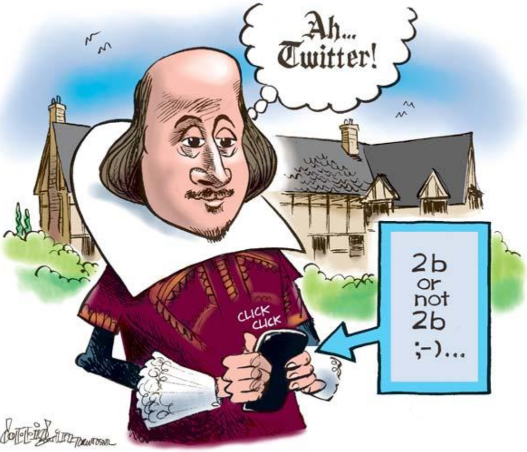
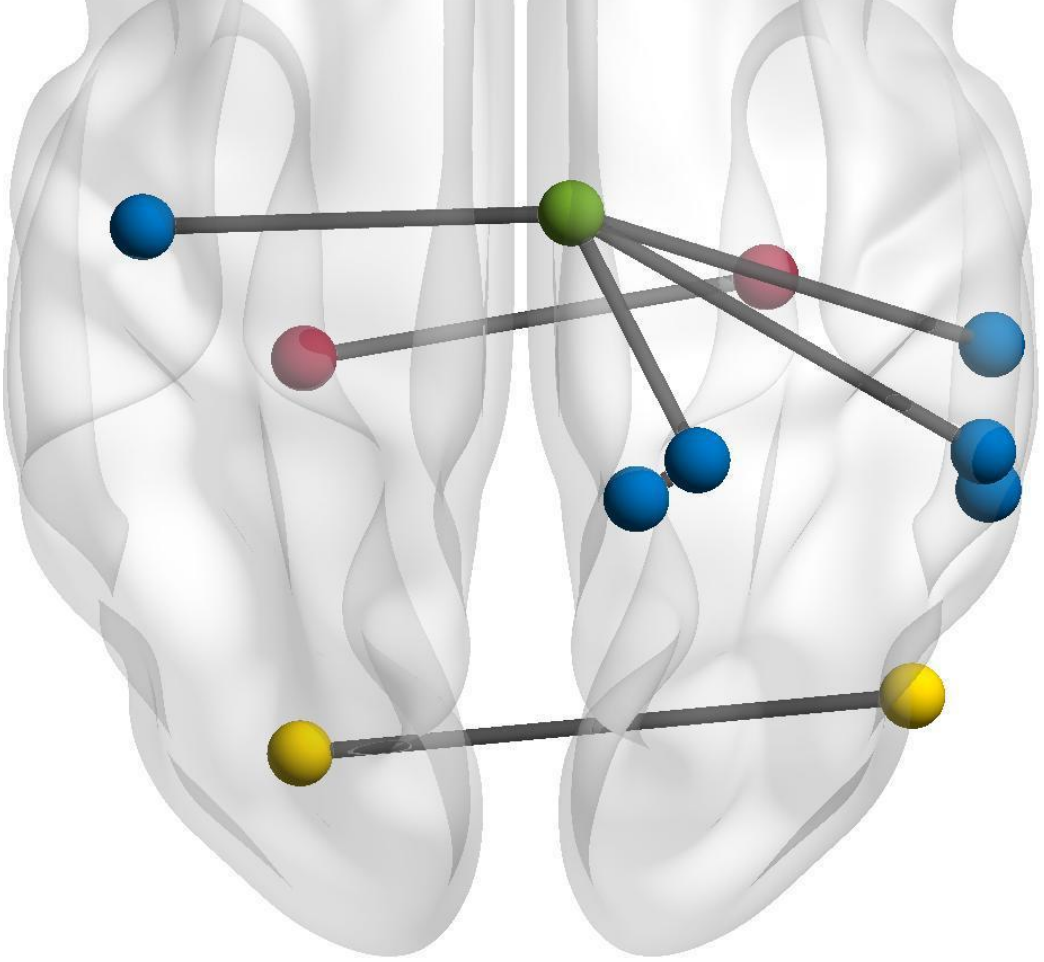


To be connected, or not to be connected...

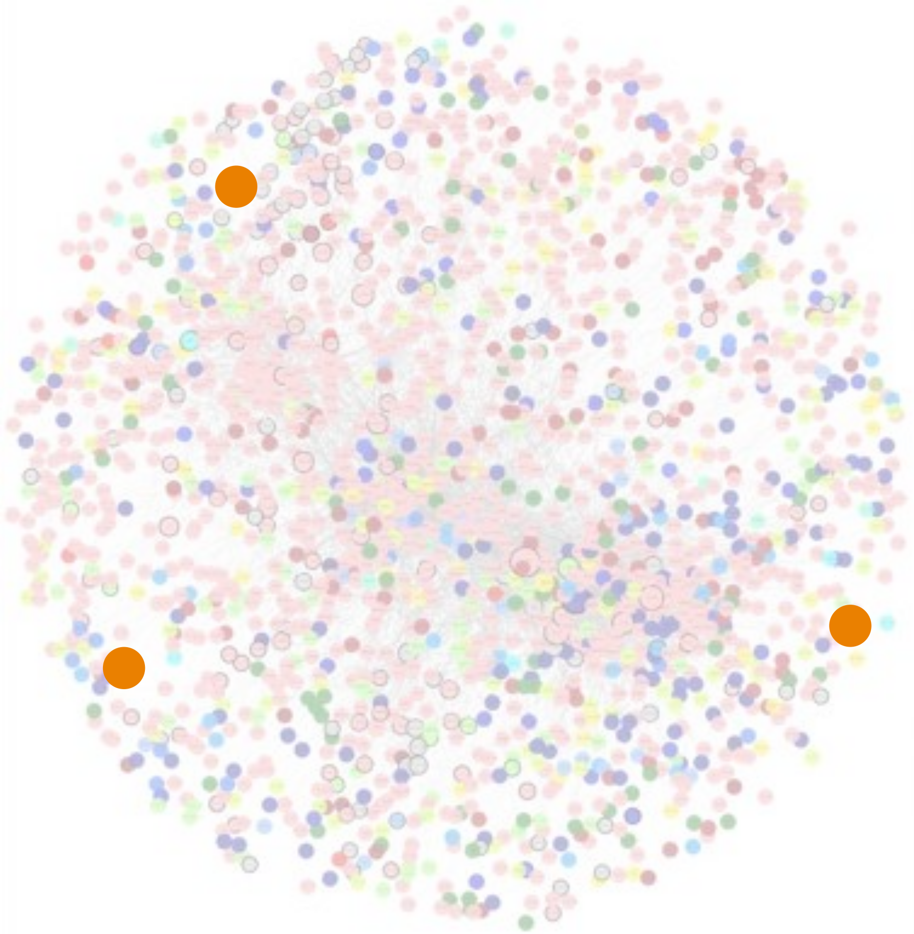
That is the **Minimum Inefficiency Subgraph** Problem

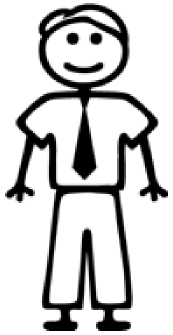
Natali Ruchansky
Francesco Bonchi
David Garcia-Soriano
Francesco Gullo
Nicolas Kourtellis



