Parallel Processing for Large Scale Community Detection in Social and Bio-Medical Networks

Boleslaw K. Szymanski[a,c]

Konstantin Kuzmin[a], Mingming Chen[a], Chris Gaiteri[b]

[a]NeST Center & SCNARC, RPI, Troy, NY

[b]Rush Medical College, Rush University, Chicago, IL

[c]Engine EU Project, Wroclaw University of Technology, Poland

PPAM, Krakow, Poland, September 7, 2015

Complex Systems

Complex

[adj., v. kuh m-pleks, kom-pleks; n. kom-pleks]

–adjective

1.

composed of many interconnected parts; compound; composite: a complex highway system.

2.

characterized by a very complicated or involved arrangement of parts, units, etc.: complex machinery.

3.

so complicated or intricate as to be hard to understand or deal with: a complex problem.

Source: Dictionary.com

Complexity, a **scientific theory** which asserts that some systems display behavioral phenomena that are completely inexplicable by any conventional analysis of the systems' constituent parts. These phenomena, commonly referred to as emergent behaviour, seem to occur in many complex systems involving living organisms, such as a stock market or the human brain.

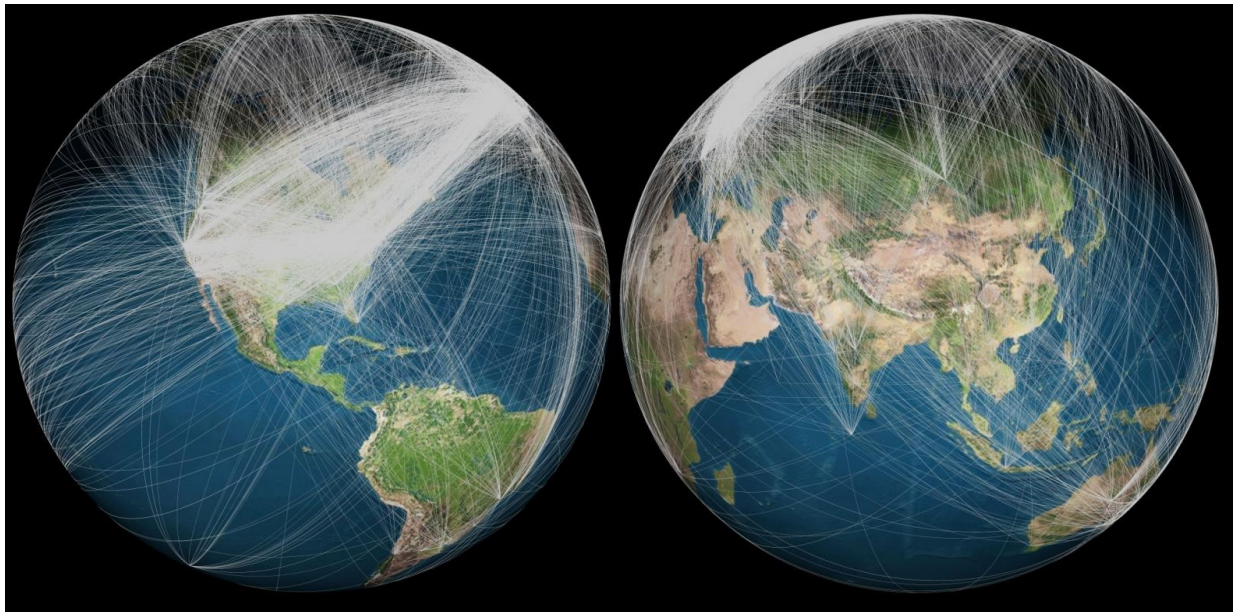
Source: John L. Casti, Encyclopædia Britannica

Complexity

SCNARC

RENSSELAER POLYTECHNIC INSTITUTE 1824

Networks of Complex Interactions



Behind every complex system there is a **network**, that defines the interactions between the components.

Facebook, a global social network

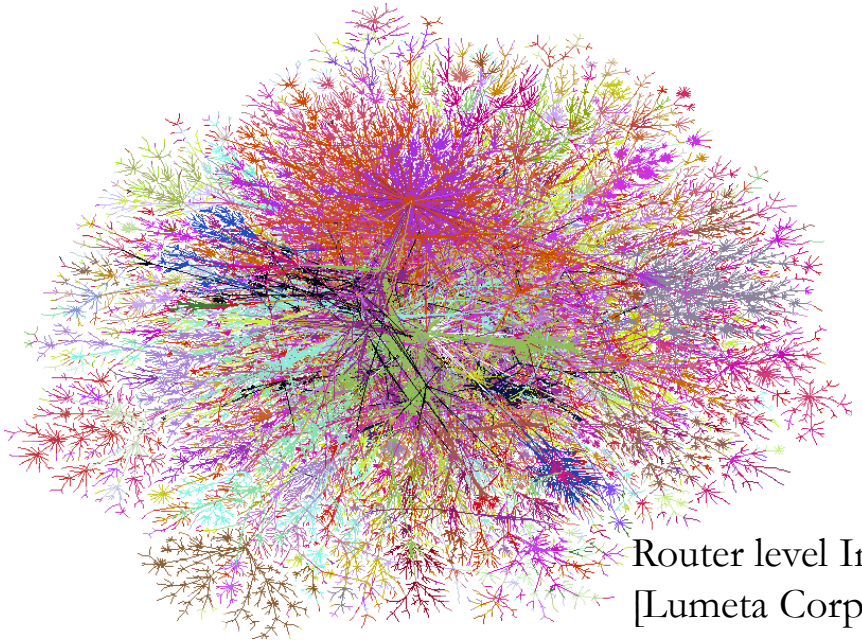
We will never **understand complex system** unless we map out and **understand the networks** behind them.

A.-L. Barabasi, *Network Science Book Project (*2013)

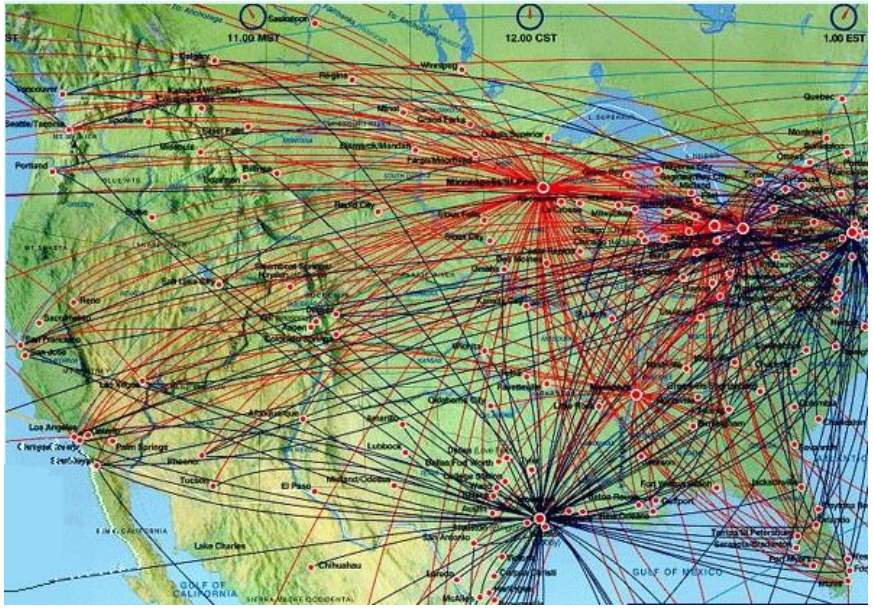
SCNARC

RENSSELAER POLYTECHNIC INSTITUTE 1824

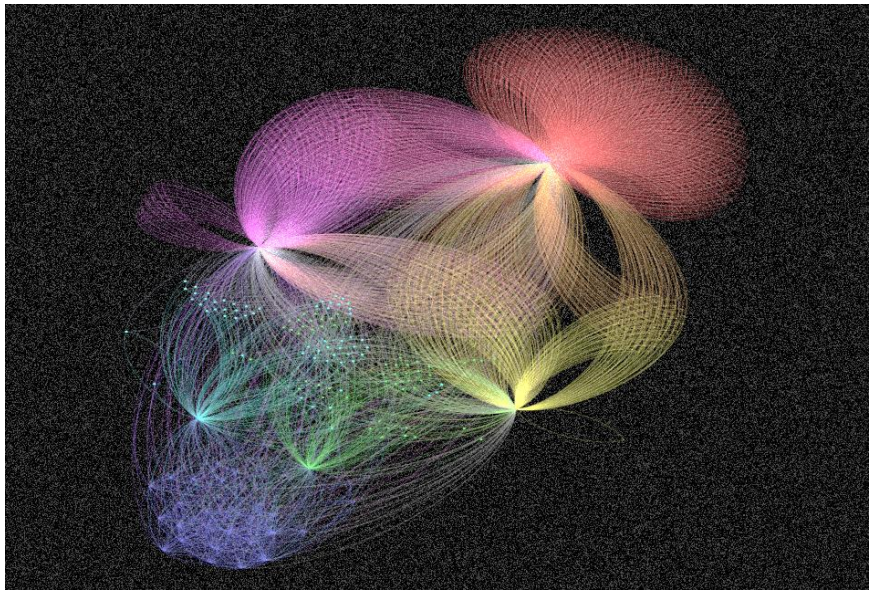
Networks Everywhere



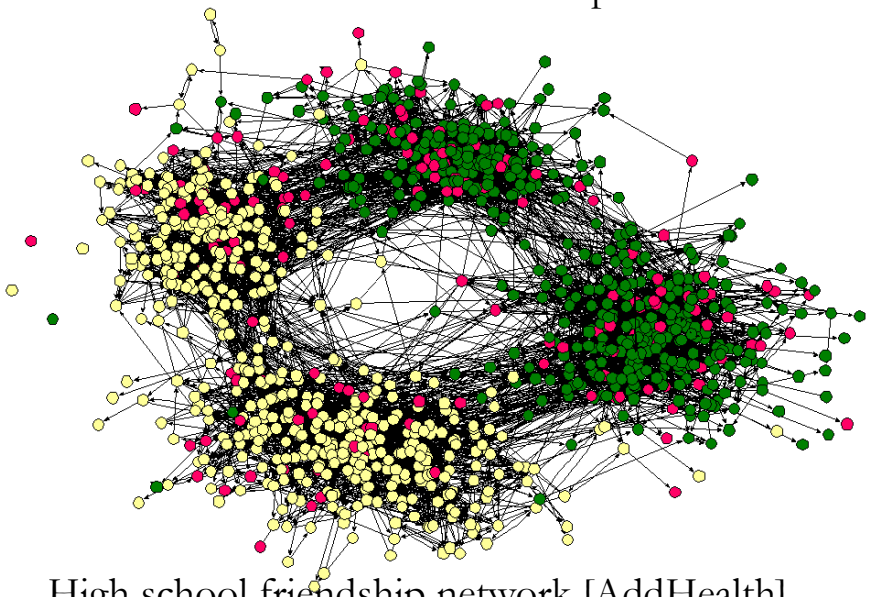
Router level Internet
[Lumeta Corp.]



airline transportation network



Communities in Gowalla Social Network
Tommy Nguyen et al., (2012)



High school friendship network [AddHealth]

