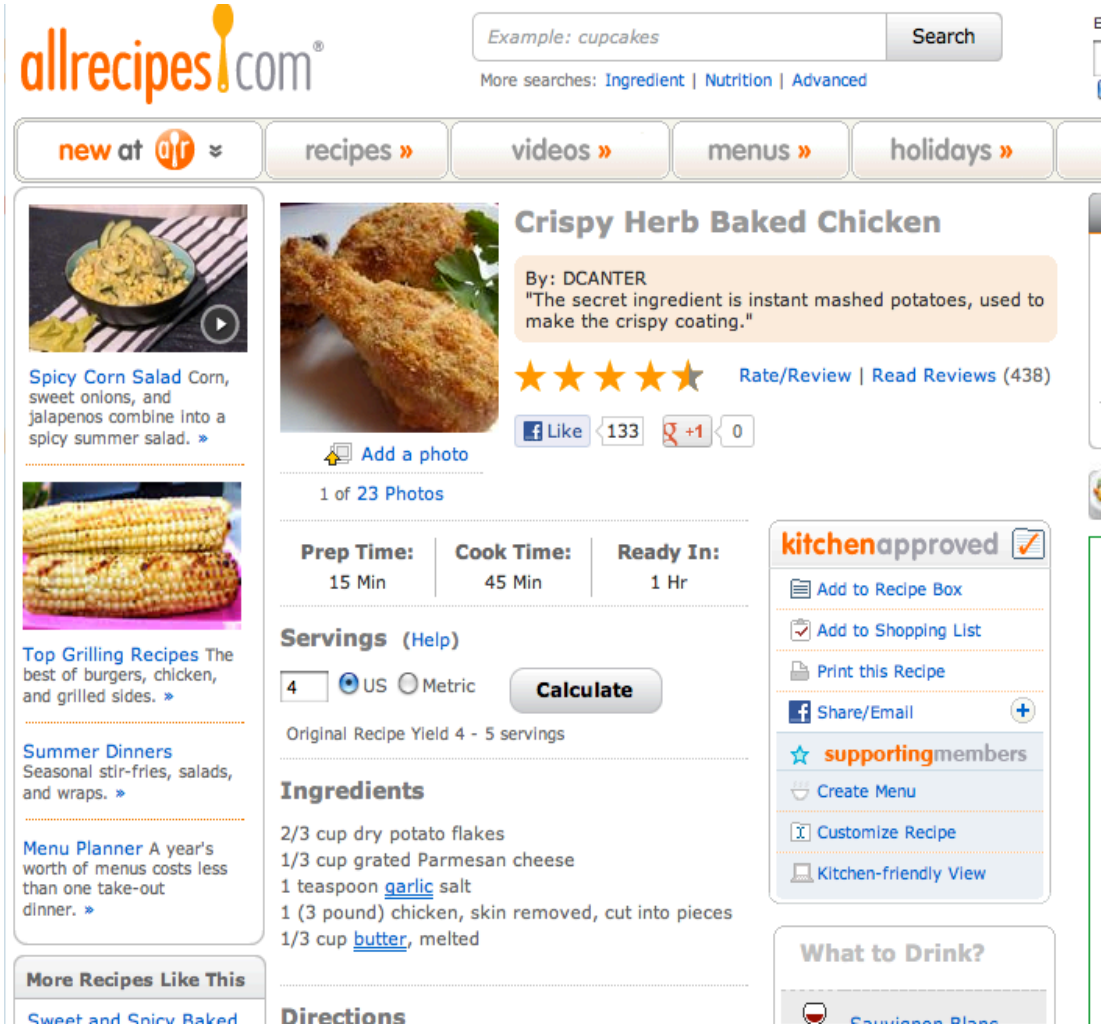
[1]School of Information, University of Michigan

[2]IQSS, Harvard University

[3]CCS, Northeastern University

WebSci'12

Online recipes



Our online recipes

- allrecipes.com®
- 46,337 recipes
- Each recipe includes directions, ingredients, nutrition info, cooking time, and regional information
- 1,976,920 reviews include ratings and text

Research questions

- What patterns emerge from the collective cooking knowledge aggregated in recipes?
- How can ingredient networks be used for predicting recipe ratings?

Recipe mining

- Cooking methods
  - Regional preferences
- Ingredients
  - Combination of ingredients
  - Modification of ingredients
- Predicting ratings:

OR

Credit: sonomaorganics.com

Ingredients

Combining ingredients

+

✔

+

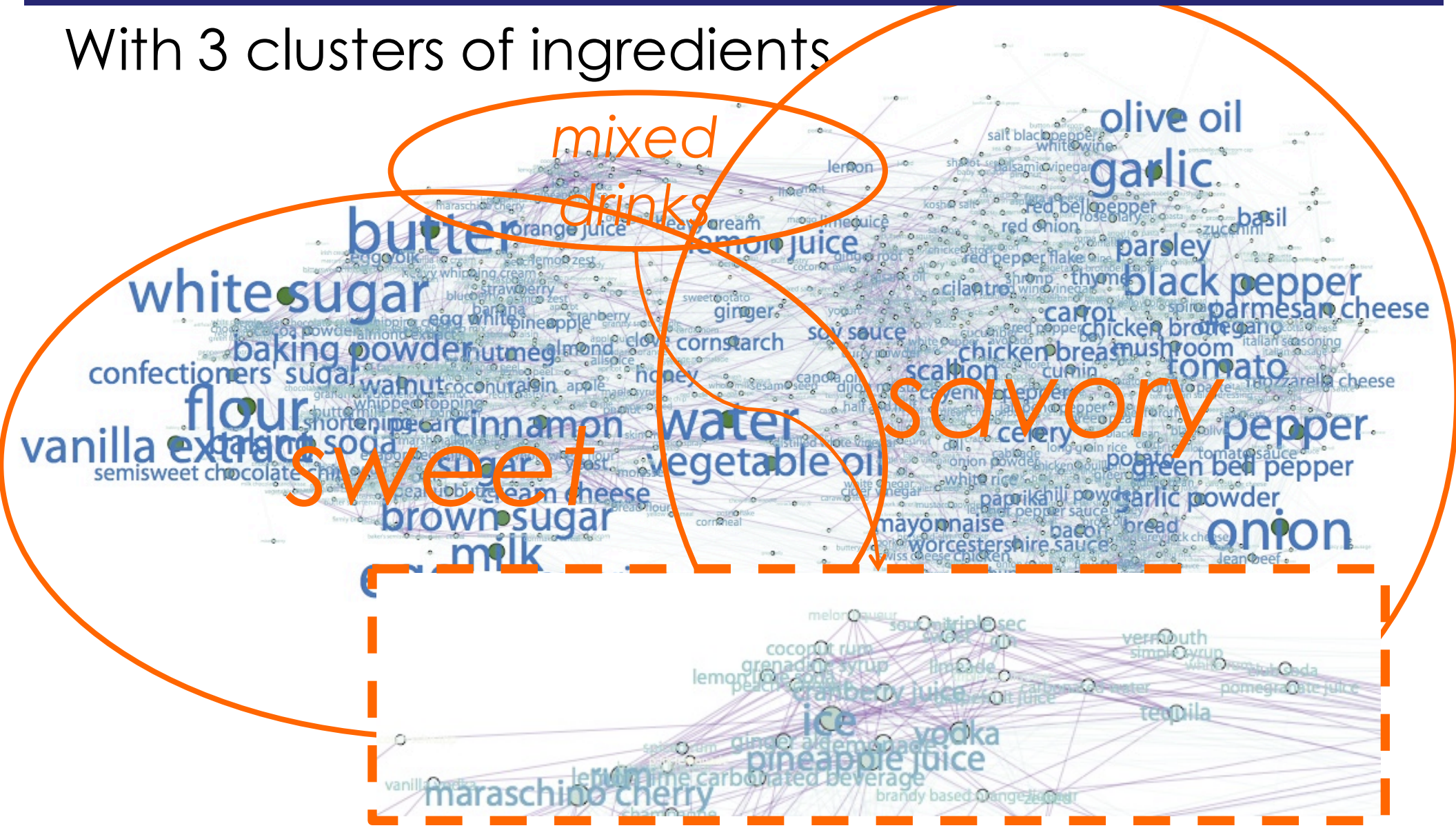
✗

Complement network

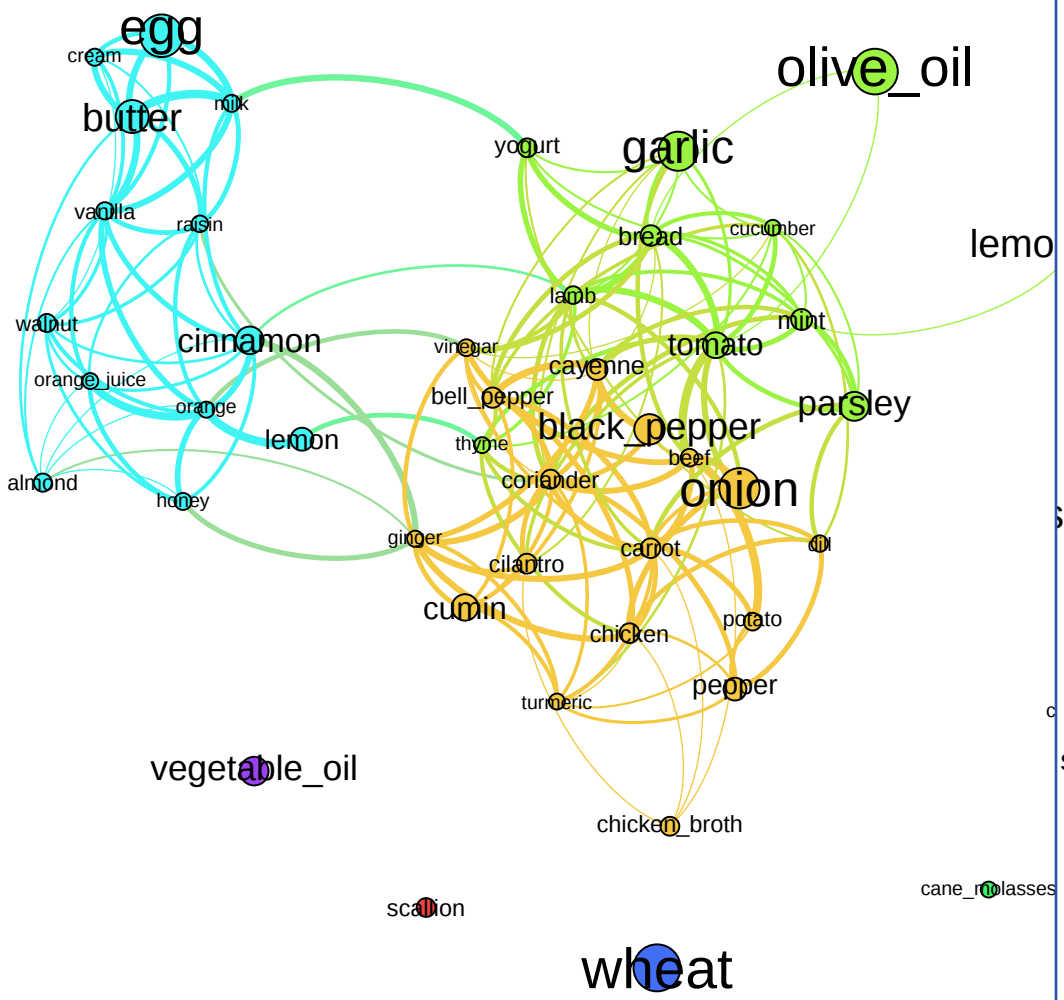
- Nodes: ingredients
- Undirected edges:
  - Weighted by pointwise mutual information
    PMI = log (P(a,b) / P(a)P(b))
    - P(a,b) = (# of recipes containing a and b)/(# of recipes)
    - P(a) = (# of recipes containing a)/(# of recipes)
    - P(b) = (# of recipes containing b)/(# of recipes)
- Recipe rating and PMI of its ingredient pairs
  - Mean and minimum of PMI (no correlation with rating)
  - Max of PMI (ρ=0.09, $p < 0.001$)