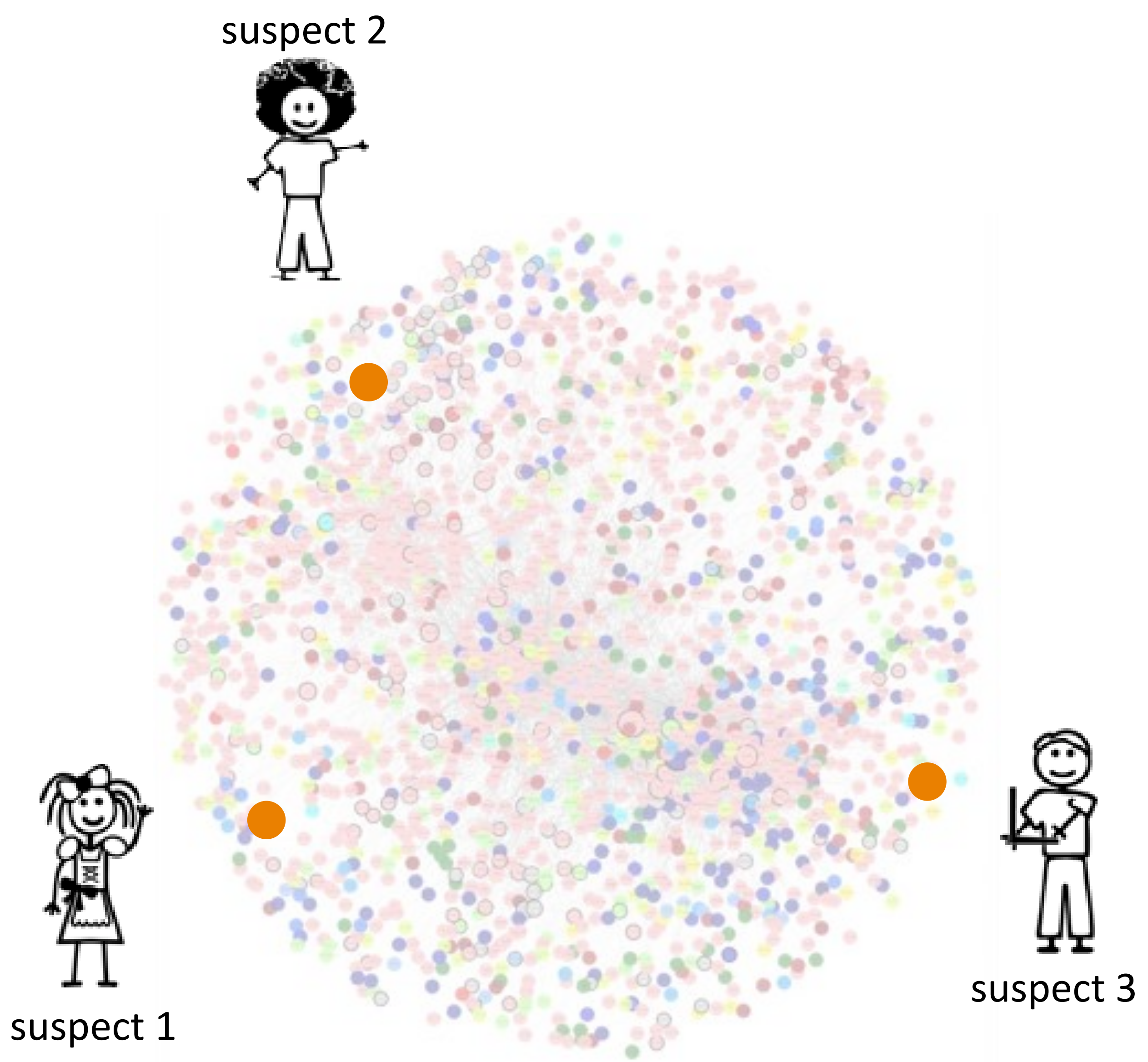
Biologists in Lab X have constructed a large protein-protein interaction network (PPI).

Biologists in Lab X have constructed a large protein-protein interaction network (PPI).

