Healthcare Analytics

Aryya Gangopadhyay

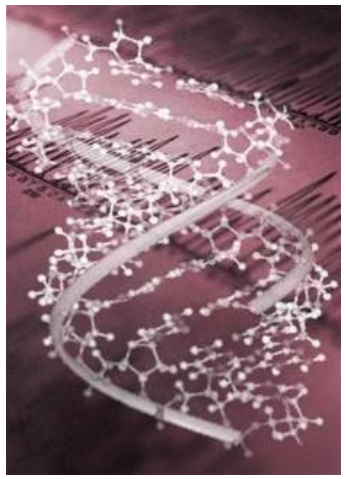
UMBC

Two of many projects

- Integrated network approach to personalized medicine
  - Multidimensional and multimodal
  - Dynamic
  - Analyze interactions
- HealthMask
  - Need for sharing data
  - Protecting privacy
  - Protecting data utility

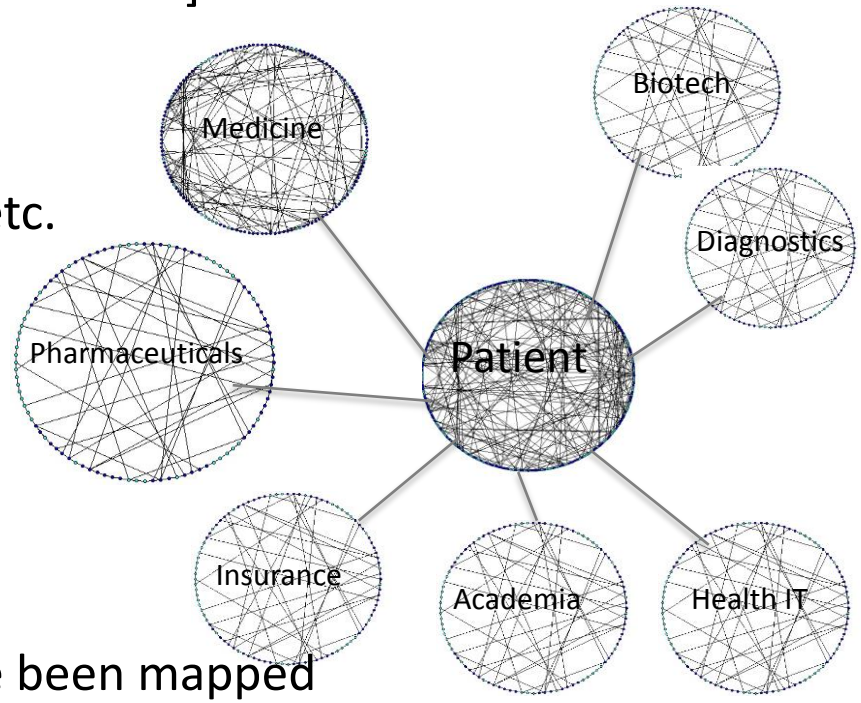
Outline

- Systems approach to healthcare (P4 medicine)
  - Components
  - Interactions
  - dynamics
- CIDeR biological network
- Topological properties
- Link analysis
- Perturbation analysis
- Implications for personalized medicine
- Challenges, and future work

Motivation

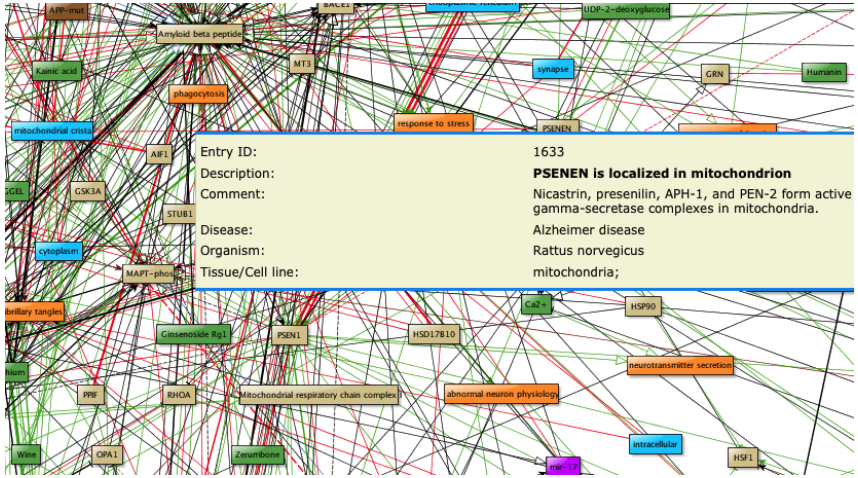
- "Within 10 years every healthcare consumer will be surrounded by a virtual cloud of billions of data points" [Hood *et al.* 2013]