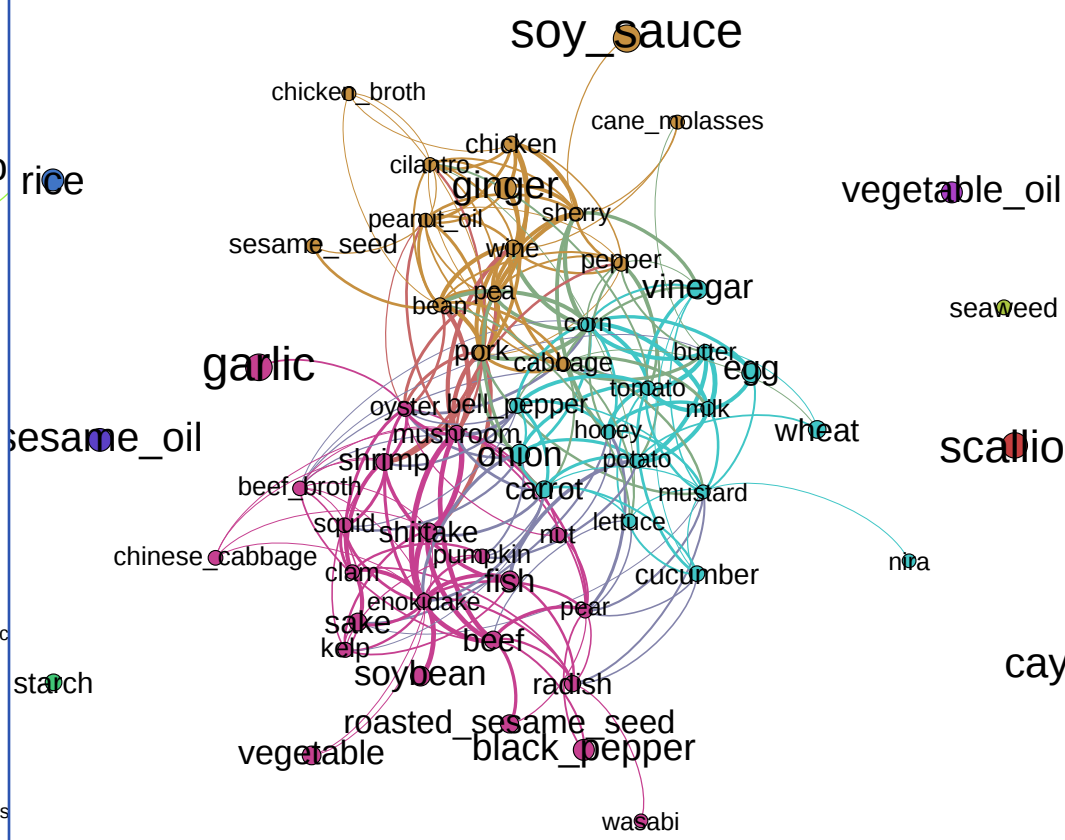
Complement network

With 3 clusters of ingredients



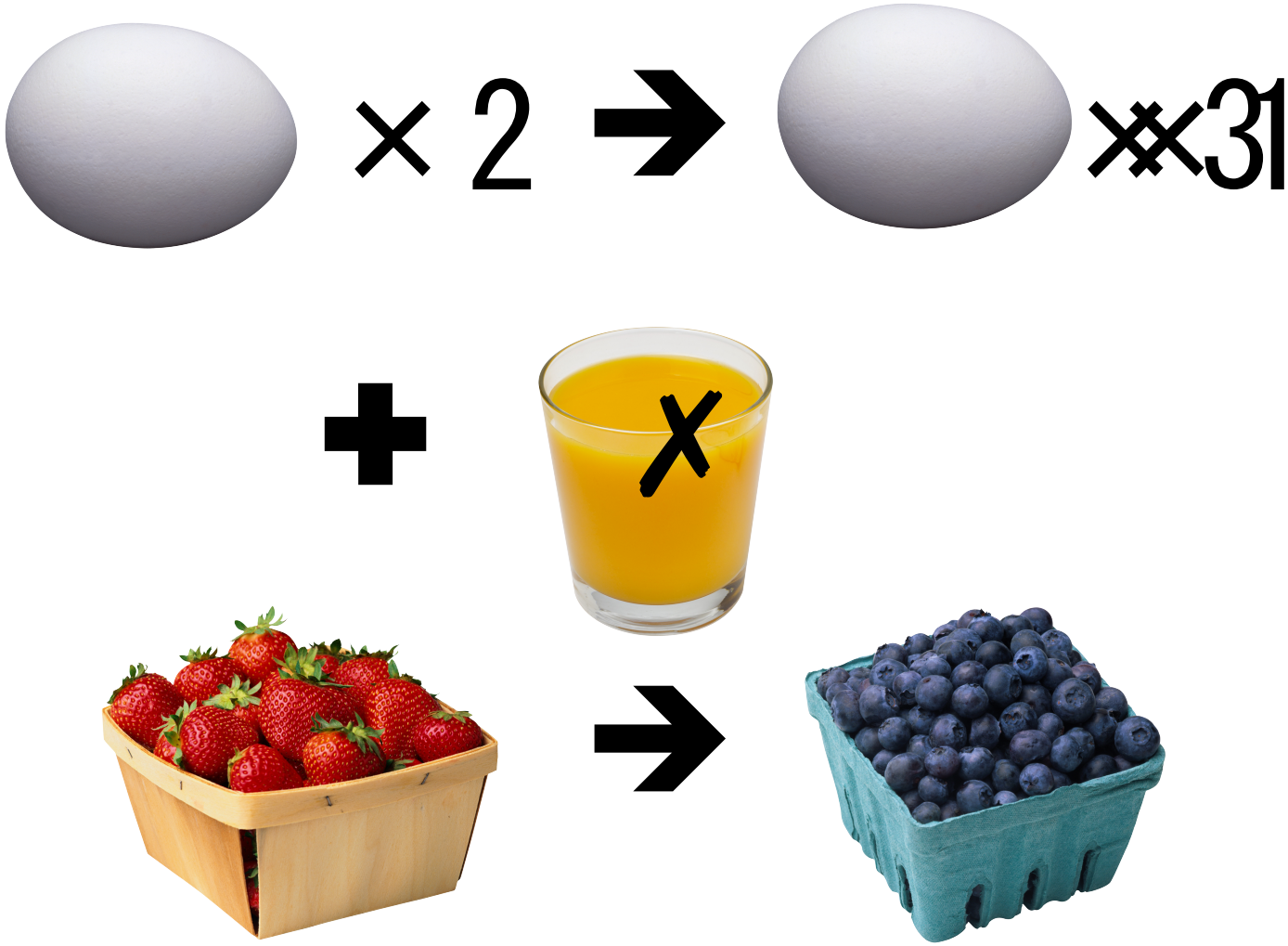
Guess the cuisine!