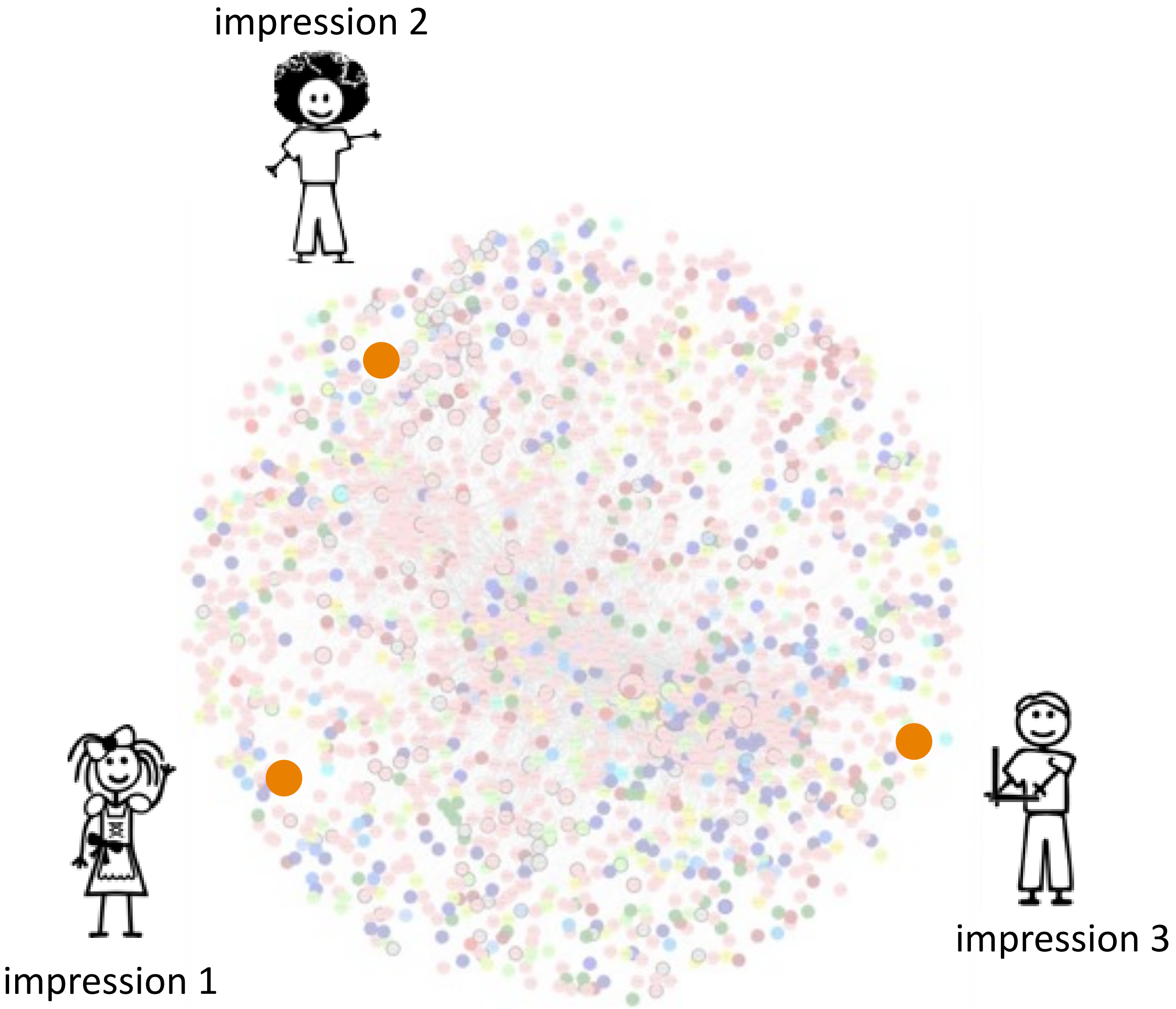




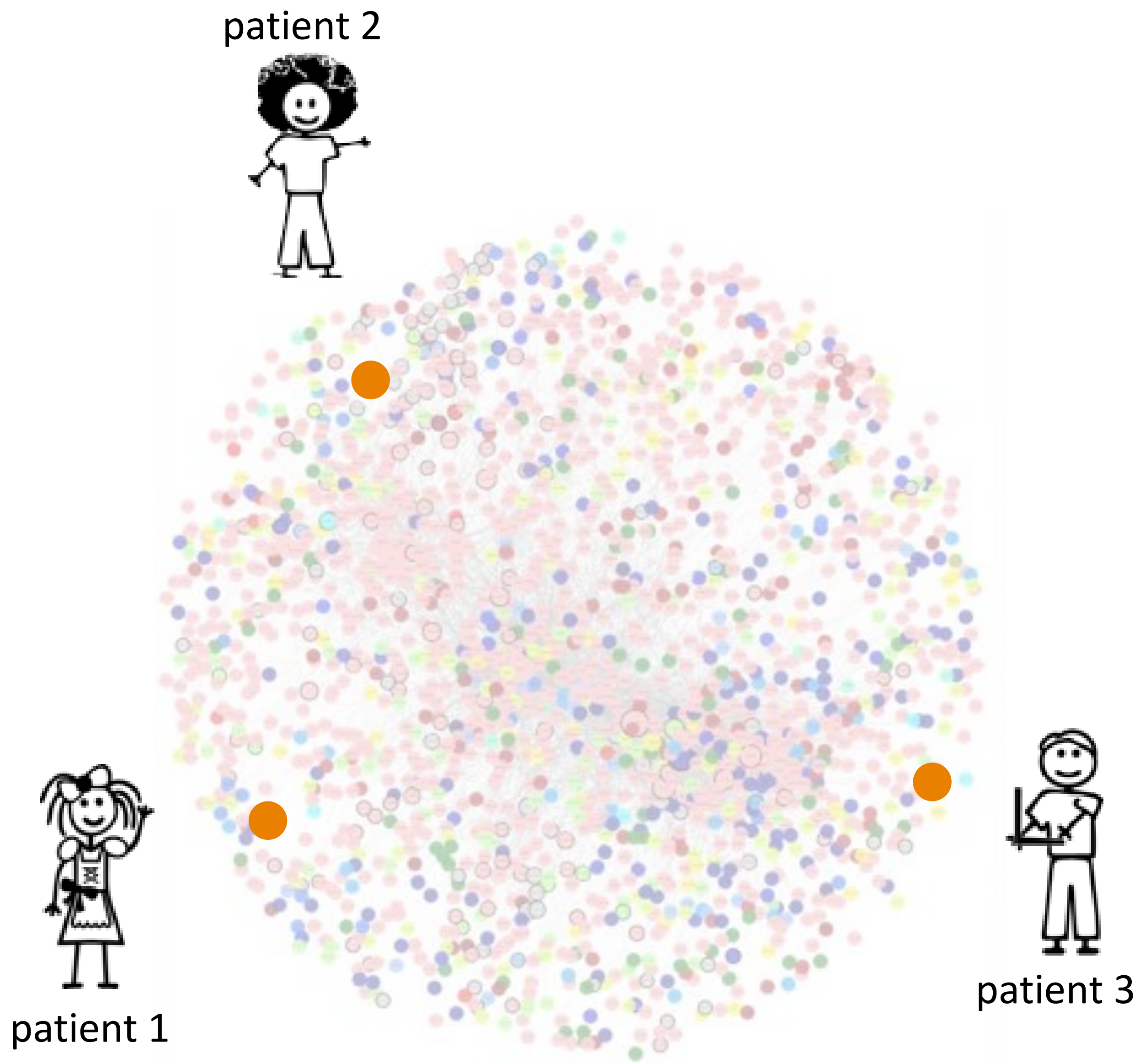
The PI has tasked them with making an amazing discovery about relationship among **specific proteins** P1, P2, and P3.



Given a set of **subjects in a terrorist network** suspected of organizing an attack.
Which other subjects, likely to be involved, should we keep under control?



Given a set of **users who clicked on an ad**, who else should the ad be displayed to?



Given a set of **patients infected** with a viral disease, which other people should we monitor?

Community search / seed set expansion

- General class of problems of the form:

> Given a graph $G=(V,E)$ and a set of vertices $Q \subset V$,
> find a subgraph H of G that "explains" the connections among Q.
> (H minimizes/maximizes some objective function)

- Several approaches in the literature
  - H must be a connected subgraph
  - Mostly based on random-walks
  - Tend to return rather large solutions
  - Solutions get very large when query nodes belong to different communities
  - Have parameters

The Minimum Wiener Connector Problem
(SIGMOD 2015)

Our proposal: find the connected subgraph H containing Q and minimizing the **Wiener Index** (the sum of pairwise distances)

$$\boxed{H^* = \mathop{\arg\min}_{G[S]:Q\subseteq S\subseteq V} \sum_{\{u,v\}\in S} d_{G[S]}(u,v)}$$

- Parameter-free
- Returns smaller and denser subgraphs

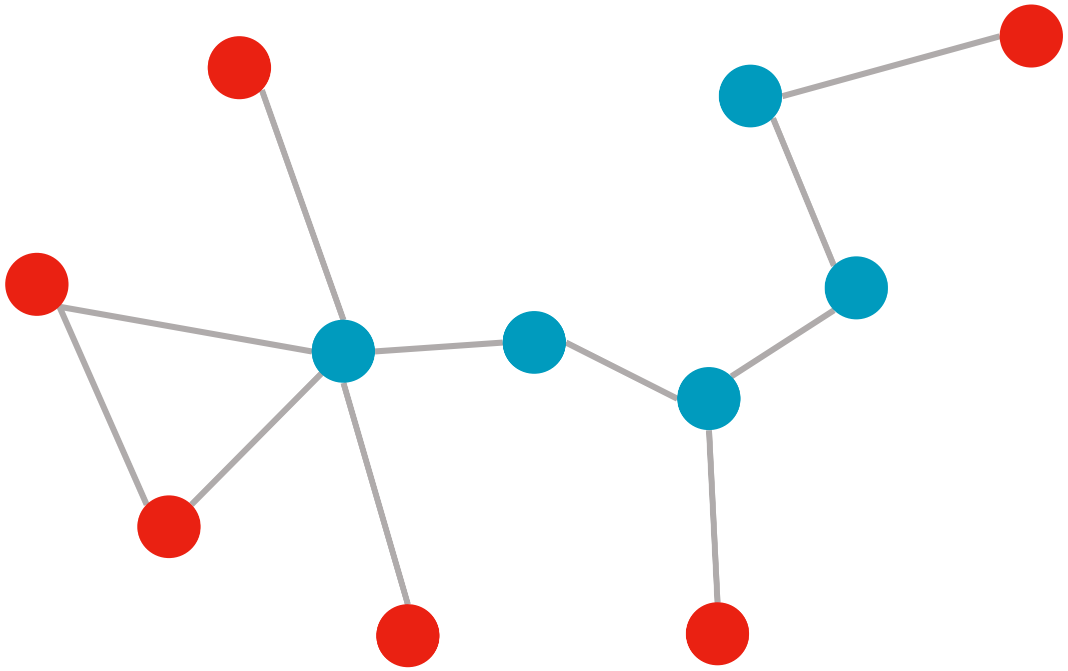
  No matter whether the query nodes belong to the same community or not
- Add "important" nodes (high centrality)
- Efficient algorithm with approximation guarantees