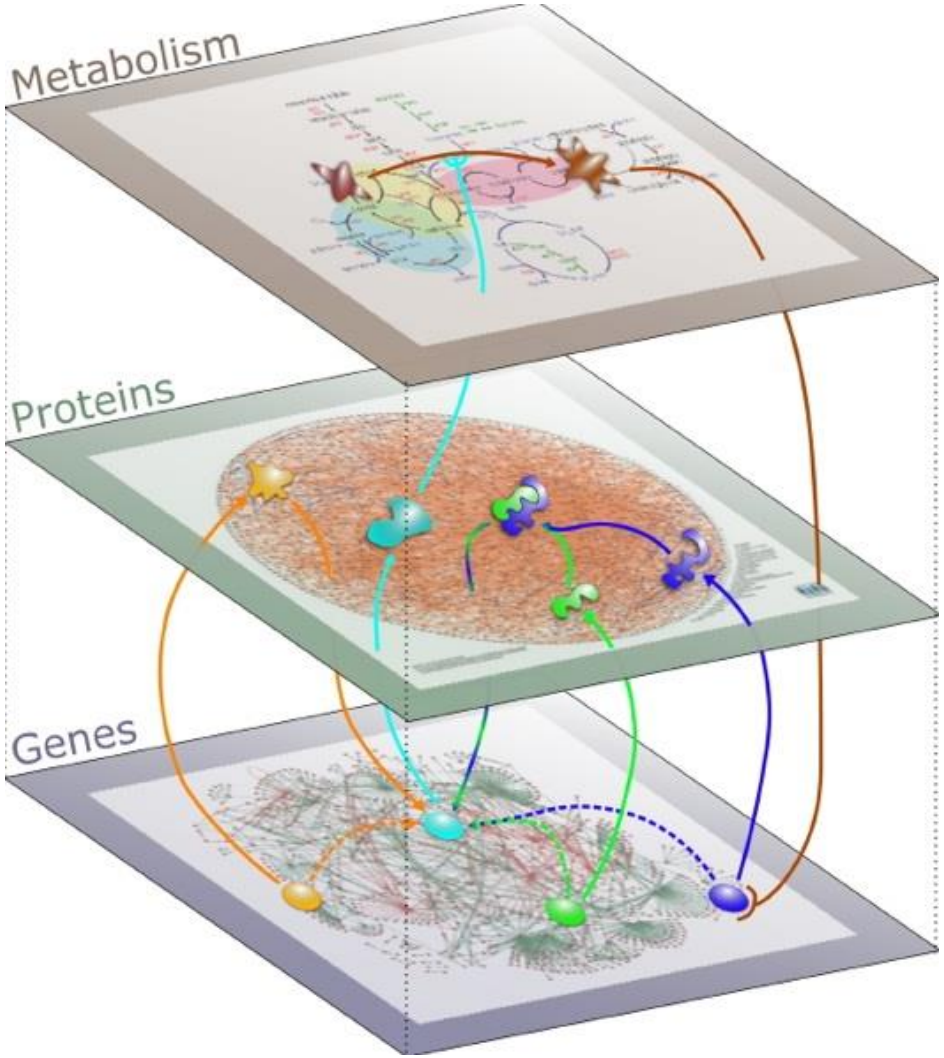
- Structural model of interactions
  - Network of nodes: genes, proteins, etc.
  - Interaction: edges
- Multiple, interdependent networks
  - Gene regulatory networks
  - Protein-protein interaction networks
  - Metabolic networks
- Some of the fundamental networks have been mapped
- Why networks?
  - The behavior of complex systems arises from the coordinated actions of its interactive components