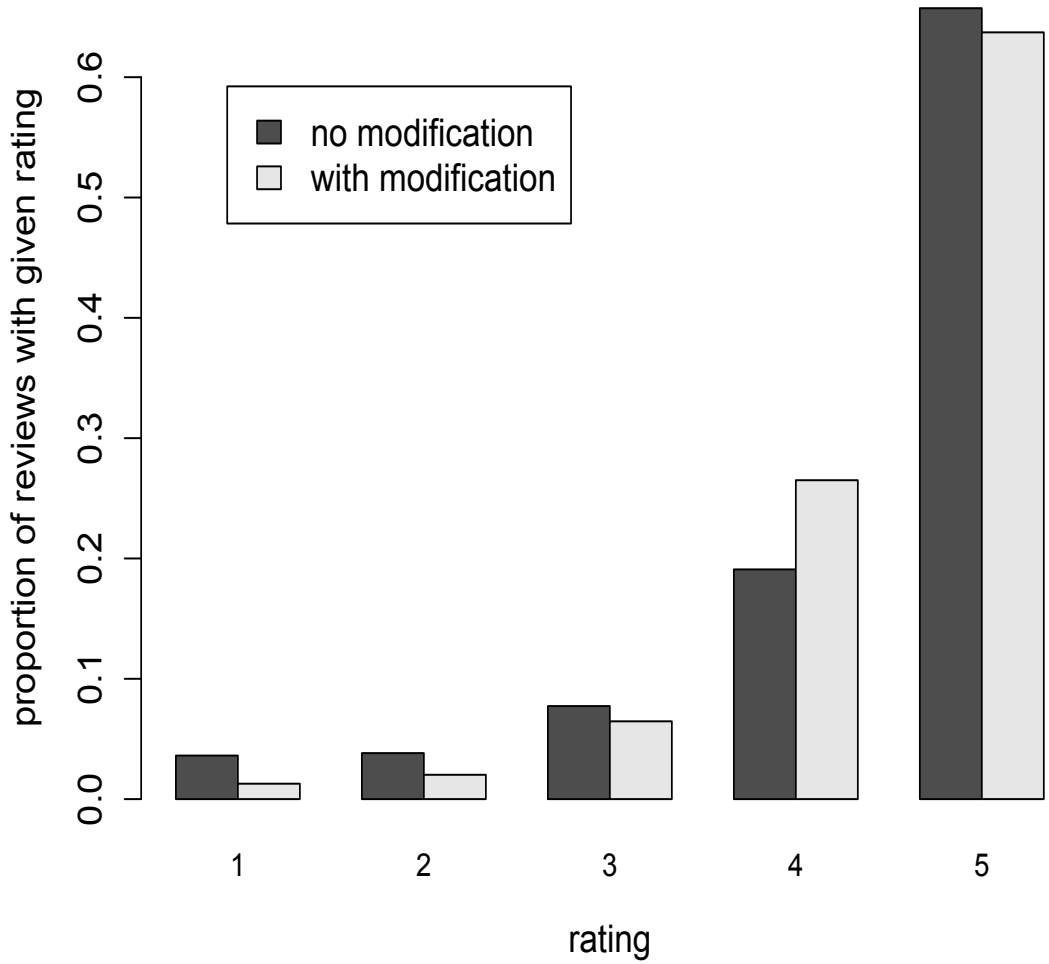
What do you notice about sweet vs. savory ingredients?