Smaller, denser, and more central vertices

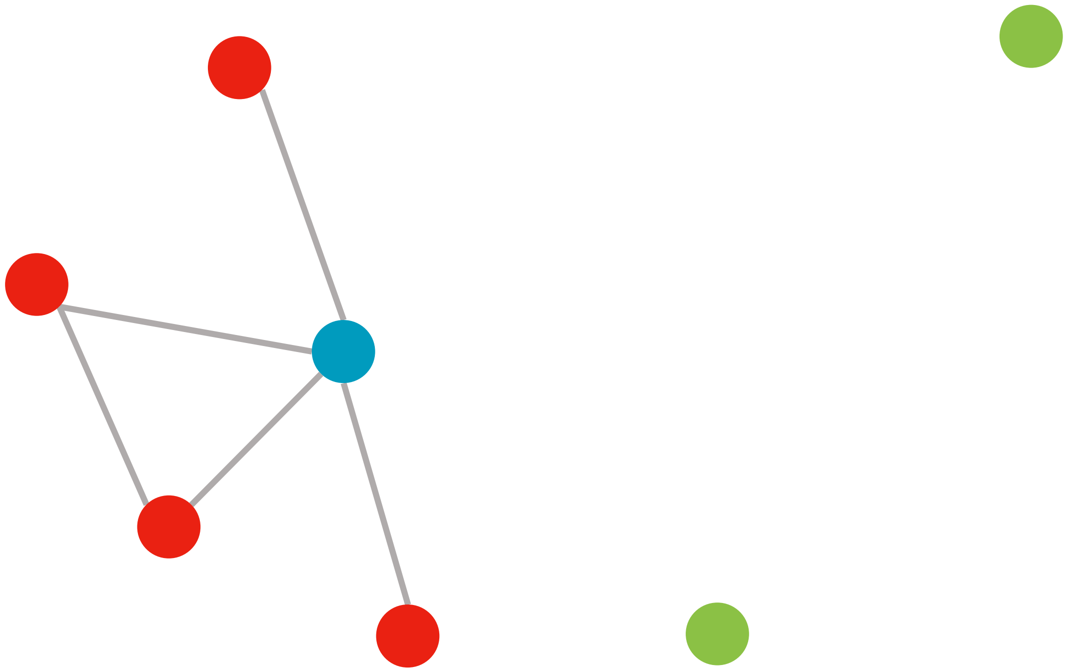
		email	yeast	oregon	astro	dblp	youtube	
$\lvert V[H]\rvert$		671	819	9028	12758	11804	17865	CTP
		155	188	4556	1735	7349	5615	CPS
		137	100	1846	598	842	684	PPR
		26	**24**	26	26	25	19	ST
		24	**24**	**23**	**23**	**23**	**17**	WS-Q
$\delta(H)$		0.016	0.016	0.01	<0.01	<0.01	0.01	CTP
		0.047	0.028	0.02	0.019	0.01	<0.01	CPS
		0.029	0.039	0.02	0.07	0.01	0.02	PPR
		0.080	0.088	0.090	0.09	0.08	0.1	ST
		0.093	**0.091**	**0.106**	**0.13**	**0.11**	**0.13**	WS-Q
$bc(H)$		<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	CTP
		0.03	0.02	<0.01	<0.01	<0.01	<0.01	CPS
		0.03	<0.01	<0.01	0.02	0.01	<0.01	PPR
		0.09	0.07	0.10	0.11	0.10	0.13	ST
		0.11	**0.11**	**0.12**	**0.14**	**0.12**	**0.18**	WS-Q
$\mathbf{W}(H)$		$\approx 750k$	$\approx 2M$	$\approx 137M$	$\approx 292M$	$\approx 400M$	$\approx 1.5G$	CTP
		54 598	69 296	$\approx 50M$	$\approx 8.3M$	$\approx 12.6M$	$\approx 561M$	CPS
		52 222	15 838	$\approx 7.5M$	40 079	$\approx 1.2M$	$\approx 1.3M$	PPR
		1 200	1 259	1 164	1 318	3 371	1 324	ST
		968	**931**	**923**	**1 007**	**2 043**	**956**	WS-Q

Relaxing connectivity

instead of forcing connectivity



relax the constraint



Desired Properties

Parsimonious vertex addition

- vertices should be added iff they help forming a more **cohesive** subgraph

Outlier Tolerance

- query vertices which are far from others should remain disconnected

Multi-community awareness

- if the query vertices span multiple communities, connectedness should not be imposed among them

Cohesiveness

- As with the Wiener Connector, we leverage shortest path distances; however, the distance between disconnected vertices is infinite.
- Idea: use the reciprocal of the shortest-path distance! This has the useful property of handling disconnection neatly ($\infty^{-1} = 0$)

Network Efficiency (Latora and Marchiori): $\mathcal{E}(G) = \frac{1}{|V|(|V|-1)} \sum_{\substack{u,v \in V \\ u \neq v}} \frac{1}{d_G(u,v)}$

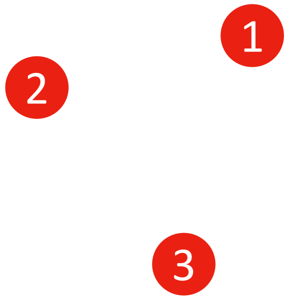
Harmonic Centrality (Boldi and Vigna): $c(u) = \sum_{v \in V} \frac{1}{d_G(v,u)}$

What about these problem statements?

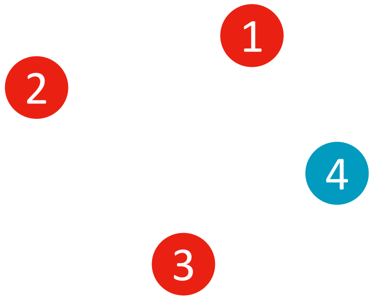
Given a graph $G=(V,E)$ and a set of vertices $Q \subset V$,
find a (not-necessarily connected) subgraph H of G, with
$Q \subset V(H)$ that maximizes network efficiency $E(H)$

Given a graph $G=(V,E)$ and a set of vertices $Q \subset V$,
find a (not-necessarily connected) subgraph H of G, with
$Q \subset V(H)$ that maximizes the total harmonic centrality $C(H)$

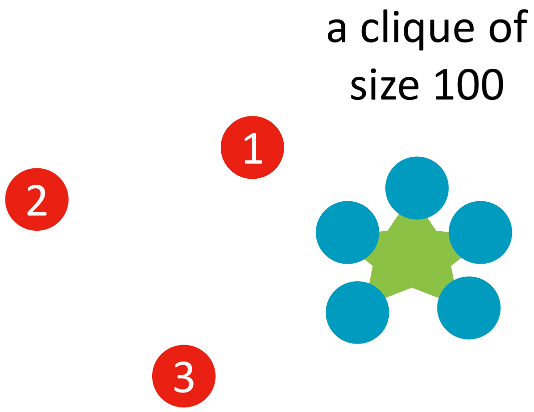
These do not work…



C(G[Q])=0
E(G[Q])=0



C(H)=0
E(H)=0



C(H)=9900
E(H)=0.942

Minimize Network Inefficiency

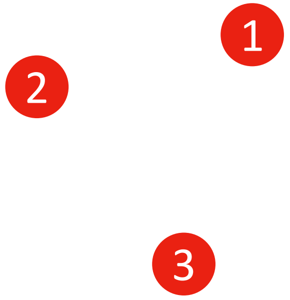
Given a graph G=(V,E), we define its inefficiency as:

$$\mathcal{I}(G) = \sum_{\substack{u,v \in V \\ u \neq v}} 1 - \frac{1}{d_G(v,u)}$$

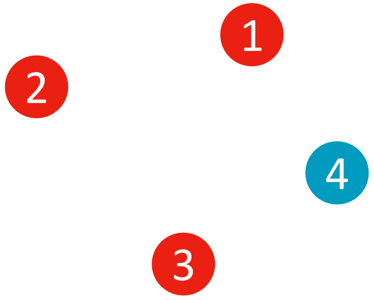
Note:

$$\begin{aligned} \mathcal{E}(G) &= C(G)/(n(n-1)) \\ \mathcal{I}(G) &= n(n-1) - C(G) \end{aligned}$$