Motivation

- Biological processes as interconnected systems
- Analyze interactions
- Resilience against random perturbations
- Vulnerable to targeted attacks

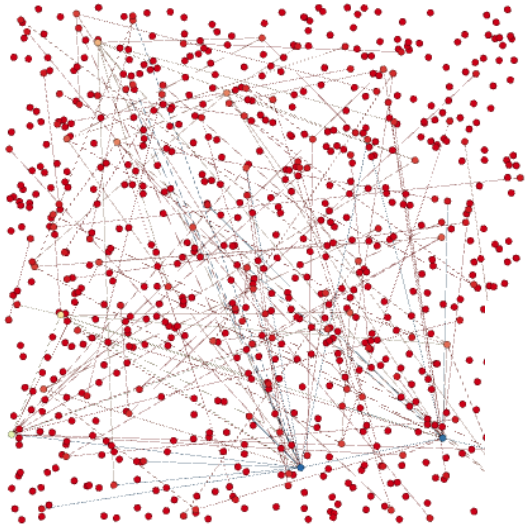


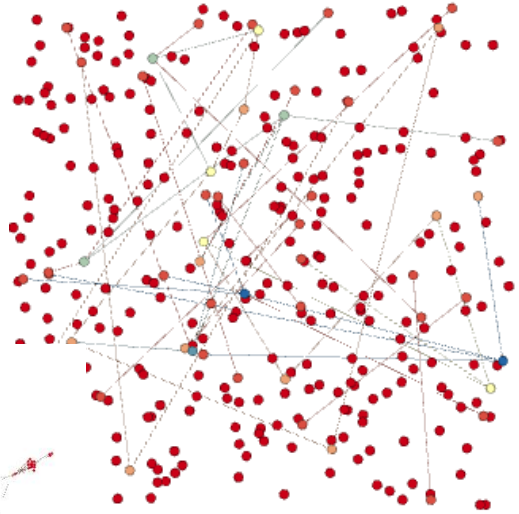
Large, multi-dimensional, multimodal, dynamic