Recipe modification

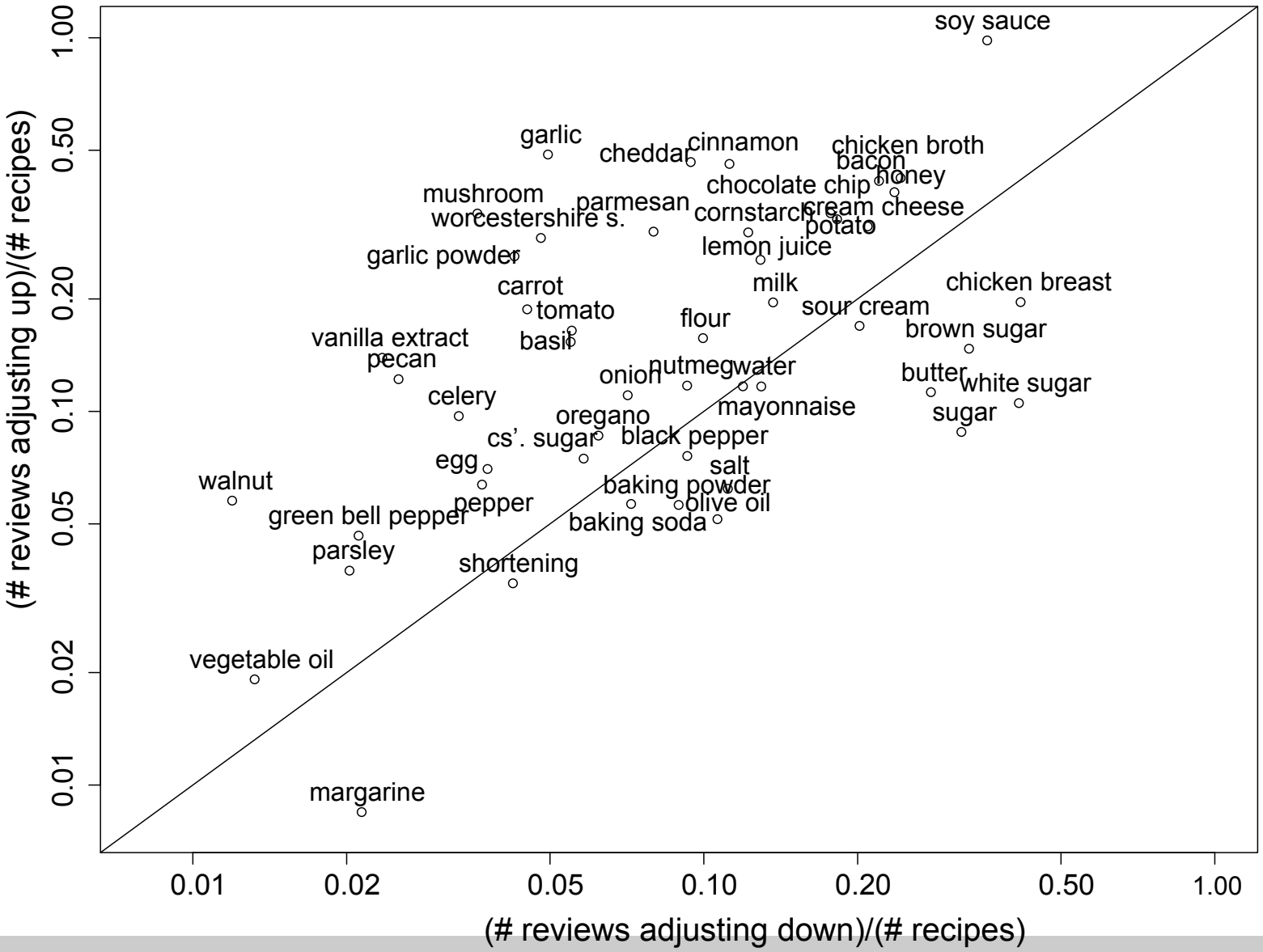


Recipe modification

- 60% of reviews contain "add", "omit", "instead", "extra", and 14 others.
- Reviews that include changes assign higher star ratings (4.49 vs. 4.39, $p<10^{-10}$)
- almost perfect but not quite (4 star) reviews often suggest modifications