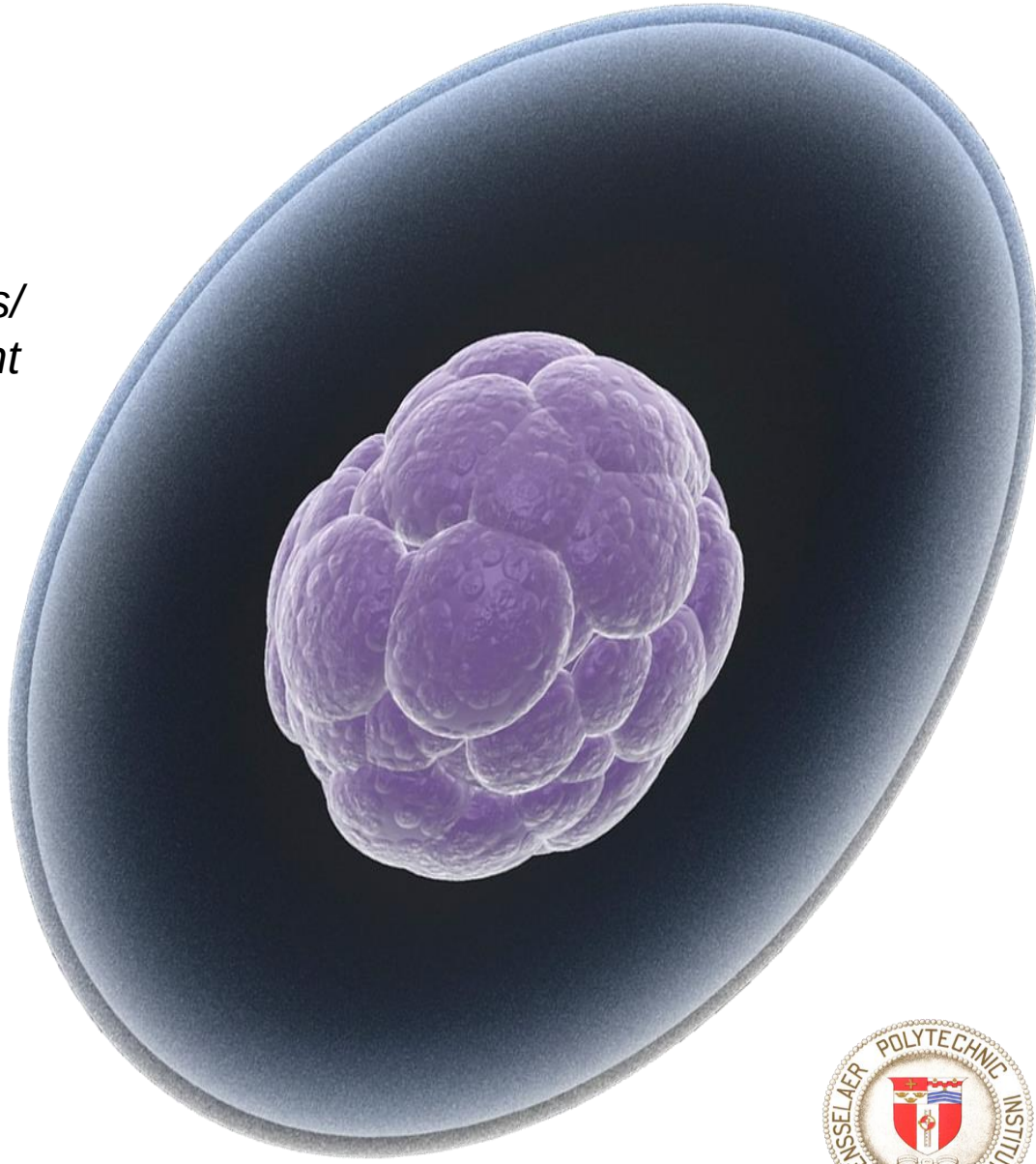
Human Genome

How Many Genes are in the Human Genome?

23,299

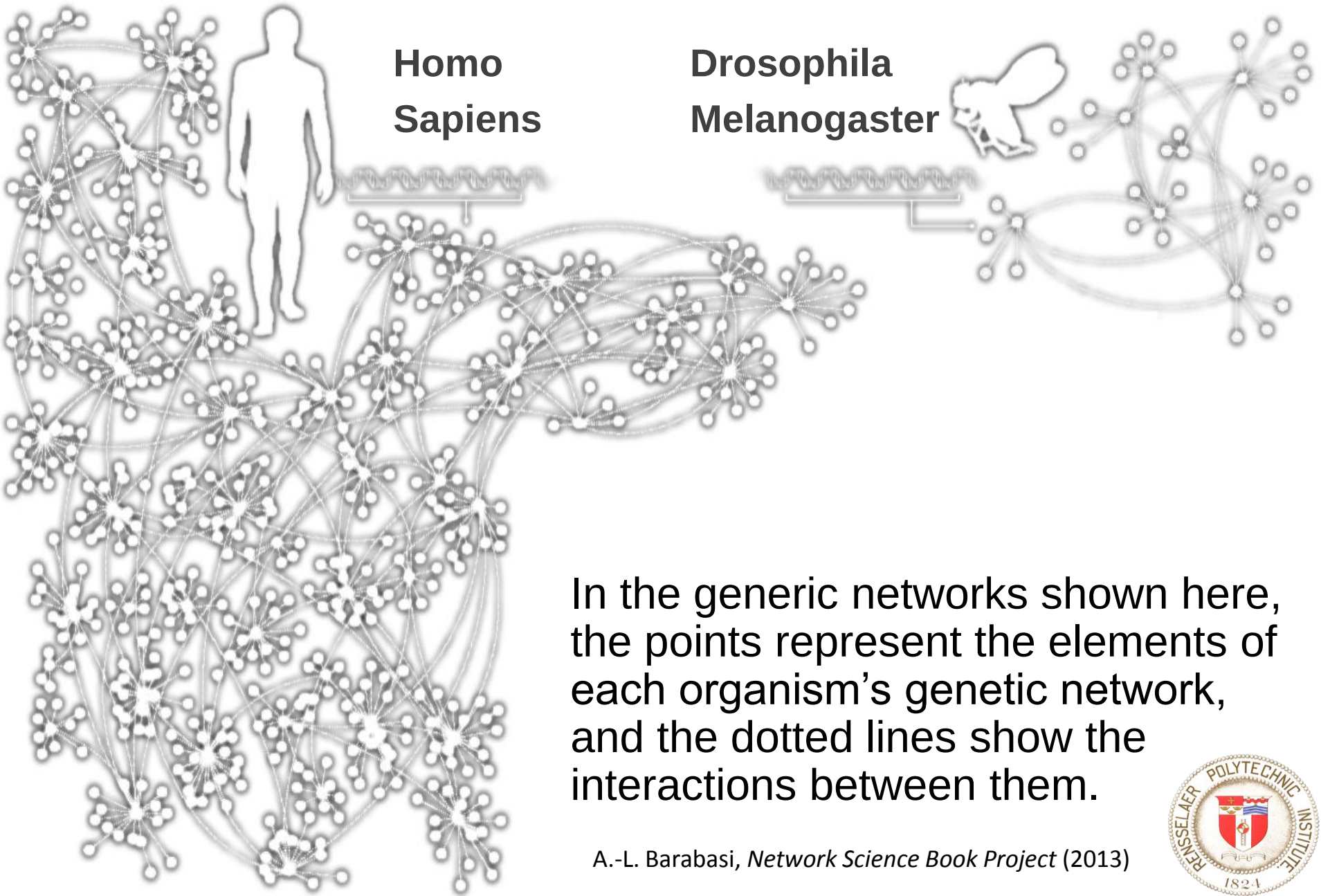
http://www.ornl.gov/sci/techresources/Human_Genome/faq/genenumber.shtml

Humans have only about three times as many genes as the fly, so human complexity seems unlikely to come from a sheer quantity of genes. Rather, some scientists suggest, each human has a network with different parts like genes, proteins and groups.



A.-L. Barabasi, *Network Science Book Project* (2013)

Human Genes



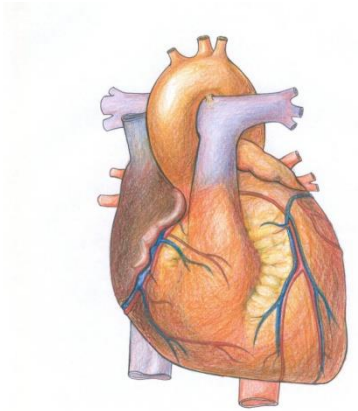
In the generic networks shown here, the points represent the elements of each organism's genetic network, and the dotted lines show the interactions between them.

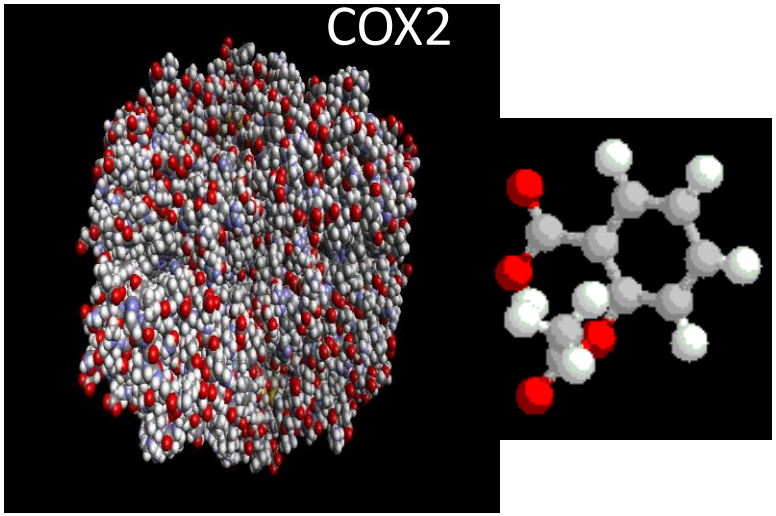
A.-L. Barabasi, *Network Science Book Project* (2013)

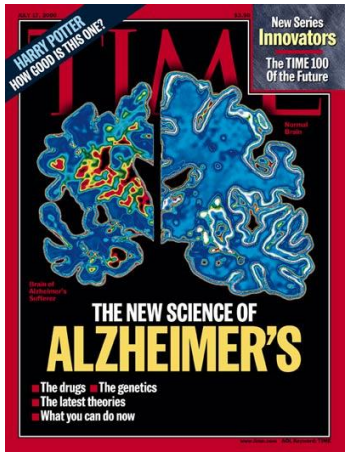
Impact on Drug Design, Metabolic Engineering