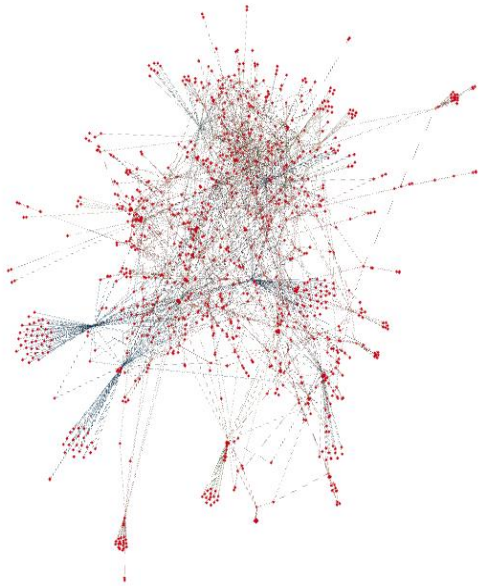
Multimodal network

Process



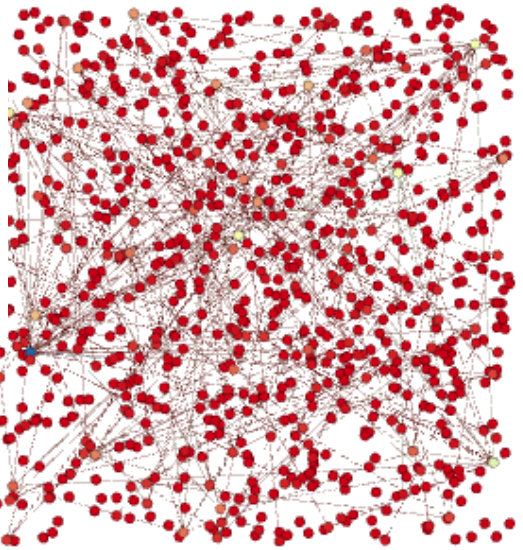


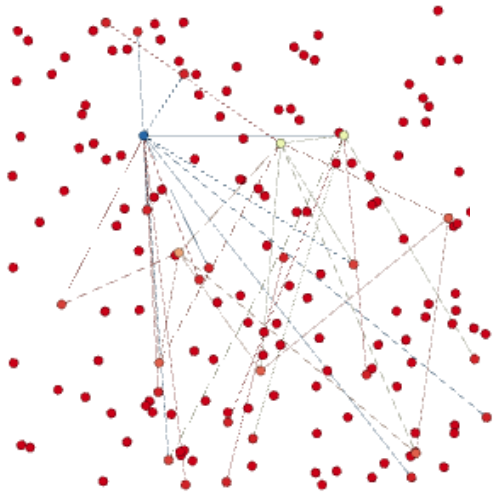
Phenotype



Genes/Proteins

Drugs

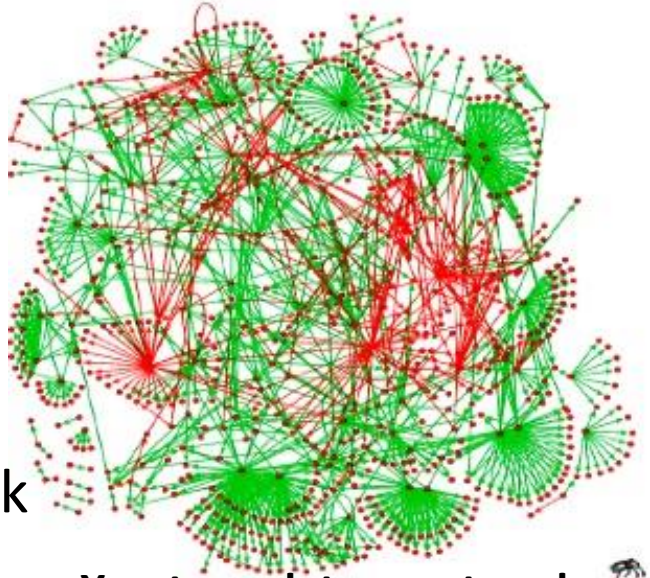




Disease

Biological Networks

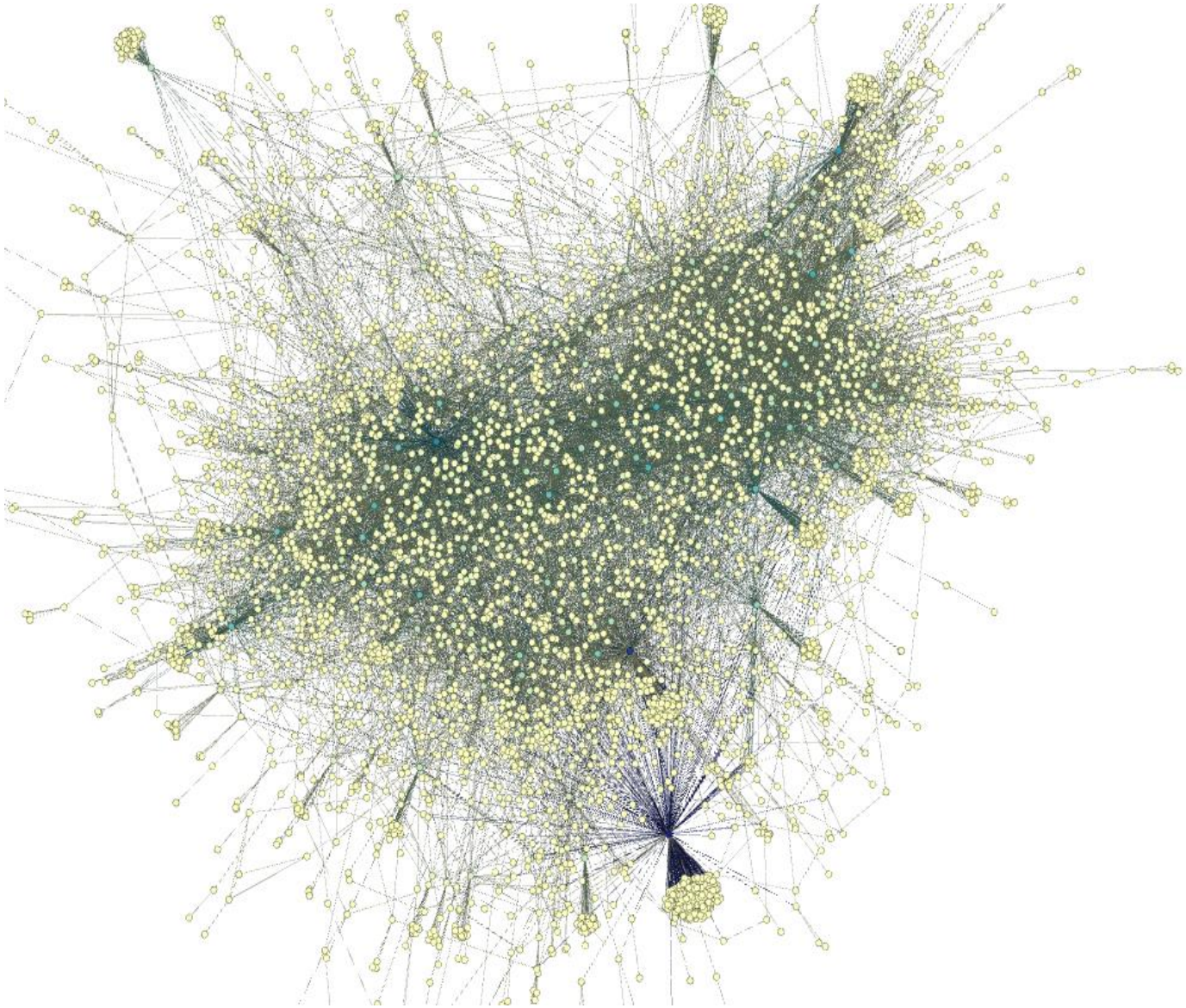
- Different types of networks
- Simple: physical binding
  - Nodes: Proteins, metabolites
  - Edges: binding, reaction
- Directed: transcriptional regulatory network
  - Nodes: transcription factors