Suggested modifications of quantities



Correlations between ingredient modifications

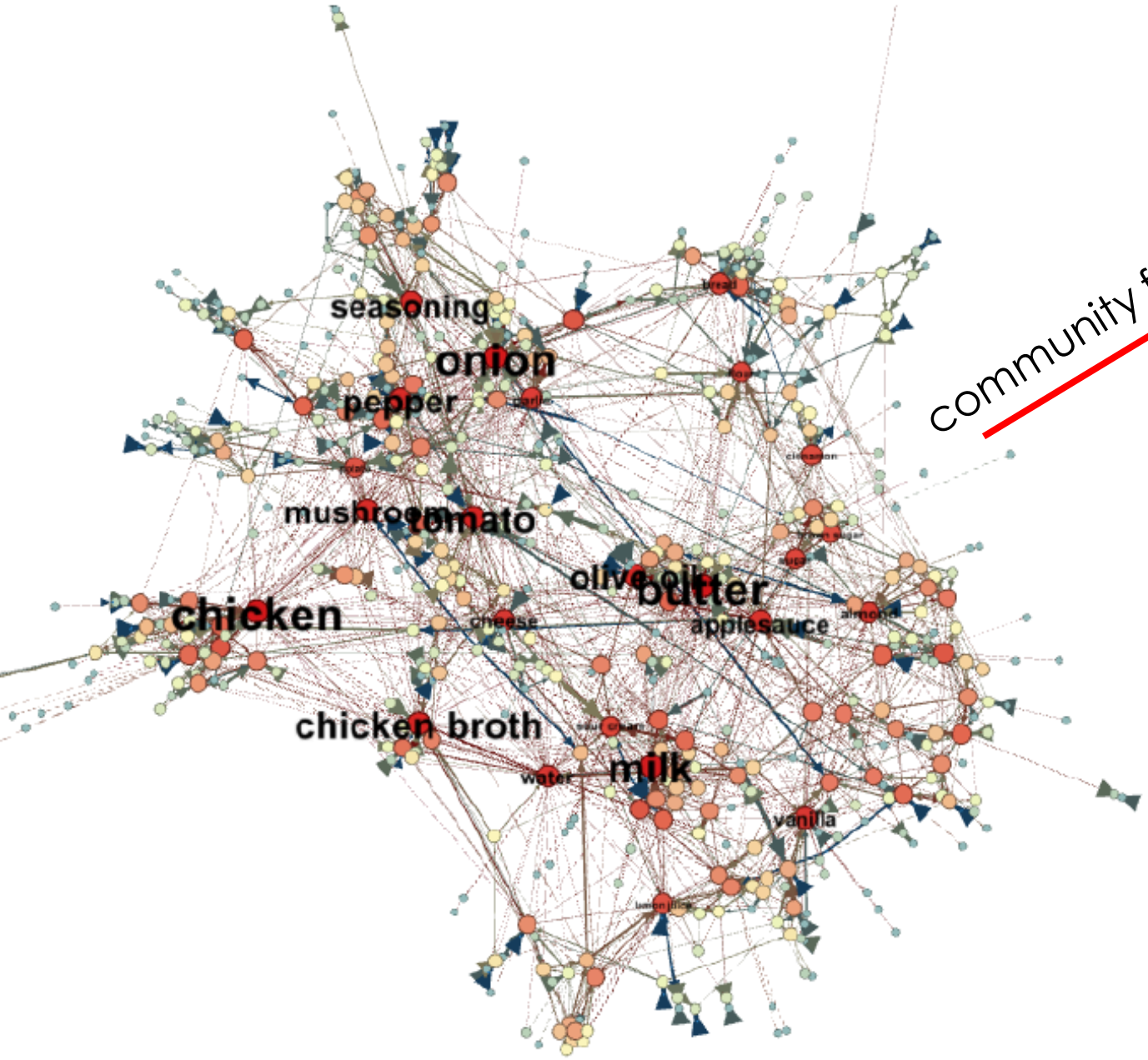
- Recipe freq. vs. deletion/recipe (ρ = -0.22)
- Recipe freq. vs. addition/recipe (ρ = -0.25)
- Recipe freq. vs. increase/recipe (ρ = -0.26)
- Correlations between ingredient modifications

	addition	deletion	increase	decrease
# recipes	0.41	0.22	0.61	0.68
addition		-0.15	0.79	0.11
deletion			0.09	0.58
increase				0.39

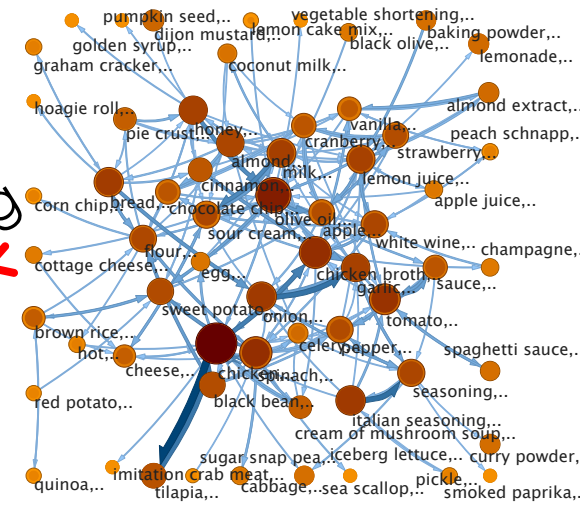
Substitution network

- Extract substitution relationships from comments
  - e.g. "I replaced the butter with sour cream"
  - "replace **a** with **b**", "substitute **b** for **a**", "**b** instead of **a**"
- Nodes: ingredients
- Edge weights = p(**b** | **a**), which is the proportion of substitutions of ingredient **a** that suggests ingredient **b**

Substitution network



community finding

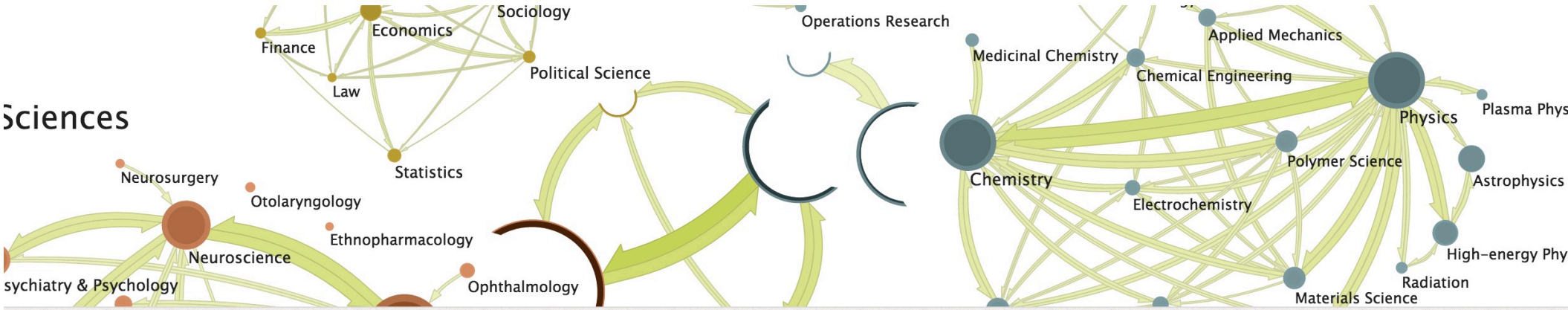


Try this at home: mapequation.org

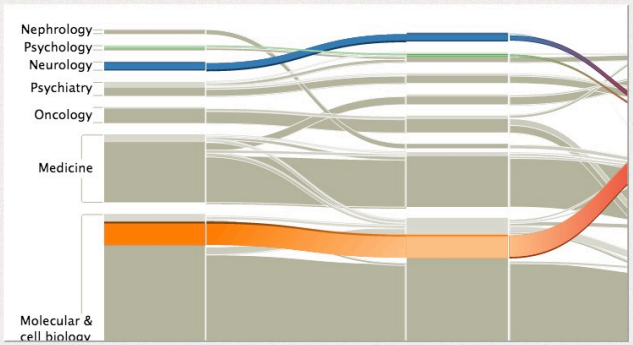
MapEquation

apps code publications about

Multilevel community detection with Infomap



Apps »