Reduces
Inflammation
Fever
Pain

Prevents
Heart attack
Stroke







Reduces the risk of
Alzheimer's Disease

Reduces the risk of
breast cancer
ovarian cancers
colorectal cancer

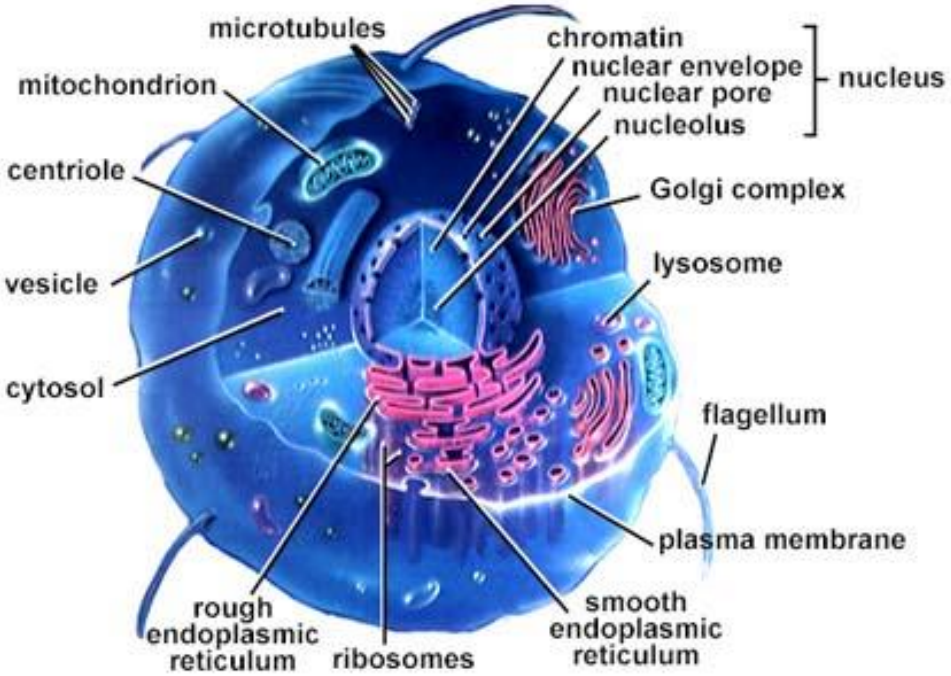
Causes
Bleeding
Ulcer

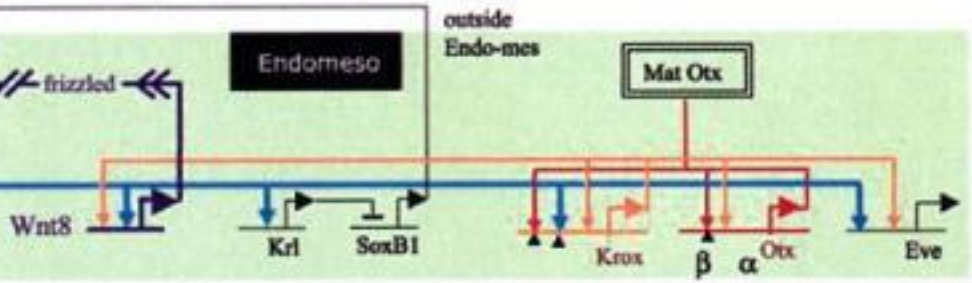
SCNARC

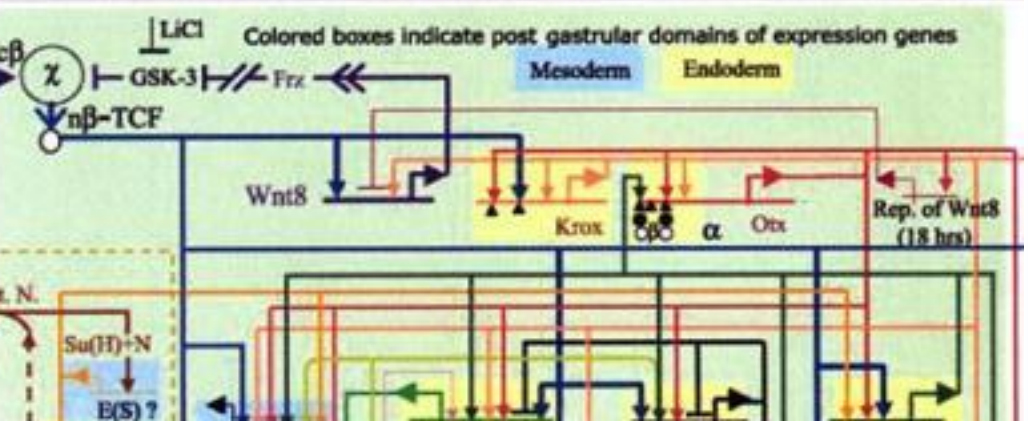
A.-L. Barabasi, *Network Science Book Project* (2013)

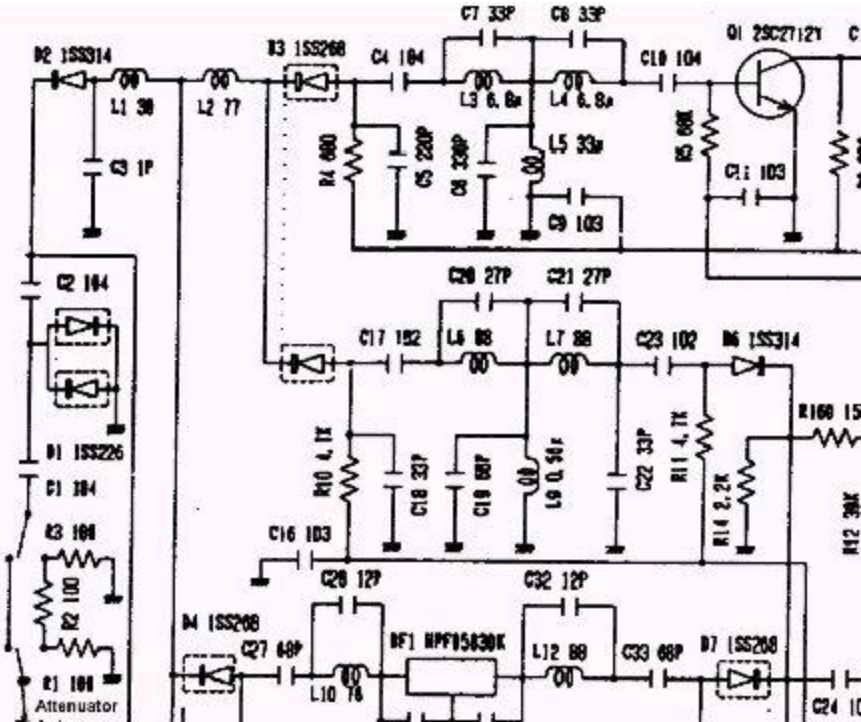
RENSSELAER POLYTECHNIC INSTITUTE 1824

A.-L. Barabasi, *Network Science Book Project* (2013)

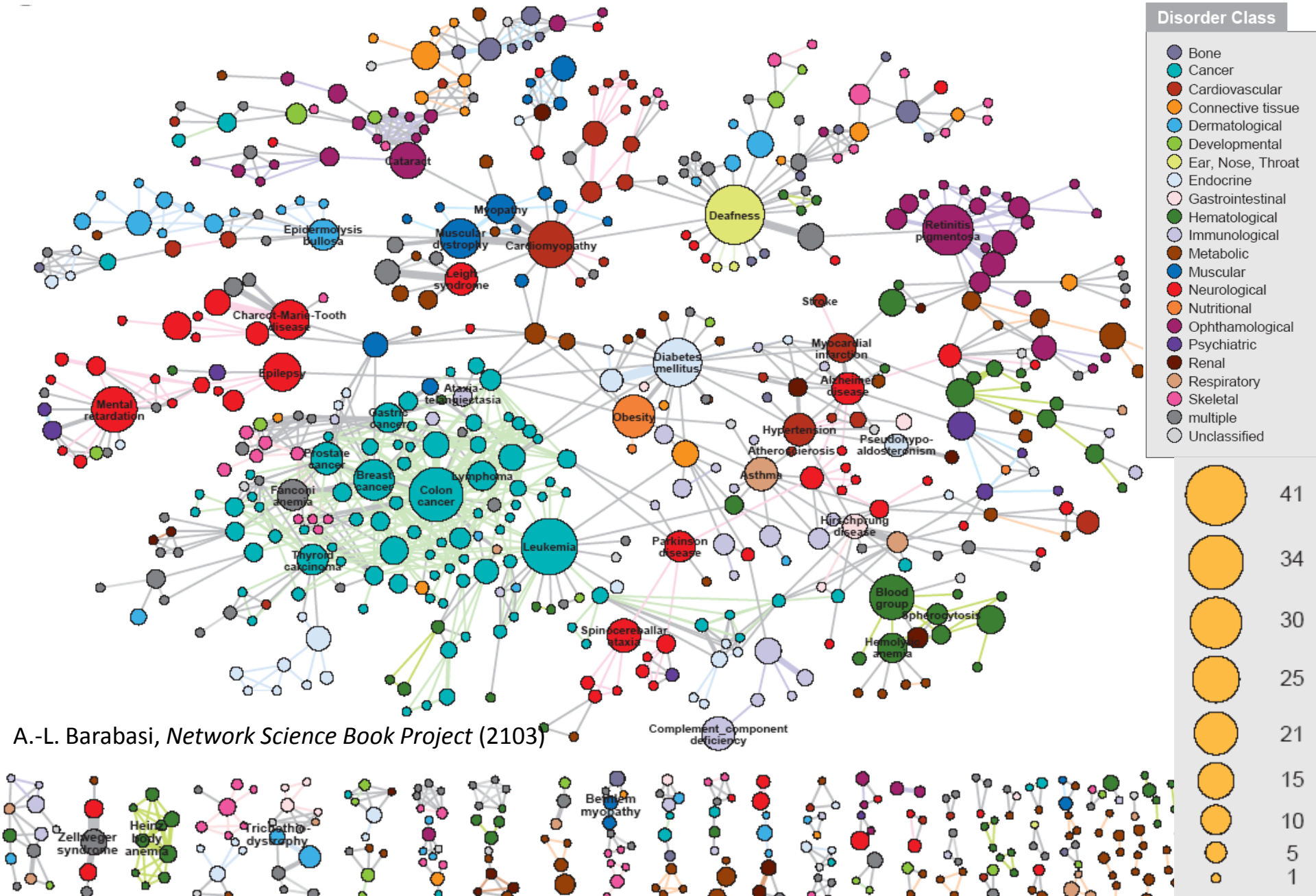






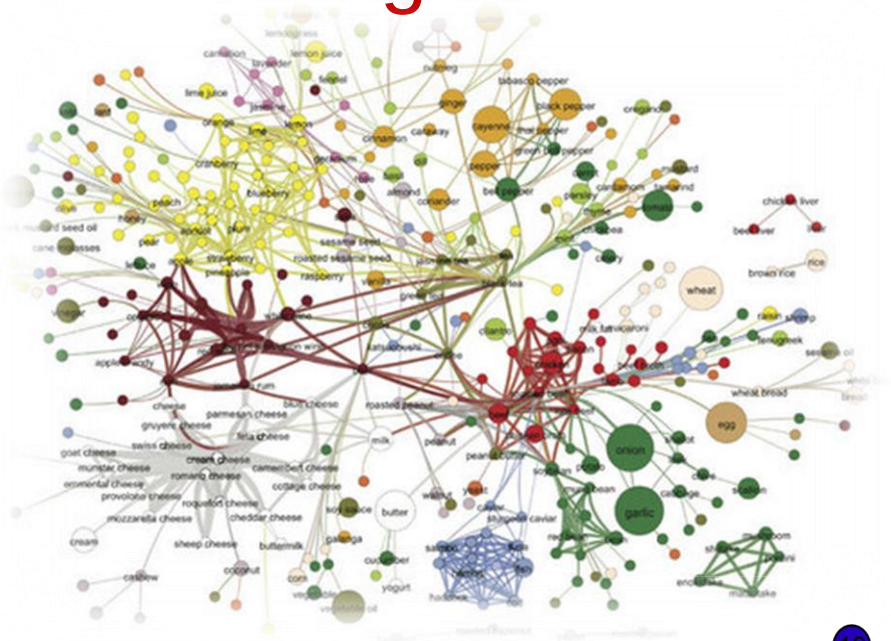


Human Disease Network



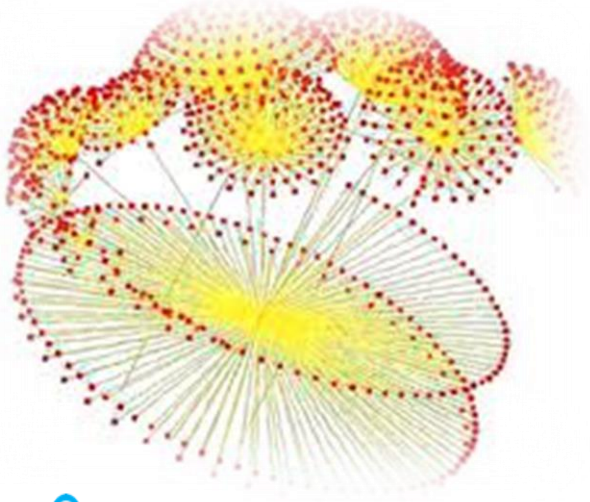
A.-L. Barabasi, *Network Science Book Project* (2103)

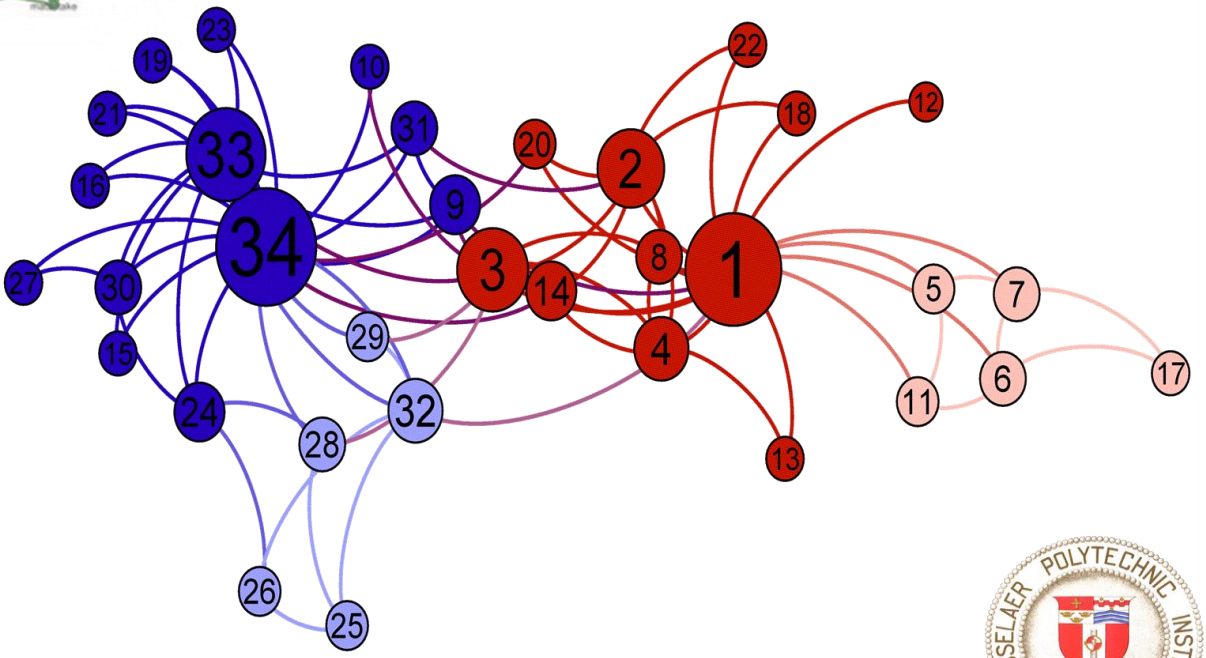
Discovering Communities in Social & Bio-networks



Clustering implies modularity
Functional modularity imposes natural boundary lines between communities.