… and this works

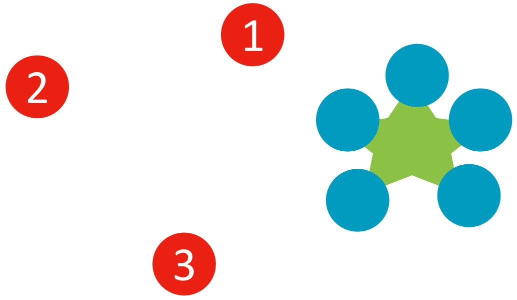


C(G[Q])=0
E(G[Q])=0
I(G[Q])=6



C(G[Q])=0
E(G[Q])=0
I(G[Q])=12

a clique of size 100



C(G[Q])=9900
E(G[Q])=0.942
I(G[Q])=606

Problem statement and hardness

Problem 1 (Min-Inefficiency-Subgraph). *Given an undirected graph $G = (V, E)$ and a query set $Q \subseteq V$, find*

$$H^* = \underset{G[S]:Q \subseteq S \subseteq V}{\arg\min} \; I(G[S]).$$

Theorem 4.1. *Min-Inefficiency-Subgraph is* **NP***-hard, and it remains hard even on undirected graphs with diameter 3.*

Greedy Algorithm

Connect	Start with the Minimum Wiener Connector for Q
Remove	Remove one vertex at a time until Q is disconnected
Choose	Choose the intermediate solution S that minimizes I(S)

Competitors

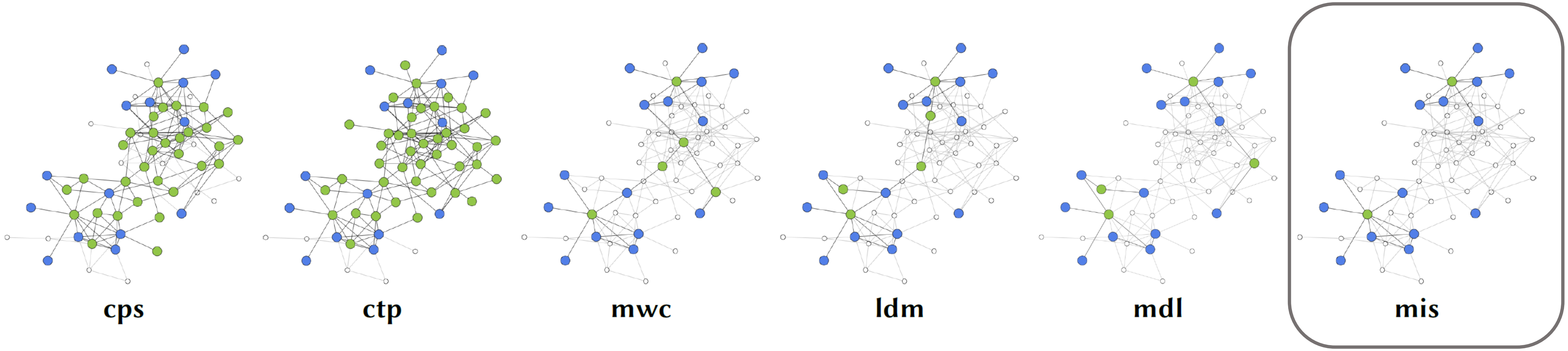
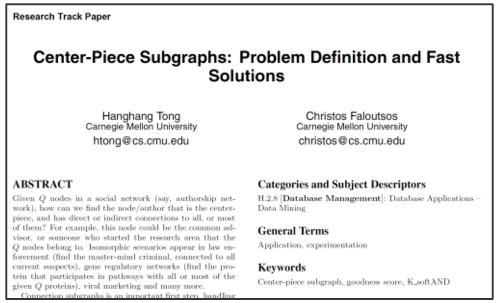


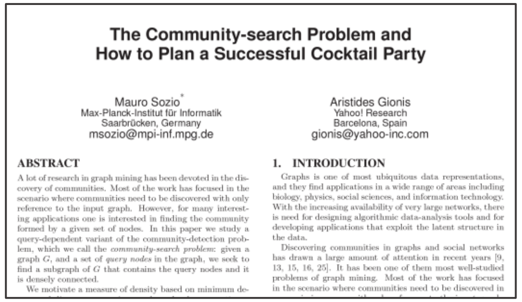
Figure 1: Comparison on the Dolphins social network: query vertices are in blue, added vertices are in green.


Research Track Paper

Center-Piece Subgraphs: Problem Definition and Fast Solutions

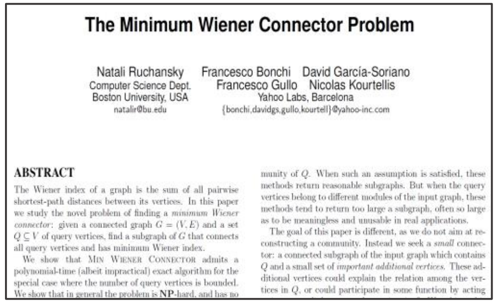
Hanghang Tong, Carnegie Mellon University, htong@cs.cmu.edu
Christos Faloutsos, Carnegie Mellon University, christos@cs.cmu.edu

KDD'06


The Community-search Problem and How to Plan a Successful Cocktail Party

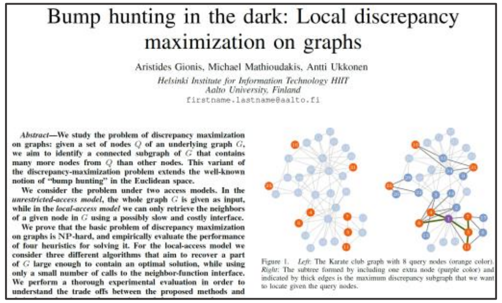
Mauro Sozio, Max-Planck-Institut für Informatik, Saarbrücken, Germany, msozio@mpi-inf.mpg.de
Aristides Gionis, Yahoo! Research, Barcelona, Spain, gionis@yahoo-inc.com

KDD'10


The Minimum Wiener Connector Problem

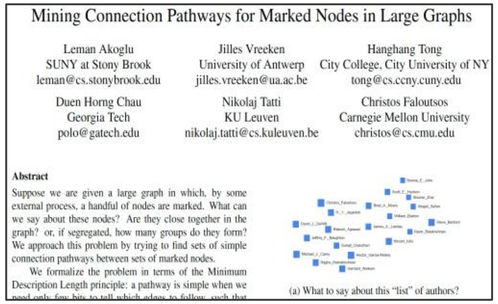
Natali Ruchansky, Computer Science Dept. Boston University, USA, natalir@bu.edu
Francesco Bonchi, David García-Soriano, Francesco Gullo, Nicolas Kourtellis, Yahoo Labs, Barcelona, {bonchi,davidgs,gullo,kourtell}@yahoo-inc.com

SIGMOD'15


Bump hunting in the dark: Local discrepancy maximization on graphs

Aristides Gionis, Michael Mathioudakis, Antti Ukkonen
Helsinki Institute for Information Technology HIIT, Aalto University, Finland
firstname.lastname@aalto.fi