Code »

```
using infomath::plogp;
for (unsigned int i = 0; i < numNodes; ++i)
{
enter_log_enter += plogp(m_moduleFlowData[i].enterl
exit_log_exit += plogp(m_moduleFlowData[i].exitFlo
flow_log_flow += plogp(m_moduleFlowData[i].exitFlo
enterFlow += m_moduleFlowData[i].enterFlow;
}
enterFlow += exitNetworkFlow;
enterFlow_log_enterFlow = plogp(enterFlow);
```

Publications »

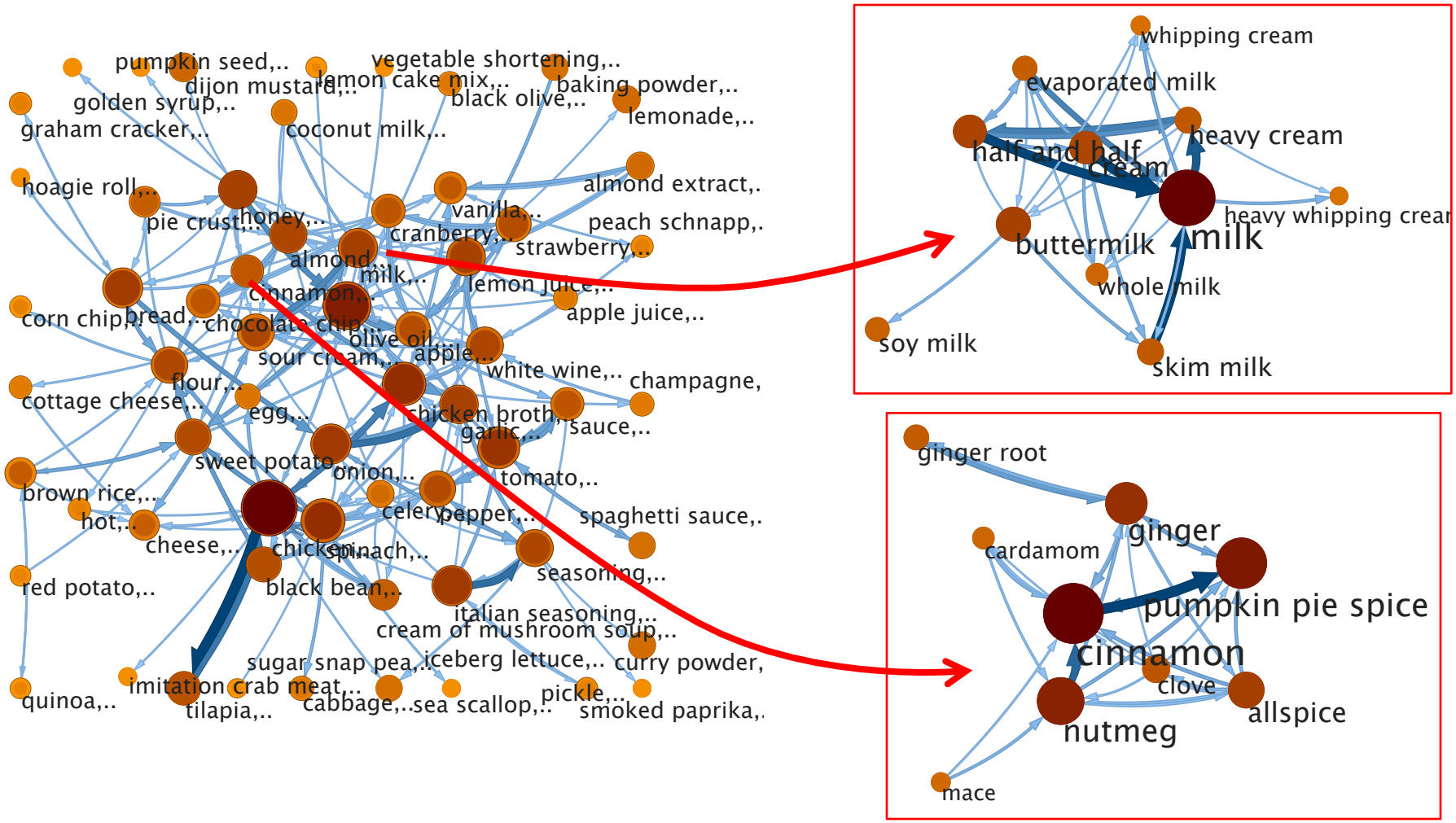
Maps of information flow reveal community structure in complex networks

Martin Rosvall and Carl T. Bergstrom

PNAS **105**, 1118 (2008). [arXiv:0707.0609]

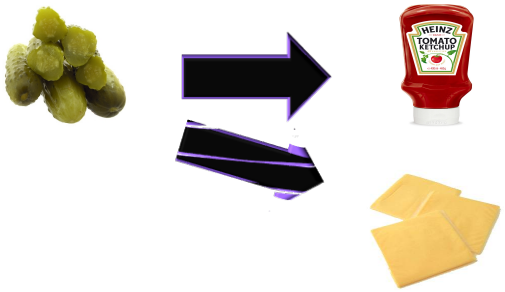
To comprehend the multipartite organization of large-scale biological and social systems, we introduce a new information-theoretic approach to reveal community structure in