Discovering community structure uncovers functionality





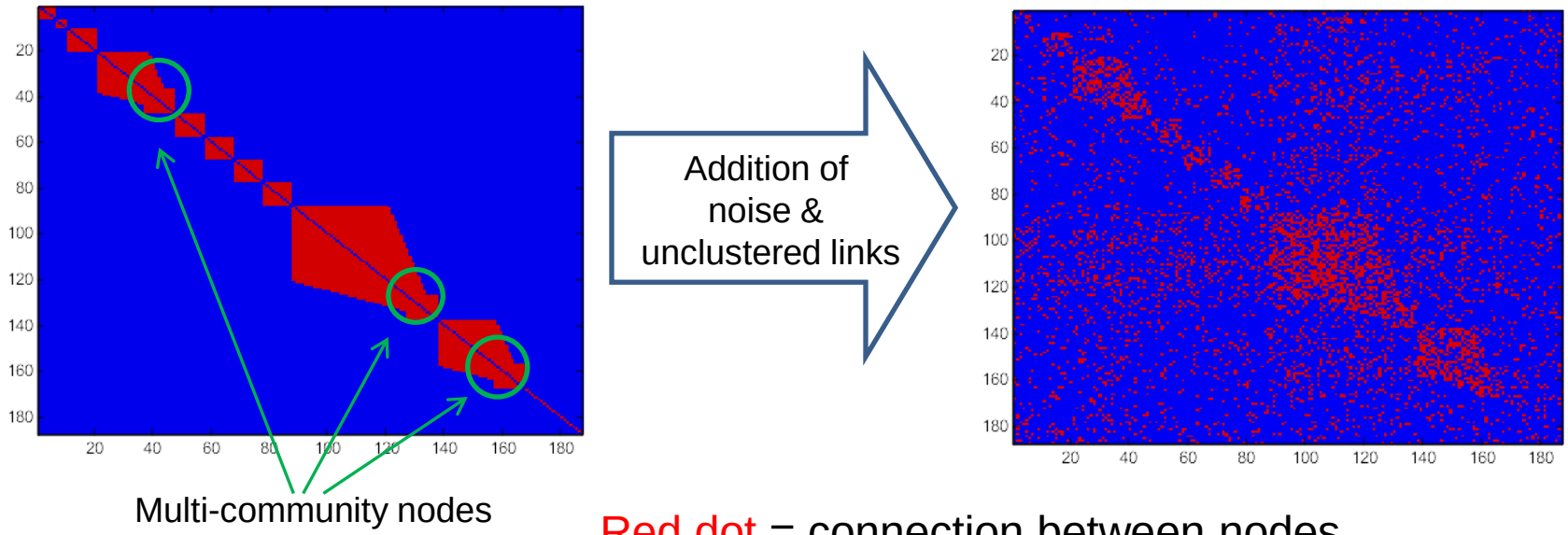
Bio (left) and social (right) networks are driven by functionality

SCNARC

RENSSELAER POLYTECHNIC INSTITUTE 1824

Motivation for Specialized Algorithms

- Biological and social networks have high level of noise and therefore have incorrect or missing links
- Biological or social functions are accomplished by communities of interacting molecules/cells or people
- Membership in these communities may overlap when humans or biological components are involved in multiple functions

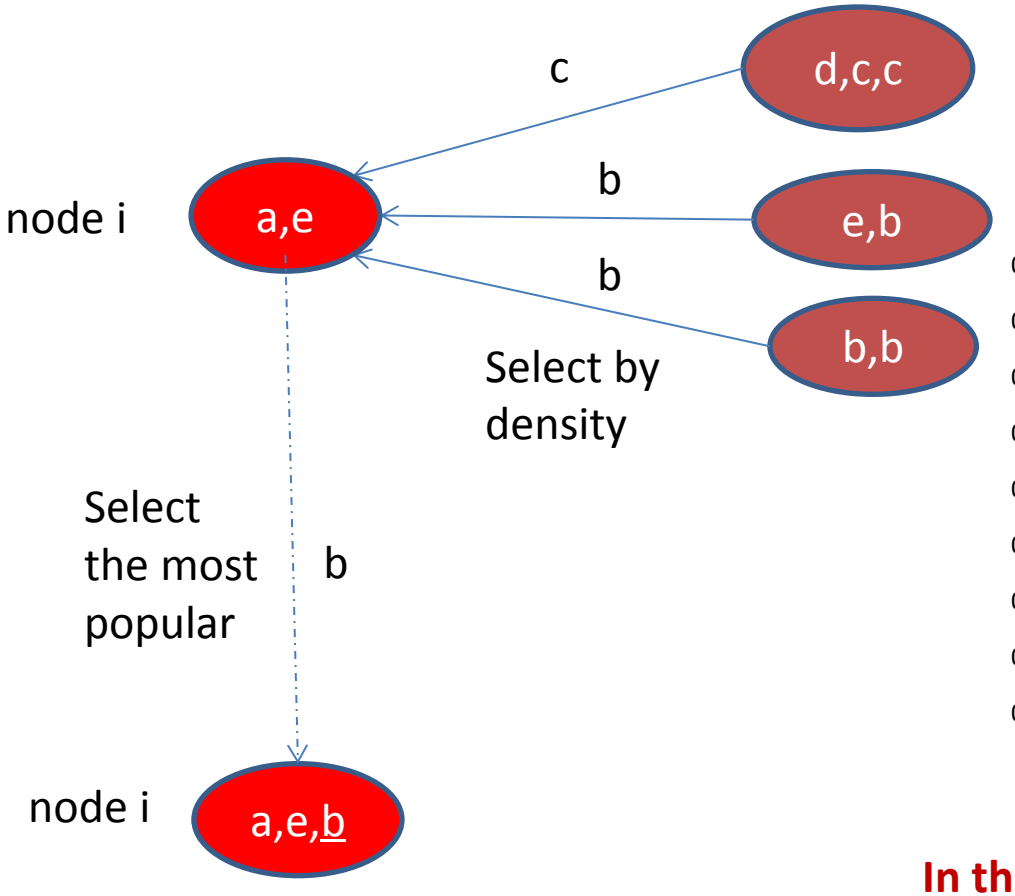


SLPA/GANXiS Community Detection Algorithm

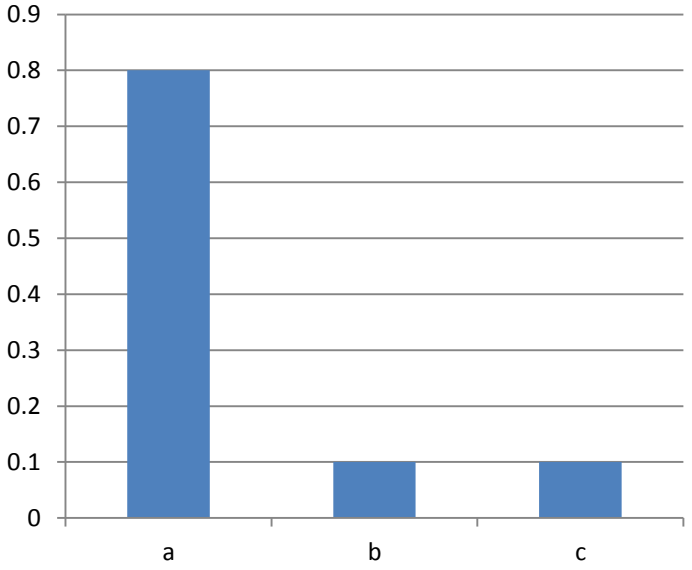
- An extension of the Label Propagation Algorithm (LPA) in which nodes send their labels to neighbors and most popular label is retained. All nodes left with the same label represent community
- **SLPA mimics pairwise interactions between nodes** (functional in bio-network, social in social networks)
- Each node broadcasts a label to its neighbors and at the same time receives a label from each of its neighbors
  - Each node has a memory of received labels which are taken into account in the next round of broadcasting
  - Linear time complexity O(m)) in the number of edges

J. Xie, B. Szymanski, *Proc. IEEE Network Science Workshop,* pp. 138 - 143. (2013).

SLPA Interaction Rules



Node i neighbors

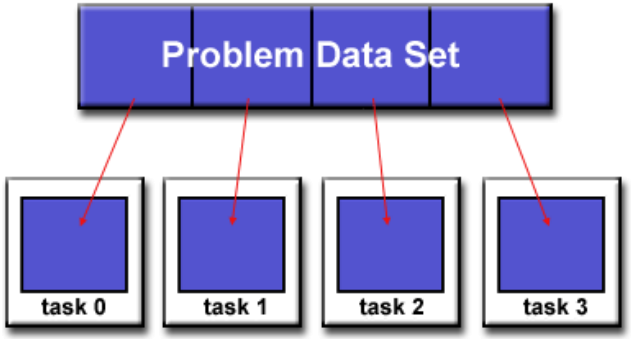


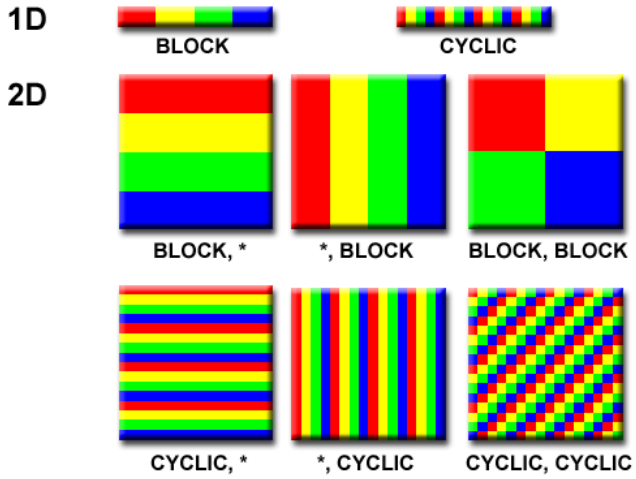
In the end, a label distribution is derived for each node.