ICDE'15

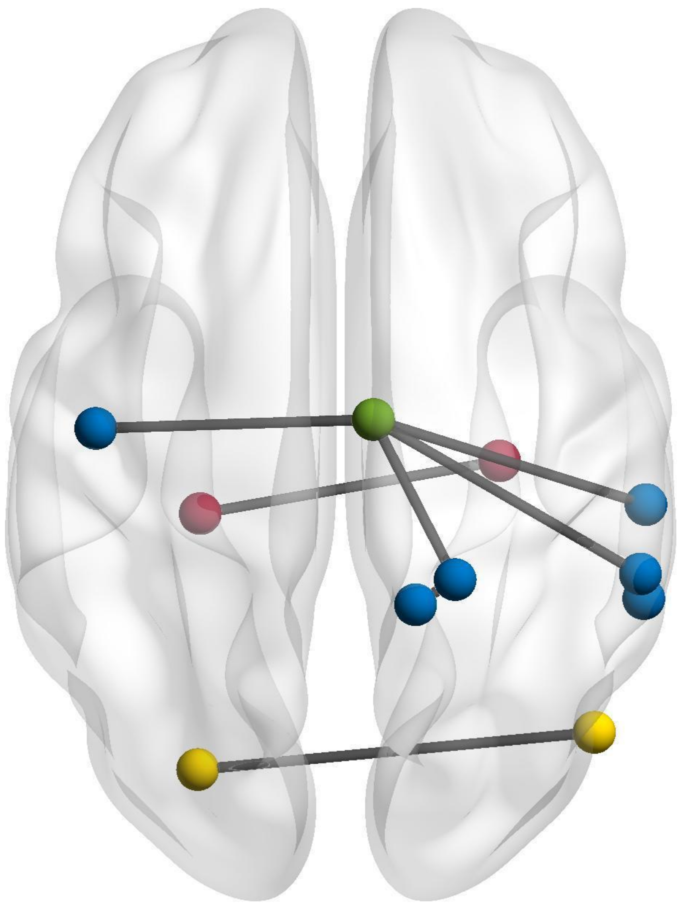

Mining Connection Pathways for Marked Nodes in Large Graphs

Leman Akoglu, SUNY at Stony Brook, leman@cs.stonybrook.edu
Jilles Vreeken, University of Antwerp, jilles.vreeken@ua.ac.be
Hanghang Tong, City College, City University of NY, tong@cs.ccny.cuny.edu
Duen Horng Chau, Georgia Tech, polo@gatech.edu
Nikolaj Tatti, KU Leuven, nikolaj.tatti@cs.kuleuven.be
Christos Faloutsos, Carnegie Mellon University, christos@cs.cmu.edu

SDM'13

Brain Co-activation Network

The data is a graph where each vertex is an area of the brain and edges are added according to co-activation in experiments. (The graph is one connected component)



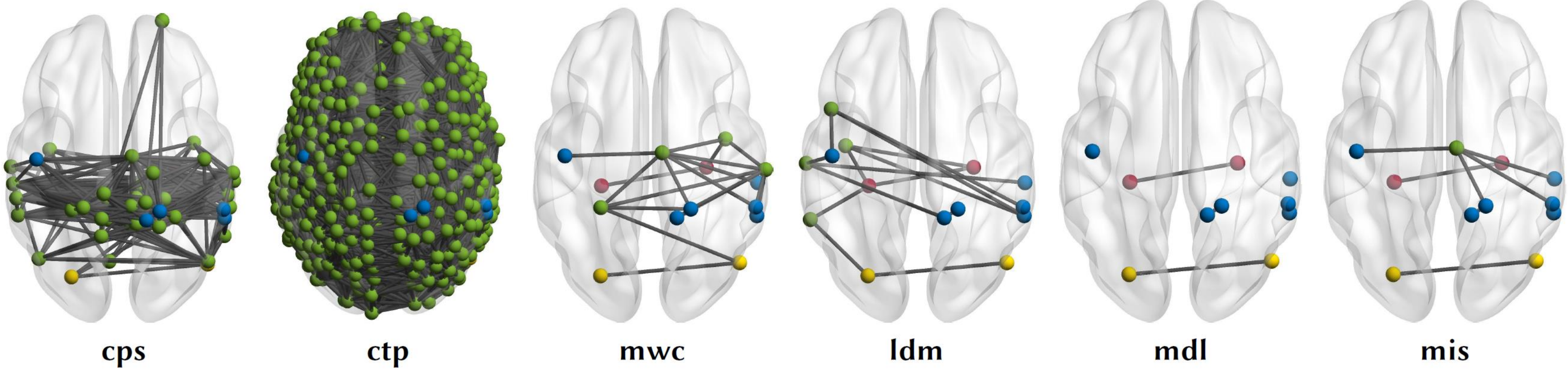
The 3 components in the solution end up corresponding to different functions: **motor**, **visual**, and **emotional**.

query vertices

extra vertices

relaxing connectivity highlights three different functional relationships and gives a smaller, more interpretable solution

Brain Co-activation Network: competitors



Experimental Results

Parsimonious vertex addition

- vertices should be added iff they help forming a more **cohesive** subgraph

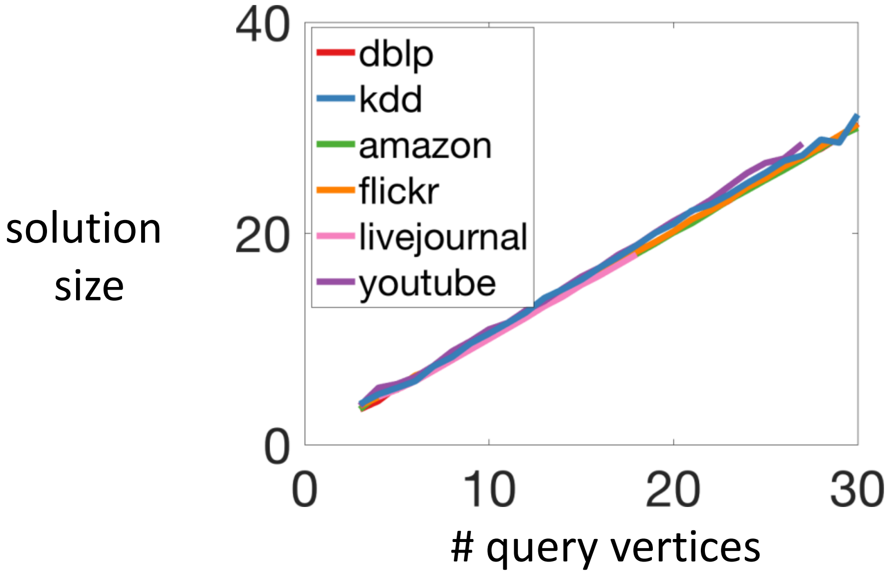
Outlier Tolerance

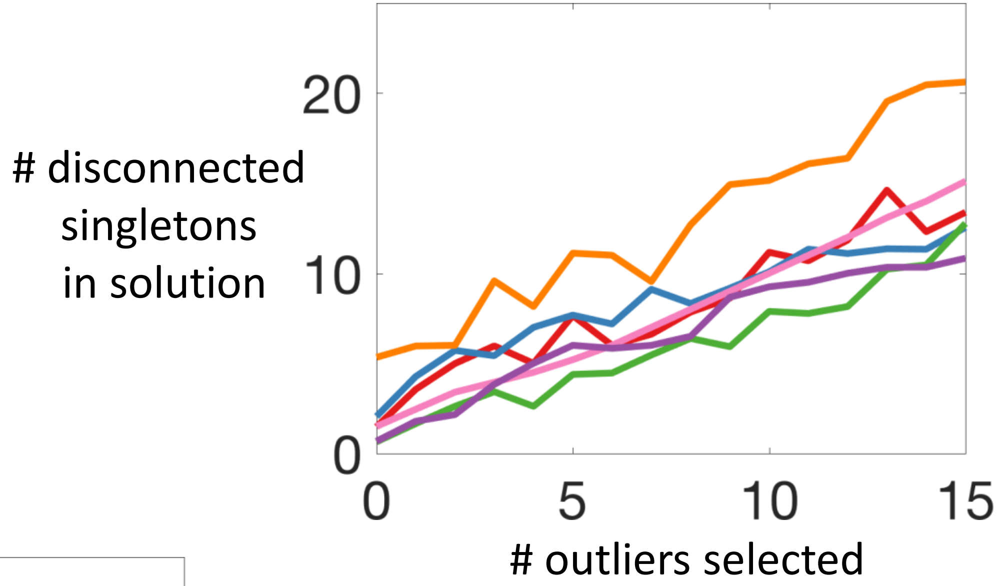
- query vertices which are far from others should remain disconnected