Substitution network: communities



Substitution network and users' preference

- Preference network
  - Create an edge from ingredient **a** to **b** if rating(**a**) < rating(**b**)
  - ex:
    - Recipe X contains 
    - Recipe Y contains 
    - Rating(X) > Rating(Y)



Substitute network and users' preference

- Weight of preference network
  - PMI(a->b)=log(p(a->b)/p(a)p(b))
  - where p(a->b)= (# of recipe pairs from a to b)/(# of recipe pairs)
- Correlations between preference network and substitute network (ρ = 0.72, p<0.001)

Prediction task

- Given a recipe pair with overlapped ingredients, determine which one has the higher rating

Prediction task

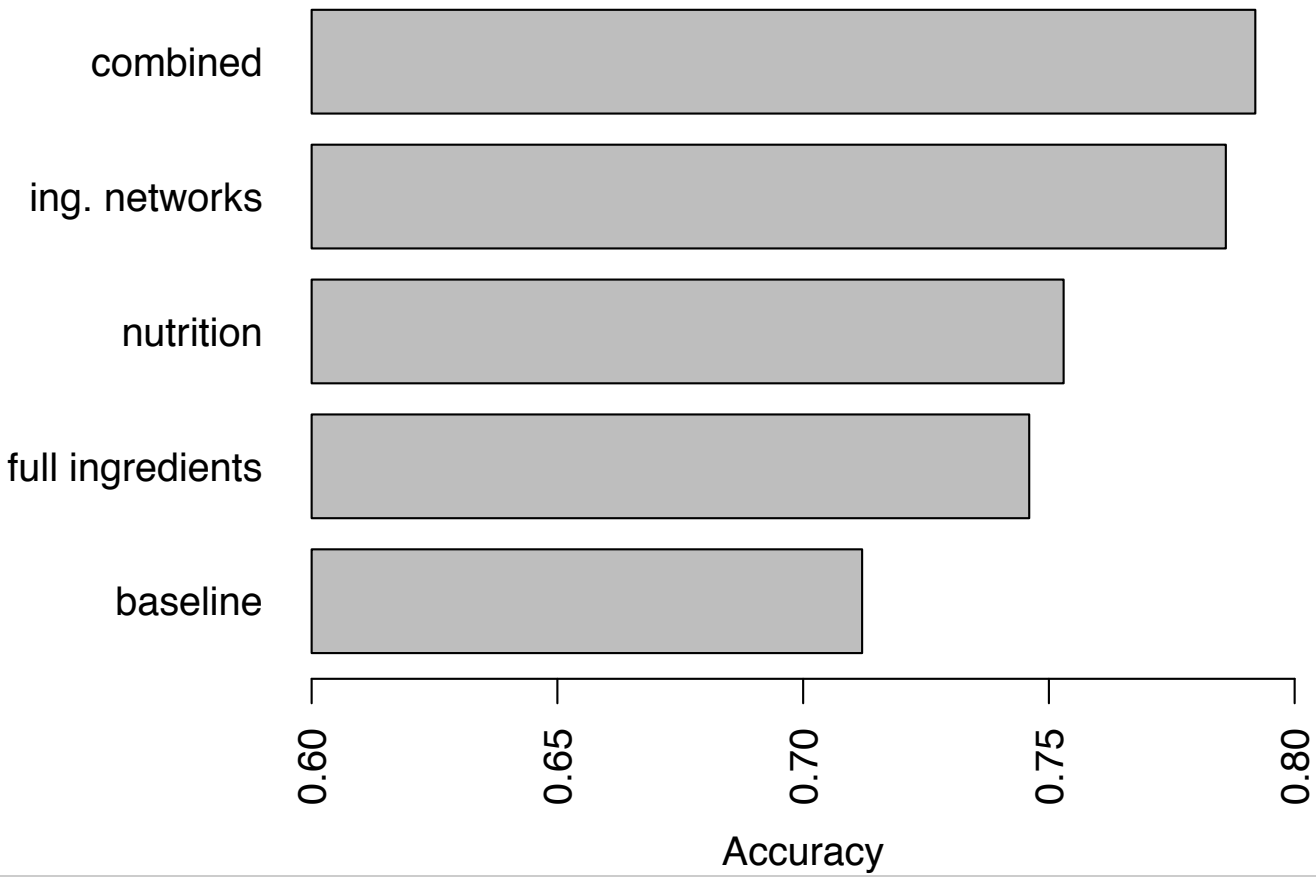
- Features
  - Baseline
    - Cooking methods, preparation time, the number of servings
  - 1000 popular ingredient list
    - Binary vector indicating the occurrence of ingredients
  - Nutrition
    - Calories, carbohydrates, fat, etc.
  - Ingredient networks
    - Network positions (centrality) and communities (SVD)
  - Combined set
    - Everything listed above

Prediction task

- 62,031 recipe pairs (X,Y)
  - where rating(X) > rating(Y)
  - >= 10 user reviews
  - >= 50% users have rated both recipes
  - Cosine similarity of ingredients (X,Y)> 0.2
- Train with gradient boosting tree
  - balanced dataset
  - 2/3 for training, 1/3 for testing
  - Evaluate based on accuracy

Prediction performance

- Ingredient network features lead to improved performance



Recipes encode our collective cooking knowledge

- complementarity of ingredients
- regional preferences in combining ingredients
- substitutability of ingredients
- complement and substitute networks encode users' preferences and can be used to effectively predict recipe ratings

Wrap up

- We saw diverse applications
  - Tie strength
  - Network diversity
  - Motifs
  - Centrality & community structure
- What makes it interesting:
  - Measurable consequence: tasty food (ratings), information (email topics), respect (survey responses)!
  - Models!