J. Xie, S. Kelley, B. Szymanski, *ACM Computing Surveys* (2013).

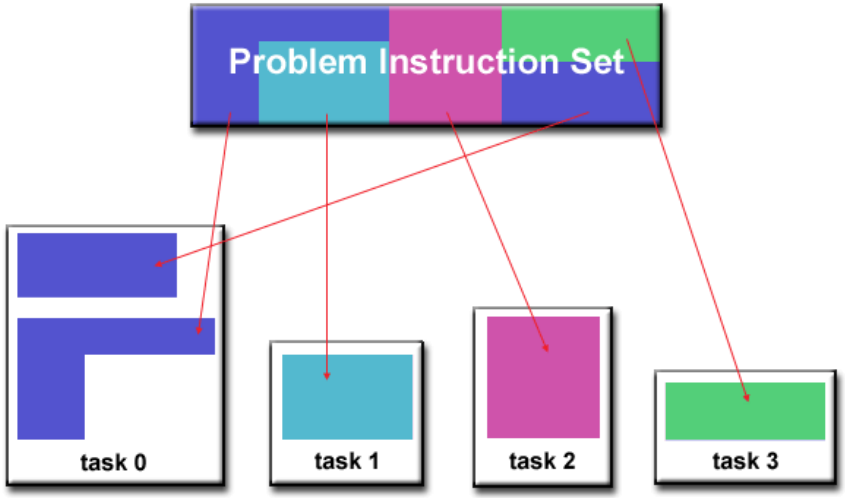
Challenges of Community Detection Parallelization Partitioning

Domain decomposition





Functional decomposition

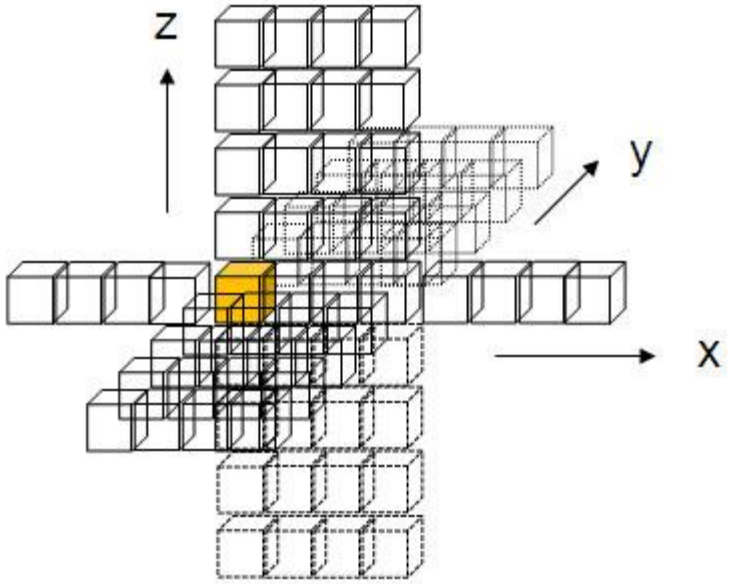


https://computing.llnl.gov/tutorials/parallel_comp/

Basic Differences in Computational Patterns

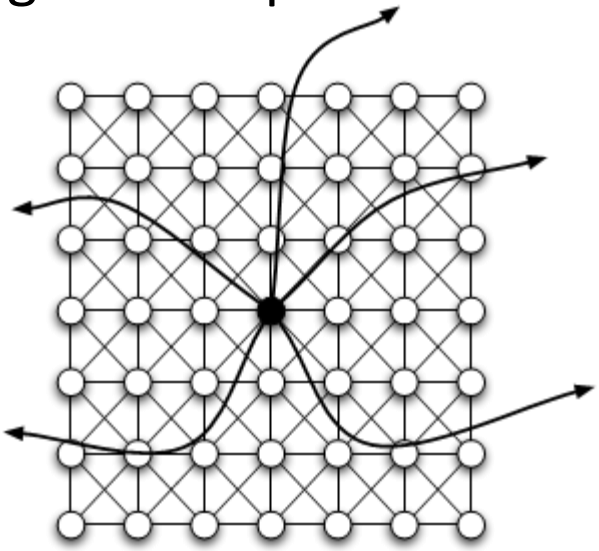
Physics Interactions

Regular grid embedded in space, all interaction are local.
Regular computational stencils



Network interactions

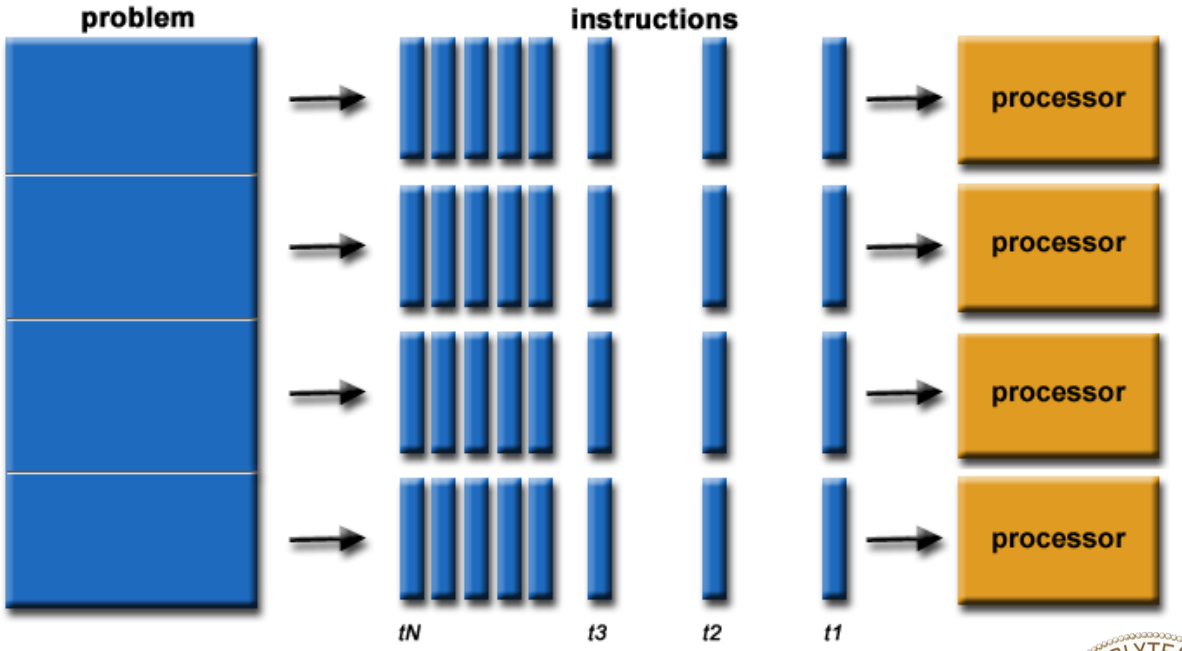
Links are independent of nodes' locations, interactions are global.
Irregular computational stencils



E. David, J. Kleinberg, *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. CUP (2010).

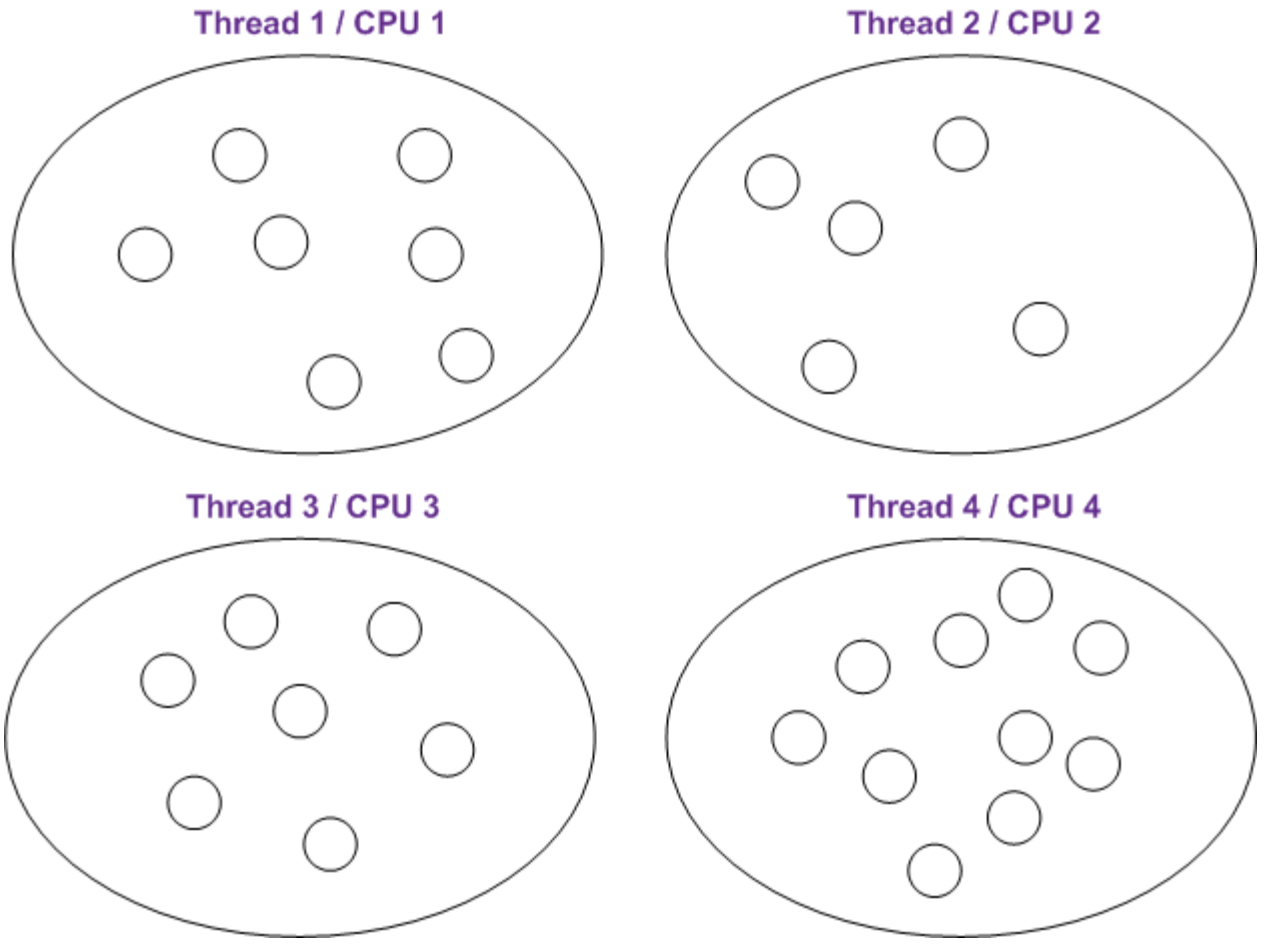
Parallel SLPA/GANXiS

- Partition the data (nodes) between processors
- Perform label propagation on each partition in parallel
- Synchronize at the end of each label propagation iteration
- Combine the results and extract communities (also in parallel)