  - Edges: regulatory interaction
- Directed but multi-typed edges: activating or inhibitory



Yeast regulatory network

E. Coli metabolic network

Entire CIDeR Network properties (Lechner et al 2012)



- SNPs: 259
- Cell. Comp: 62
- Genes: 2267
- Diseases: 140
- Process: 622
- Gene/protein mutants: 248
- Drugs: 693
- Environment: 75
- Phenotypes: 266
- microRNA: 90
- Tissue/cell: 105

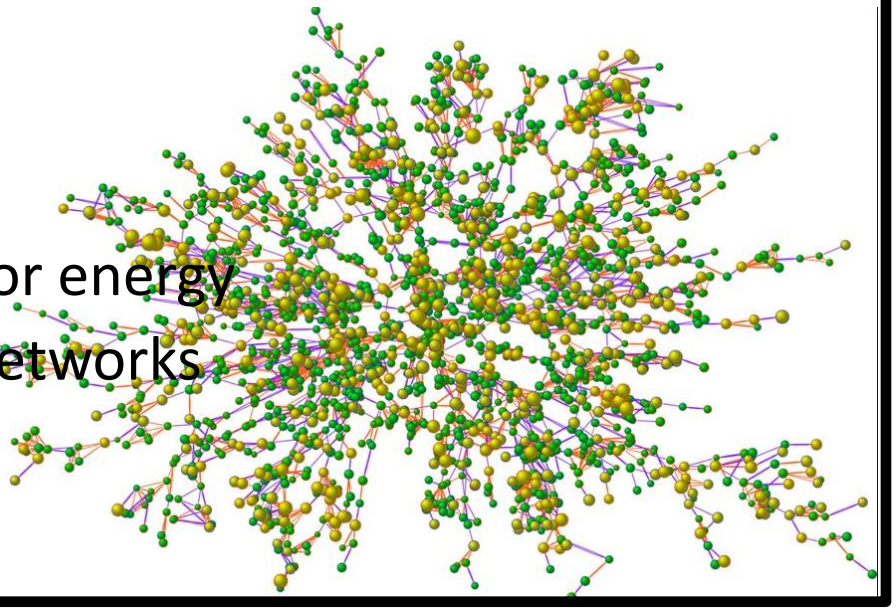
- Nodes: 5168
- Edges: 14410
- Diameter: 16
- # CC: 51
- Avg. PL: 4.46
- Avg. degree: 2.8