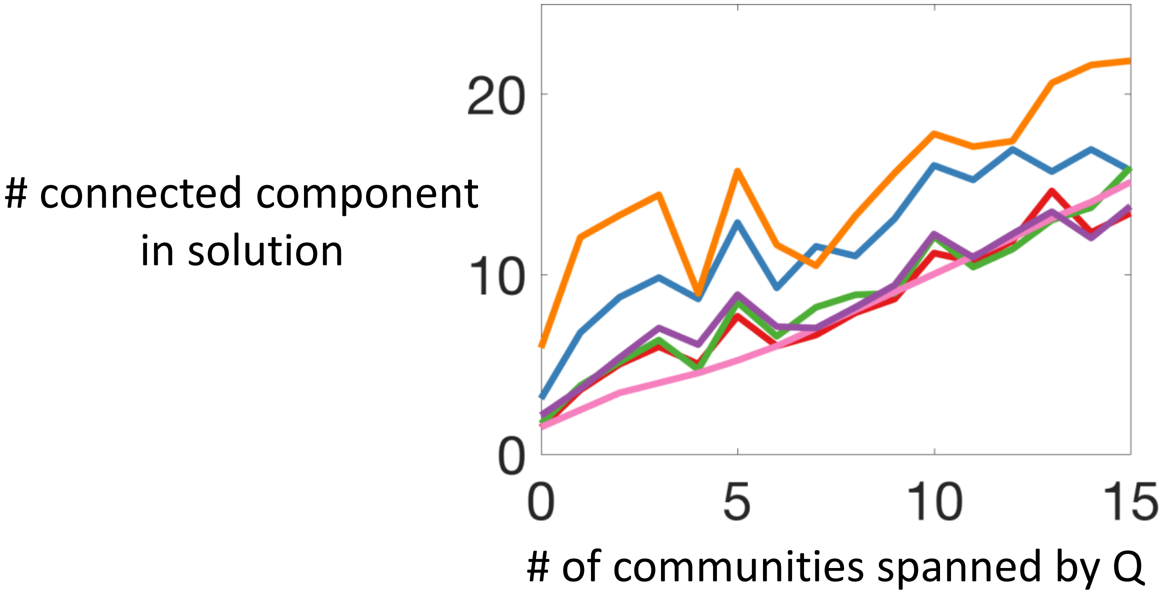
Multi-community awareness

- if the query vertices span multiple communities, connectedness should not be imposed among them