K. Kuzmin, M. Chen, B. Szymanski, *Scientific Programming* (2015).

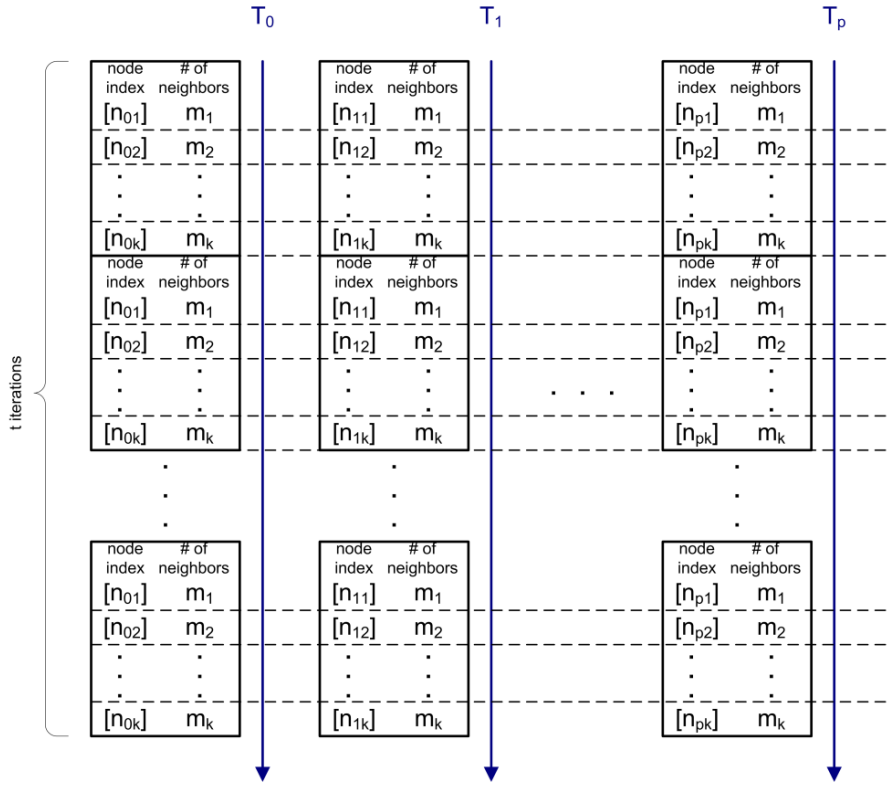
Partitioning of Network Nodes



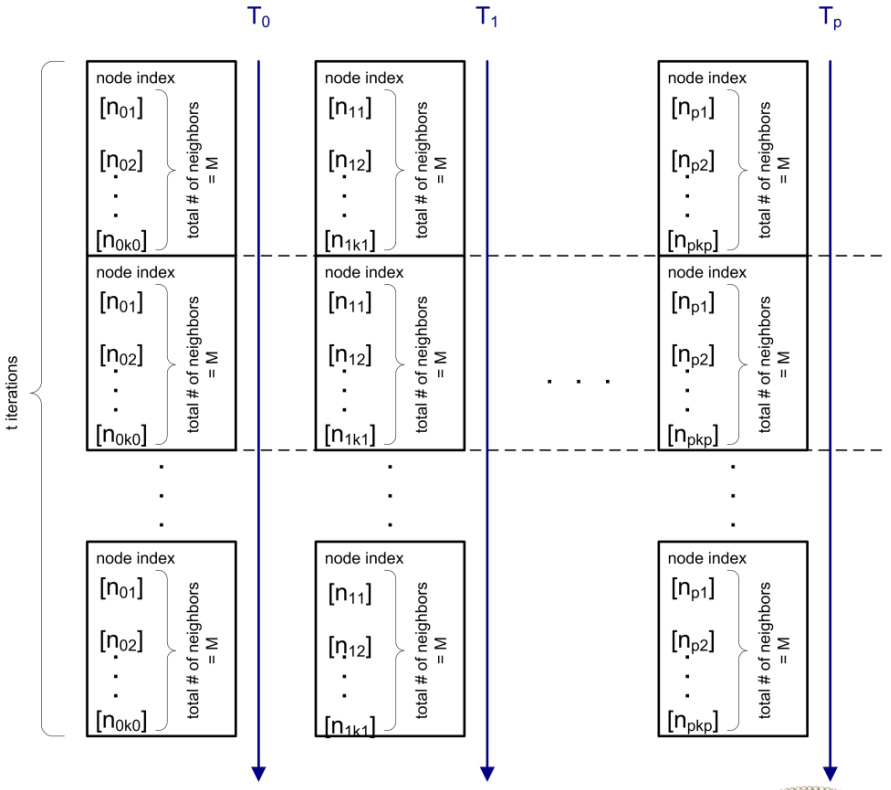
- Each thread runs on a dedicated CPU (CPU Core)
- Each thread processes a subset of nodes:
  - Each thread gets the same number of nodes
  - -or-
  - Each thread gets nodes with the same sum of degrees

Synchronization between Threads

Ideal partitioning



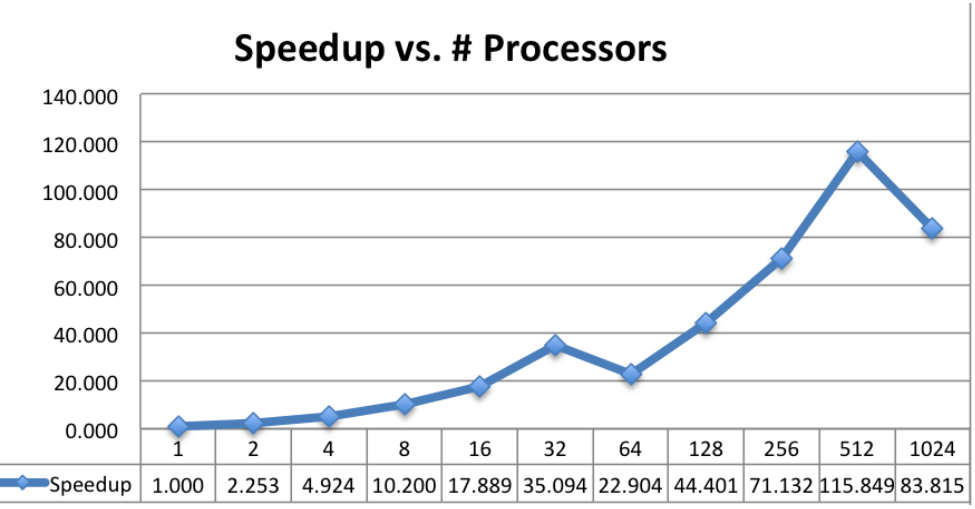
A practical partitioning



K. Kuzmin, M. Chen, B. Szymanski, *Scientific Programming* (2015).

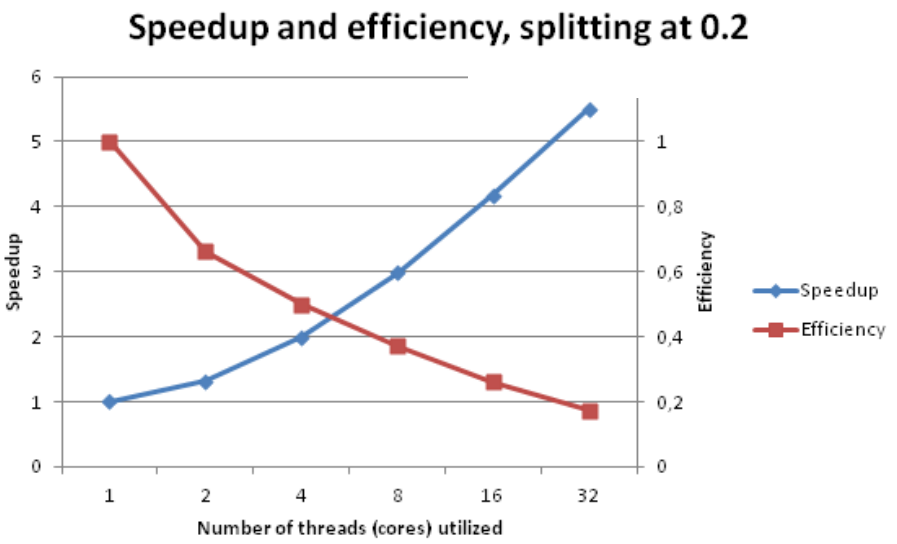
Parallel Efficiency

MPI based approach



	1	2	4	8	16	32	64	128	256	512	1024
Speedup	1.000	2.253	4.924	10.200	17.889	35.094	22.904	44.401	71.132	115.849	83.815

Multithreading approach



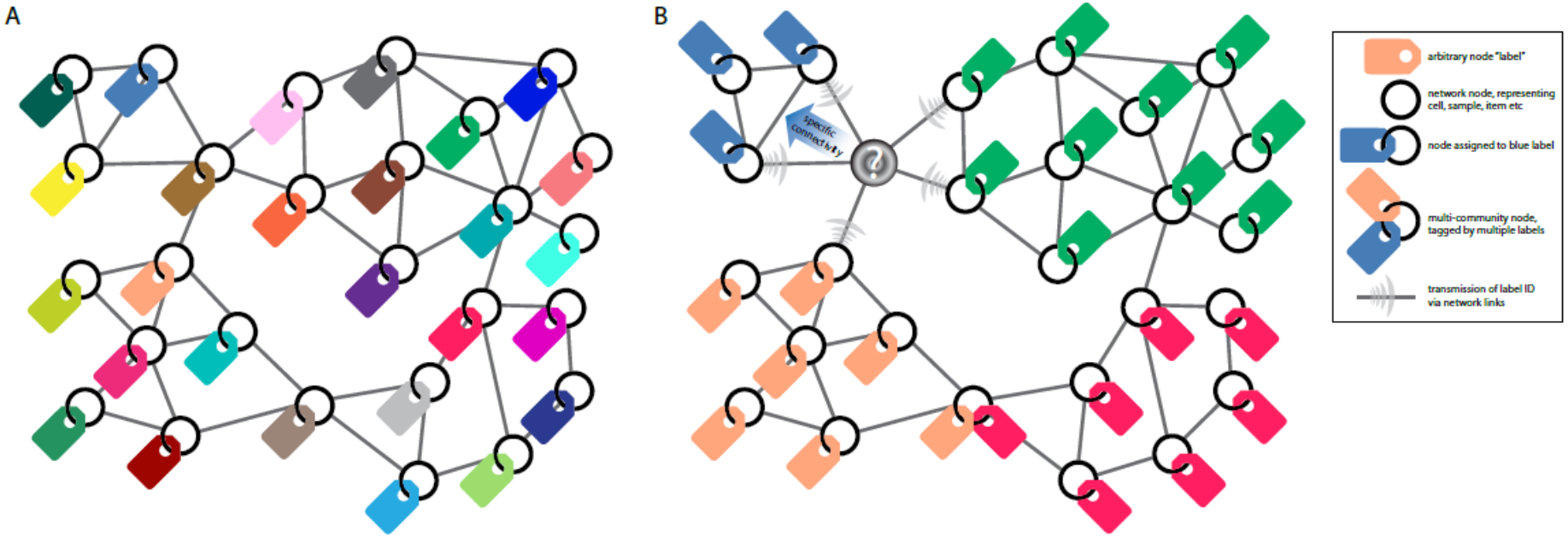
K. Kuzmin, M. Chen, B. Szymanski, *Scientific Programming* (2015).

SpeakEasy Algorithm

- **Novelty**: Identifies communities using top-down and bottom-up approaches simultaneously. Specifically, nodes join communities based on their local connections and global information about the network structure.
- **Label propagation algorithm**: each node updates its status to the label found among nodes connected to it which has the greatest specificity, i.e., the actual number of times this label is present in neighboring nodes minus its expected number based on its global frequency.
- **Consensus clustering**: the partition with the highest average adjusted Rand Index among all other partitions is selected as the representative partition to get robust community structure.
- **Overlapping communities**: overlapping communities can be obtained with co-occurrence matrix. Multi-community nodes are selected as nodes which co-occur with more than one of the final clusters with greater than a user-selected threshold. 015).

Visual Example of SpeakEasy Clustering

- Labels are represented by color tags
- Multi-community nodes are tagged with multiple colors



A. Each node is assigned with random unique label (before clustering)

B. Nodes with the same labels belong to the same community (after clustering)

C. Gaiteri, M. Chen, B.K. Szymanski, et al. ,*arXiv:*1501.04709 (2015)

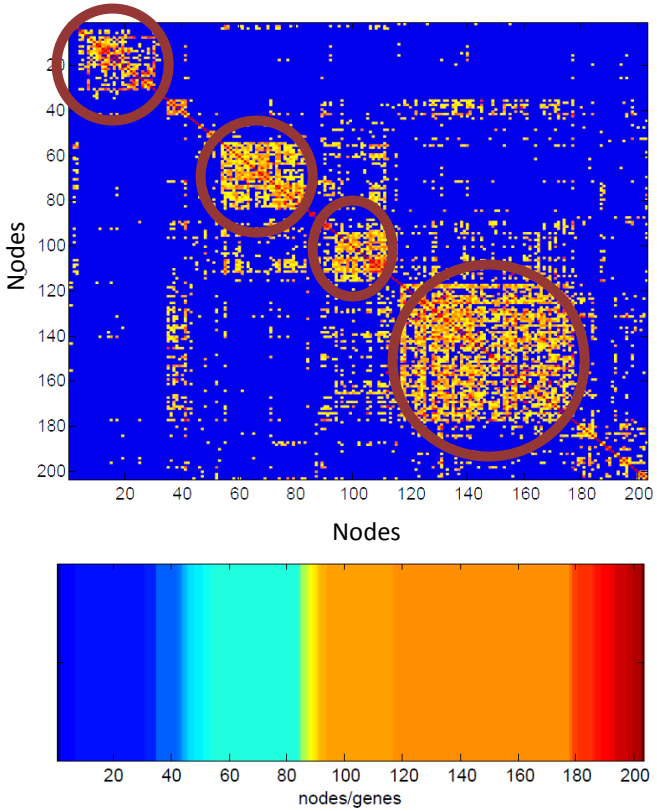
SCNARC

RENSSELAER POLYTECHNIC INSTITUTE 1824

Clustering Workflow

- Algorithm identifies communities though evolution of common labels.
- After a certain number of iterations of label propagation or if none of the nodes updates its labels in the given iteration, nodes with the same label will be clustered into the same community.
- However, because the clustering is fast and parameter-free, running the algorithm multiple times, we get an assessment of the robustness of the clusters and the identity of multi-community nodes.

Correlation matrix after clustering



Color-coded community ID

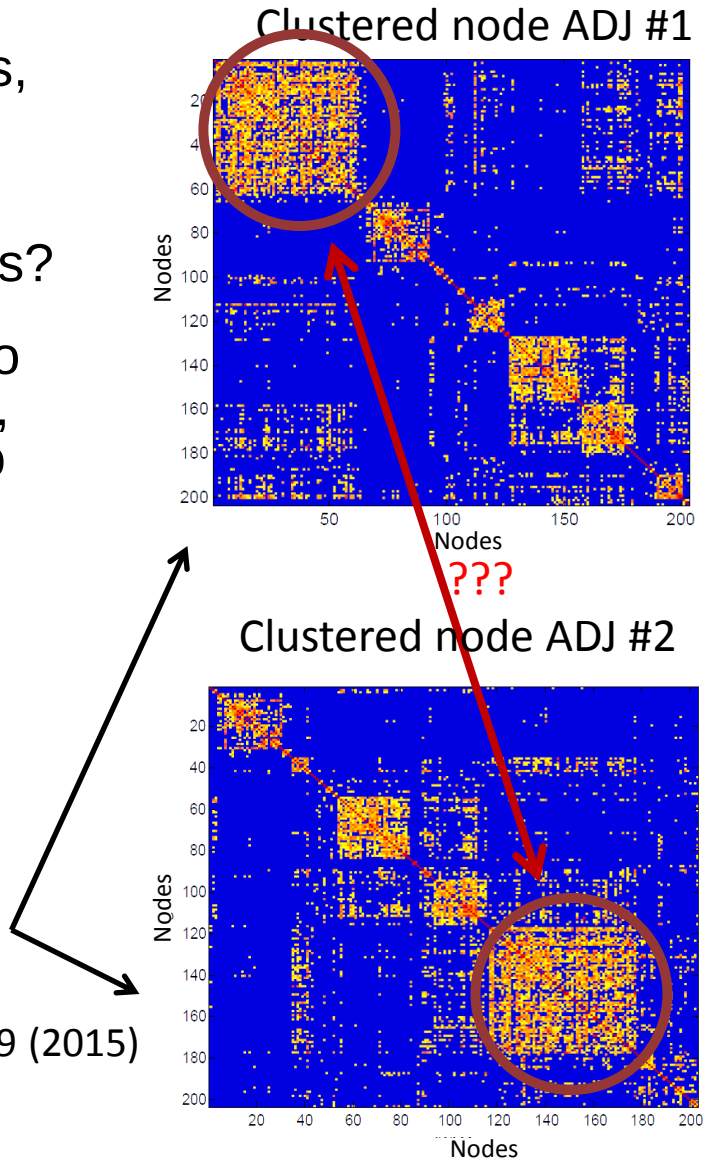
C. Gaiteri, M. Chen, B.K. Szymanski, et al. ,*arXiv:*1501.04709 (2015)

SCNARC

RENSSELAER POLYTECHNIC INSTITUTE 1824

Identifying Robust Clusters

- Individual clustering results look pretty good (dense within-community clusters, and not many between-community links.)
- However, how robust are these clusters?
- One way to test cluster robustness is to resample the data, rebuild the clusters, and compare them to the original, or to other clusters built by resampling.
- For example, how similar are the clusters from a resampled dataset?
- The sample with the highest average adjusted Rand Index among all other samples is selected as the representative sample to get robust communities.



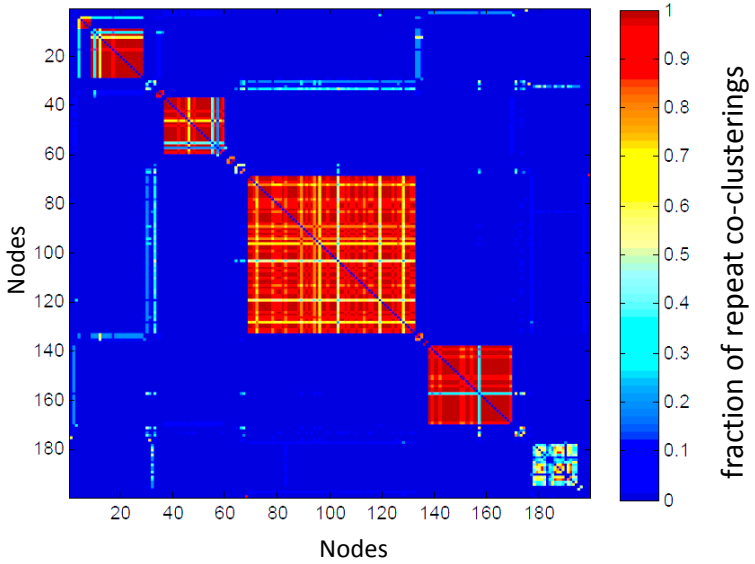
C. Gaiteri, M. Chen, B.K. Szymanski, et al. ,*arXiv:*1501.04709 (2015)

SCNARC

Identifying Multi-community Nodes

- Run SpeakEasy multiple times (e.g. 100x).
- For all pairs of nodes (*i*, *j*) the "co-occurrence" matrix records number of times they land in same cluster.
- This is useful for both identifying robust clusters and for finding nodes that link multiple communities together.

Co-occurrence matrix



Clusters in this matrix show nodes that cluster across many initial conditions

Strong non-clustered/ off-diagonal elements show multi-community nodes

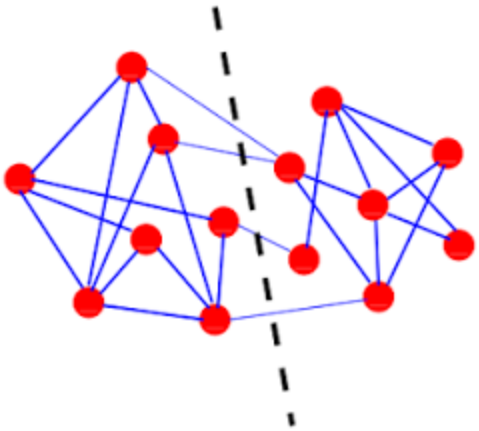
C. Gaiteri, M. Chen, B.K. Szymanski, et al. ,*arXiv:*1501.04709 (2015)

High-Level Approach to Parallelizing SpeakEasy

- Partition the data (nodes) between processors
- Perform label propagation on each partition in parallel
- Synchronize at the end of each label propagation iteration
- Exchange the global label frequencies information among the processors
- Extract community data from label histories (also in parallel)

Parallel Communication Overhead

- Sending updated label history of nodes across partitions
- Sharing the global label frequency table between all processors



Label	Frequency
1	15/575
2	1/575
3	72/575
4	3/575
5	0/575
6	12/575
…	

S. Fortunato, *Physics Reports*, pp. 75-174 (210)

Propagating Label History Updates

Scalability is affected by:

- Network properties (degree of nodes). The larger the average degree of nodes, the more likely it is that edges will go across partition cuts.
- The quality of partitioning. The larger the number of edges going across partition cuts, the more significant the parallel communication overhead.

Node Degree Considerations

If in the data scalability setting for larger networks the average node degree:

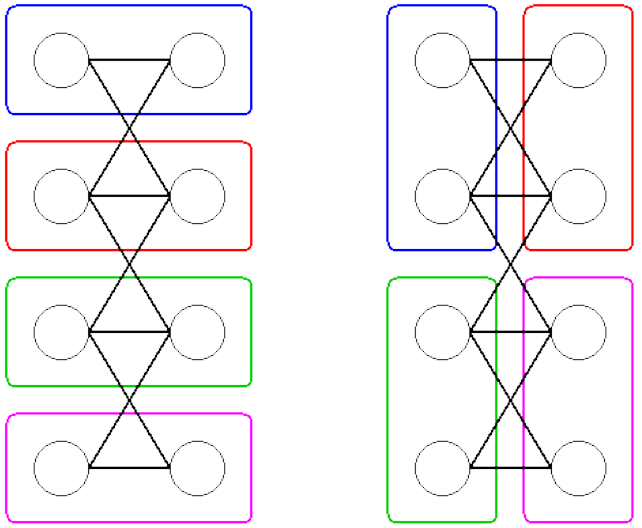
- is constant then the number of edges grows linearly with n (the number of nodes in the network) and so does the amount of computation per processor. Overall complexity O(m) = O(n). Very scalable.
- grows as log(n) then the number of edges grows as n*log(n), i.e. superlinear. Overall complexity O(m) = O(n*log(n)). Less scalable than the previous case but still practical.

Both cases happens in social and bio-medical networks

Quality of Partitioning

Ideal partitioning

- The nodes should be partitioned such that the number of edges inside a partition is maximized while the number of edges between partitions is minimized and the number of partitions which these edges reach minimized.
- Each partition should contain the same number of edges, e.g. the sum of degrees (load balancing).



for all $v \in V$ **do**
  for all $neighbor \in \mathcal{N}_v$ **do**

http://www.cs.berkeley.edu/~demmel/cs267/lecture18/lecture18.html

Global Label Frequency Table

- The total number of distinct labels is at most the number of nodes
- Most labels become extinct as processing progresses
- They are completely recalculated at every iteration of the algorithm
- Computational complexity is O(n)

Parallel Calculation of the Global Label Frequency Table

- Each processor calculates the portion of the table for the labels of all of its nodes
- Label frequencies computed by different processors are reduced to create the global table
- Each processor only needs access to the frequency of labels of all the neighbors (both inside and outside the partition) of its nodes
- Parallel communication overhead depends on the overlap of labels across different partitions and the number of edges between partitions

Sequential Fraction of SpeakEasy

Reading the input network file

- May affect the overall performance for large networks (especially if the average node degree is high)
- Can be improved in certain scenarios by using distributed or parallel I/O

Partitioning the network

- Performance penalty depends on the partitioning algorithm used
- Can be eliminated if a network is pre-partitioned

Writing the output communities file

- The size depends on the number of nodes and the degree of communities overlap
- Usually does not seriously affect the overall performance

Parallel Consensus Clustering

- Partition blocks of individual clusters of nodes (communities) between processors (requires communication)
- Compute Adjusted Rand Index (ARI) for every pair of clusters in parallel
- Determine the clustering with the highest average ARI value in parallel
- Assign each node to additional communities based on the values of the co-occurrence matrix (done in parallel)