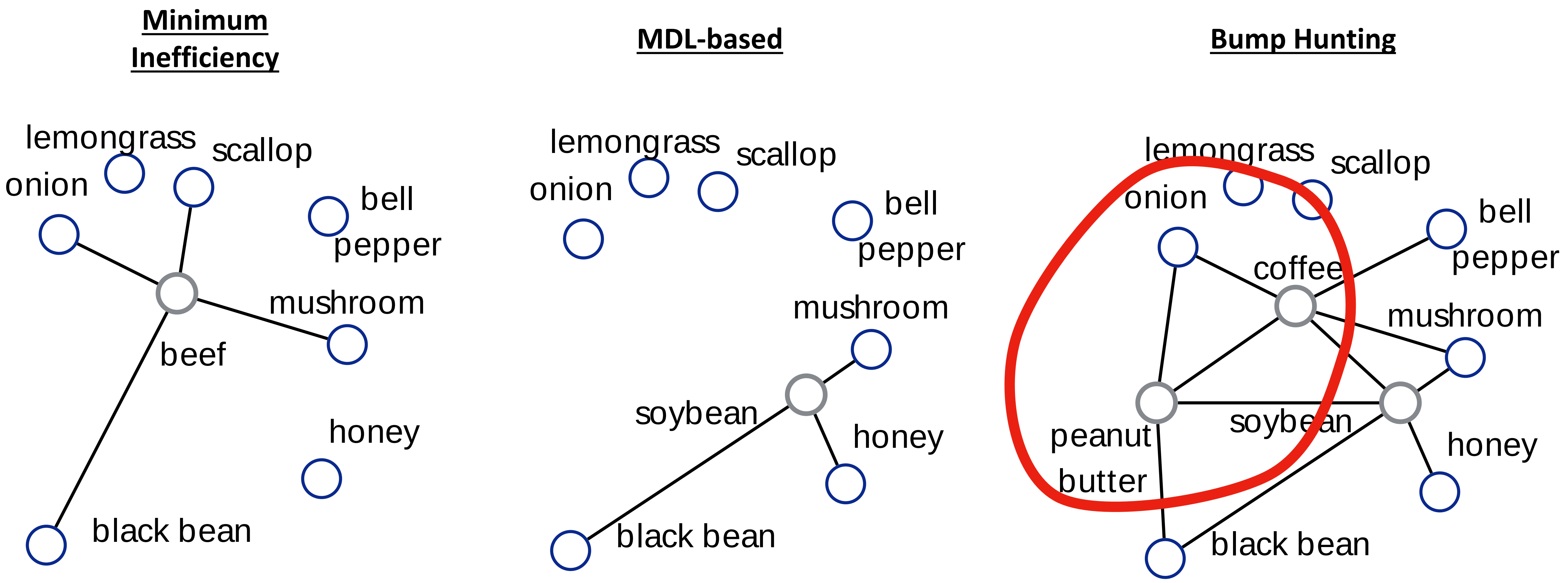
Experimental Results





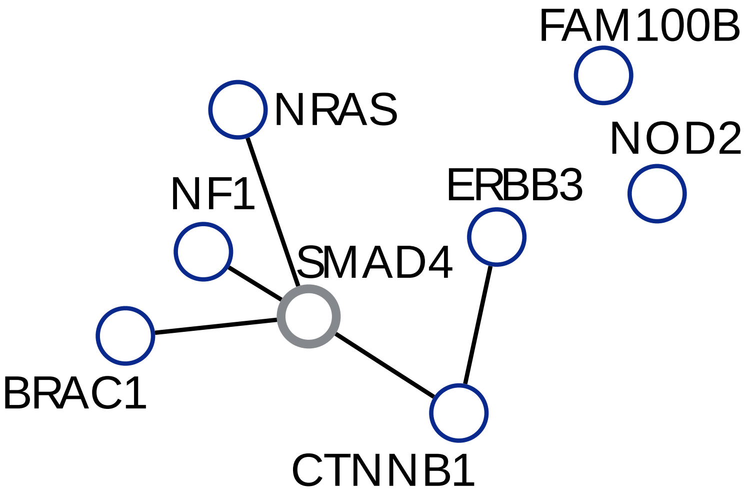


Cohesive meal creation

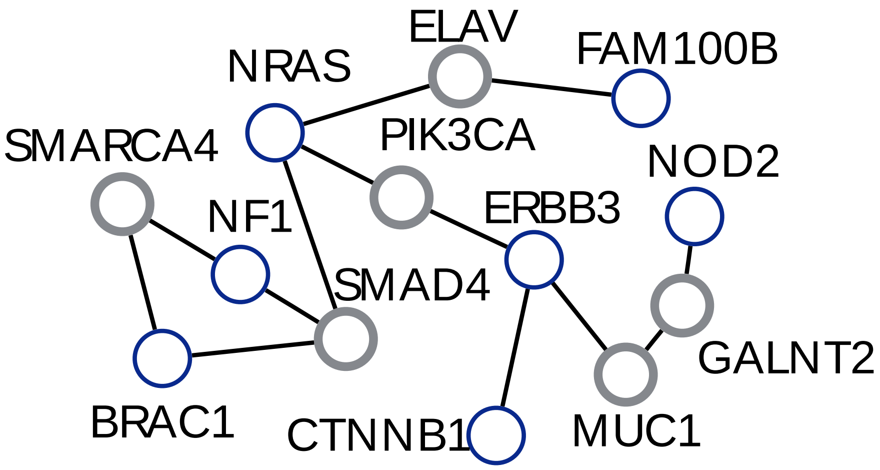


Biology

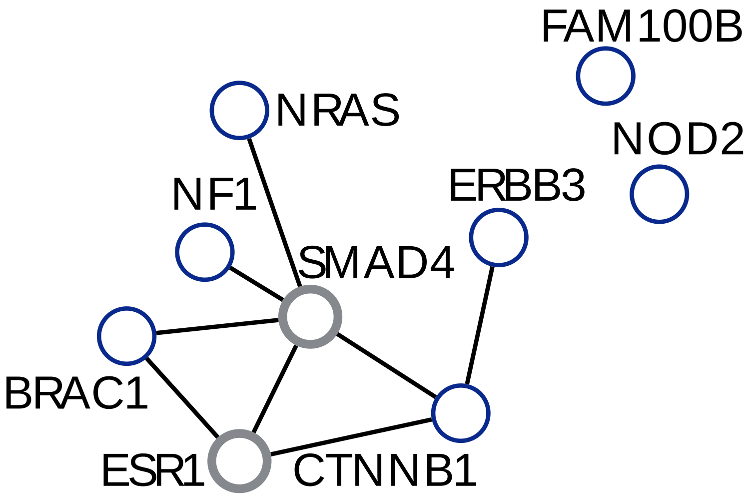
Minimum Inefficiency



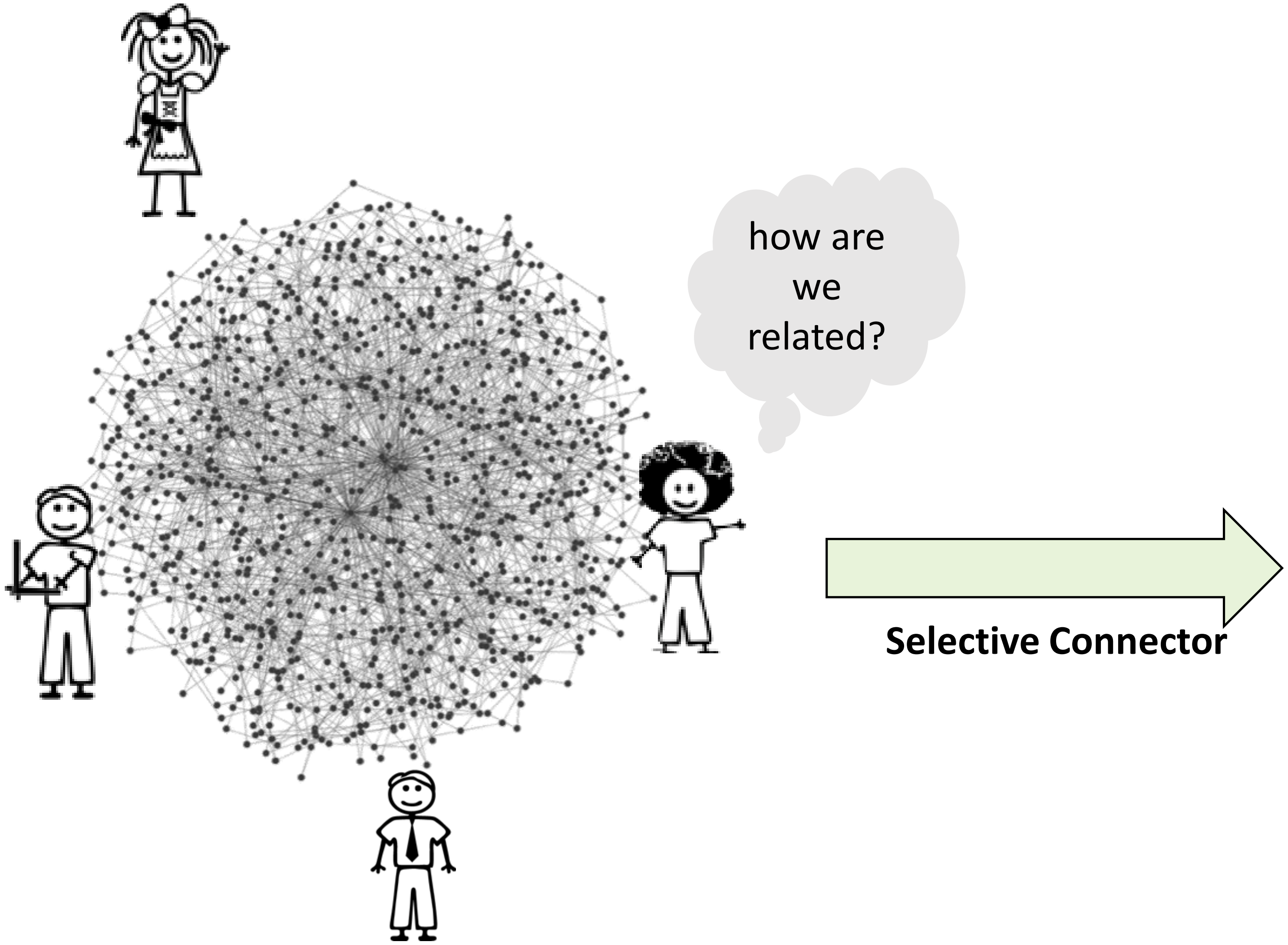
MDL-based

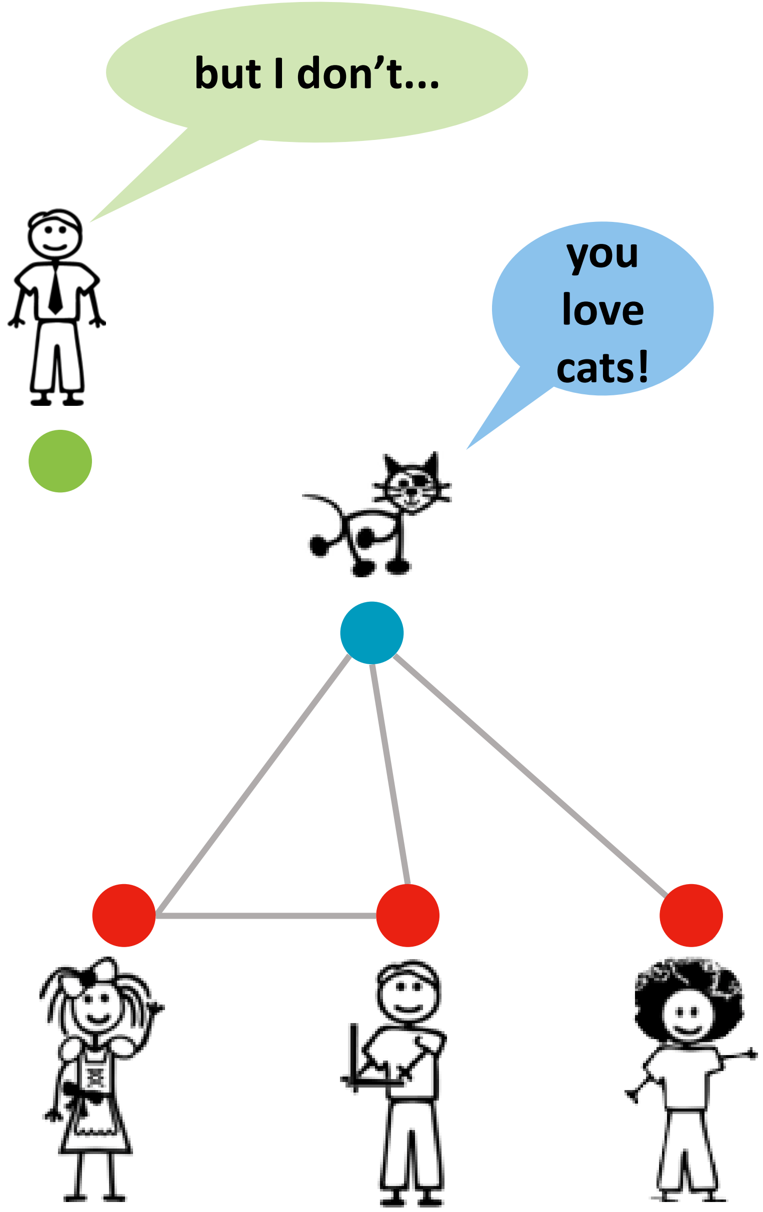


Bump Hunting



Takeaway


how are we related?
Selective Connector


but I don't...
you love cats!

Thanks!

@FrancescoBonchi
www.francescobonchi.com
francesco.bonchi@isi.it

INSTITUTE
FOR SCIENTIFIC INTERCHANGE
FOUNDATION