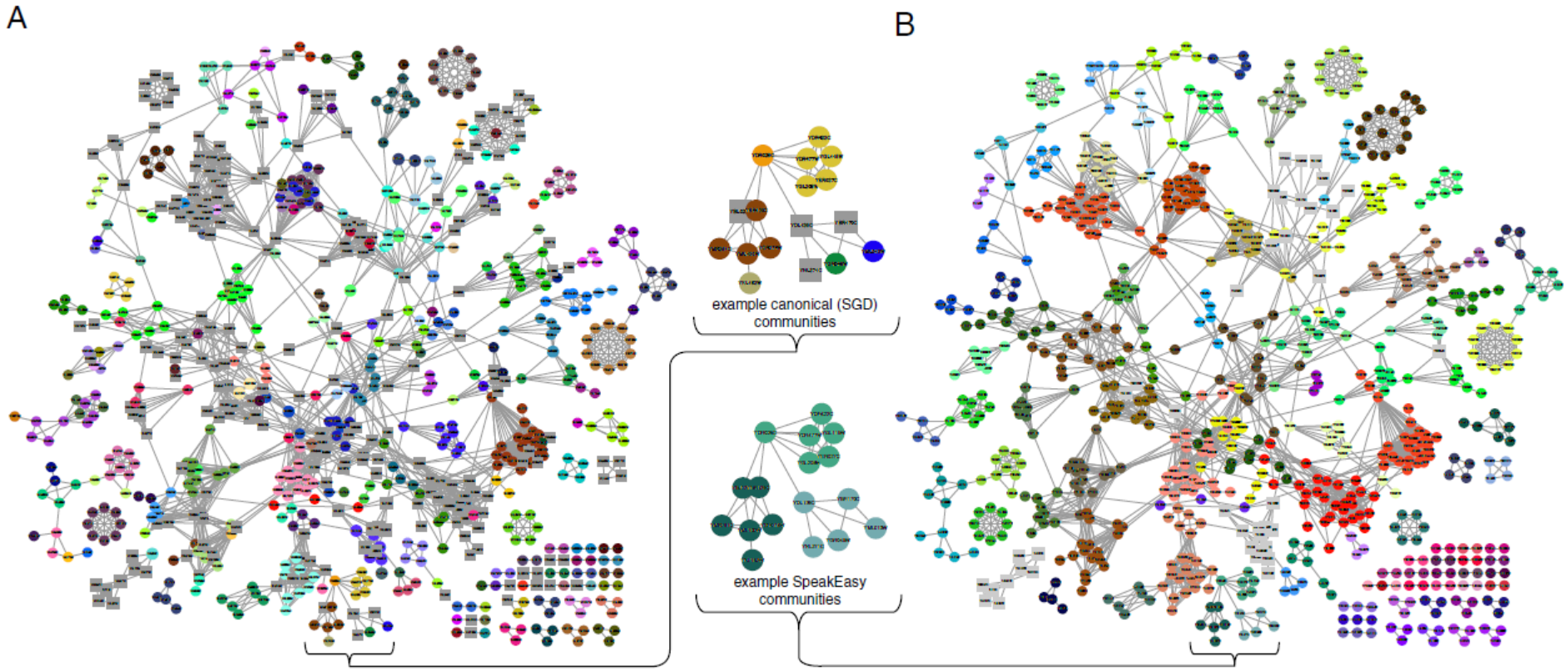
Performance on Real-world Networks

- SpeakEasy shows improved performance on 6/15 networks using the modularity (Q) metric, over other algorithms.
- SpeakEasy performs better than GANXiS on 14/15 of the networks with a mean percent difference of 28% over GANXiS.

network	*n*	*m*	GANXiS (Q)	SpeakEasy (Q)	percentage difference (Q)	GANXiS (Q_{ds})	SpeakEasy (Q_{ds})	percentage difference (Q_{ds})
karate	34	78	0.3924	**0.4198**	6.75	0.2116	**0.2302**	8.42
dolphins	62	159	0.4408	**0.5017**	12.92	0.1664	**0.2378**	35.33
Les. Mis.	77	254	0.5224	**0.5480**	4.78	0.2808	**0.3438**	20.17
pol. books	105	441	0.4831	**0.4973**	2.90	0.1634	**0.2396**	37.82
football	115	613	**0.5878**	0.5811	-1.15	0.3792	**0.4856**	24.61
Santa Fe	118	200	**0.7166**	0.4792	-39.69	0.2099	**0.2963**	34.13
jazz	198	2742	0.2816	**0.4443**	44.83	0.1917	**0.2134**	10.71
railway	297	1213	**0.6989**	0.6098	-13.61	0.2632	**0.3756**	35.20
c. elegans	453	2525	0.1706	**0.3883**	77.90	0.05151	**0.1079**	70.75
email	1133	5254	**0.5035**	0.4916	-2.39	0.05366	**0.1025**	62.55
pol. blogs	1224	19022	**0.4177**	0.3533	-16.71	0.0230	**0.0426**	59.78
net science	1461	2742	**0.9039**	0.7657	-16.55	**0.5797**	0.3600	-46.76
PGP	10680	24316	**0.8039**	0.7315	-9.43	0.1595	**0.1906**	17.77
DBLP	260998	950059	**0.6622**	0.6066	-8.76	0.2018	**0.2628**	26.29
Amazon	319948	880215	**0.7659**	0.7094	-7.66	0.2007	**0.2556**	24.04

Comparison of the quality of community structures detected with GANXiS and SpeakEasy on 15 real-world networks using modularity (*Q*) and modularity density (Q_{ds}).

Application to Protein-protein Interaction Datasets



A. The high throughput interaction dataset from Gavin et al. has nodes colored according to protein complexes found in the Saccharomyces Genome Database (SGD).

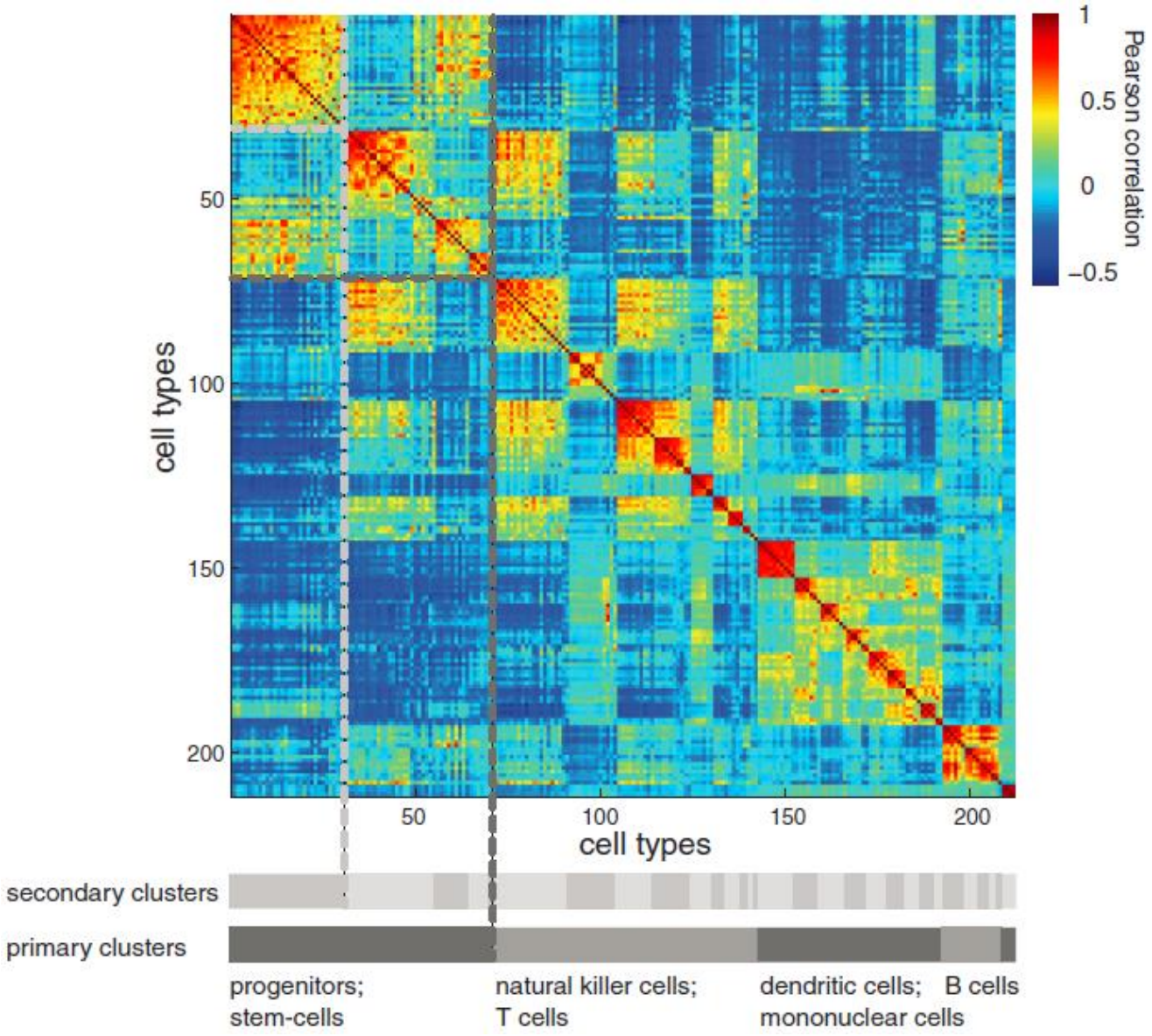
B. The communities identified with SpeakEasy on the high throughput interaction dataset from Gavin et al.

C. Gaiteri, M. Chen, B.K. Szymanski, et al., *arXiv*:1501.04709 (2015)

SCNARC

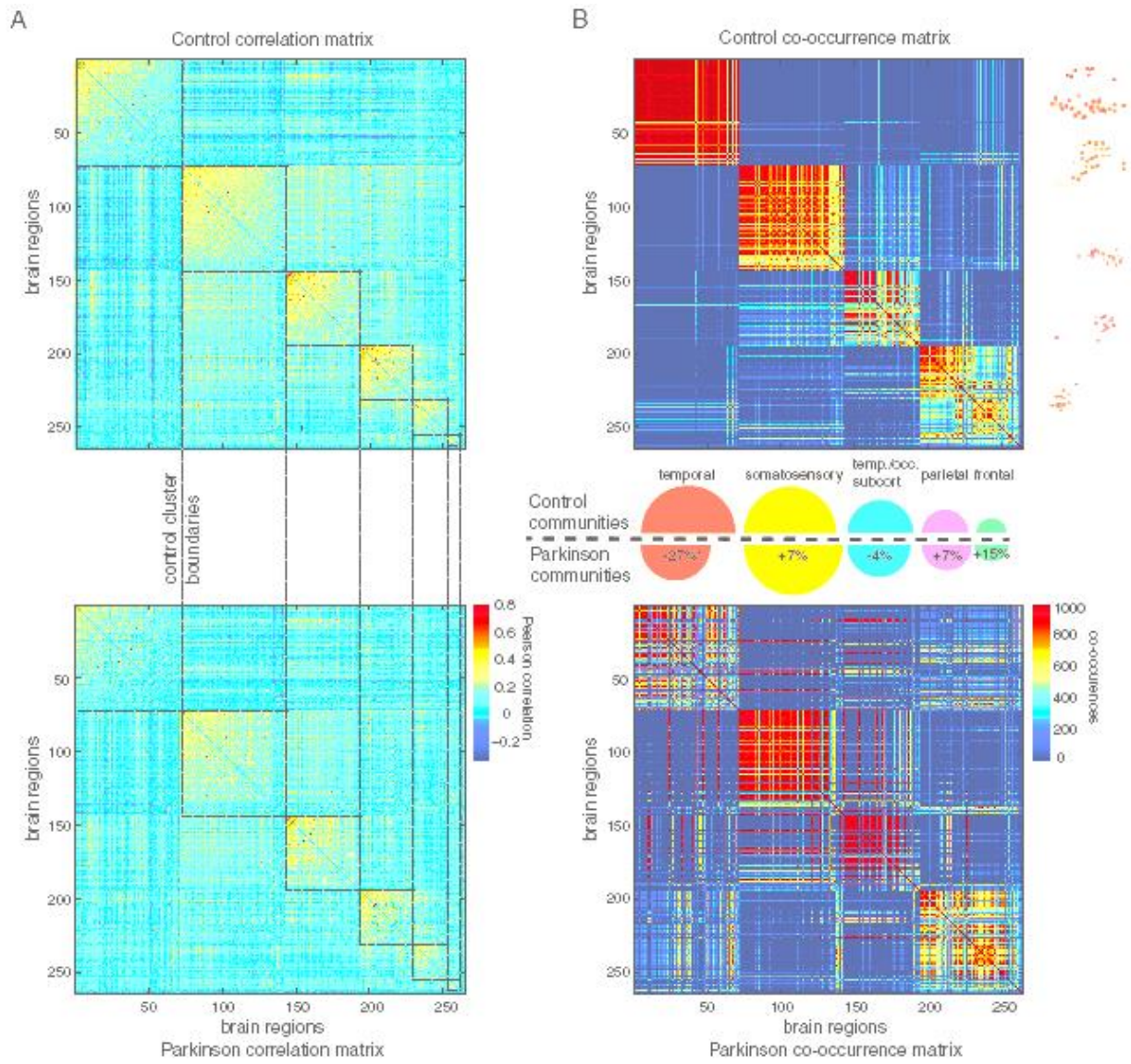
RENSSELAER POLYTECHNIC INSTITUTE 1824

Application to Cell-type Clustering



Primary and secondary biological classifications of immune cell types are reflected in primary and secondary clusters.

Application to Resting-state fMRI Data



A. Raw correlation matrices between resting state brain activity from control and Parkinson disease cohorts.
B. Co-occurrence matrices for controls and Parkinson disease cohorts.

Thank You

Questions?