Network properties

- Are biological networks random (Erdős-Rényi)?
  - A graph with *n* nodes where each pair of nodes are connected with equal probability
  - P(k): degree distribution assumed to follow the Poisson distribution
- Small world phenomenon (path length)
  - In Erdős-Rényi random networks the **average path length** is

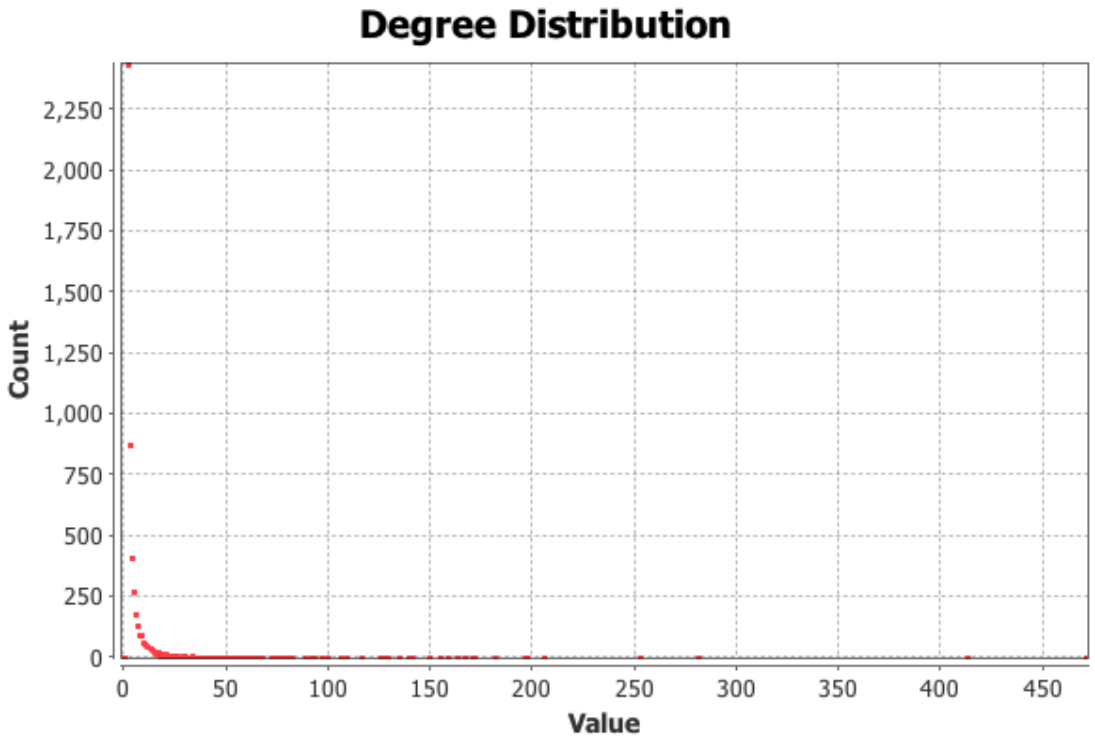
$$\frac{\ln N}{\ln \langle k \rangle}$$

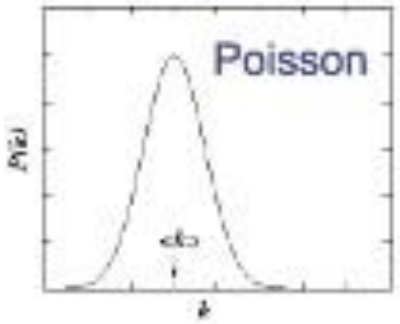
  - **Connected components**
    - Flow of Information, mass, or energy
    - Implications for biological networks



Topological property: scale-free

- Scale-free property
  - Power law degree distribution of biological networks
  - Significance
    - Most nodes with small degree (small neighborhood: 1-3)
    - A small number of nodes with very high degree: hubs (>100)



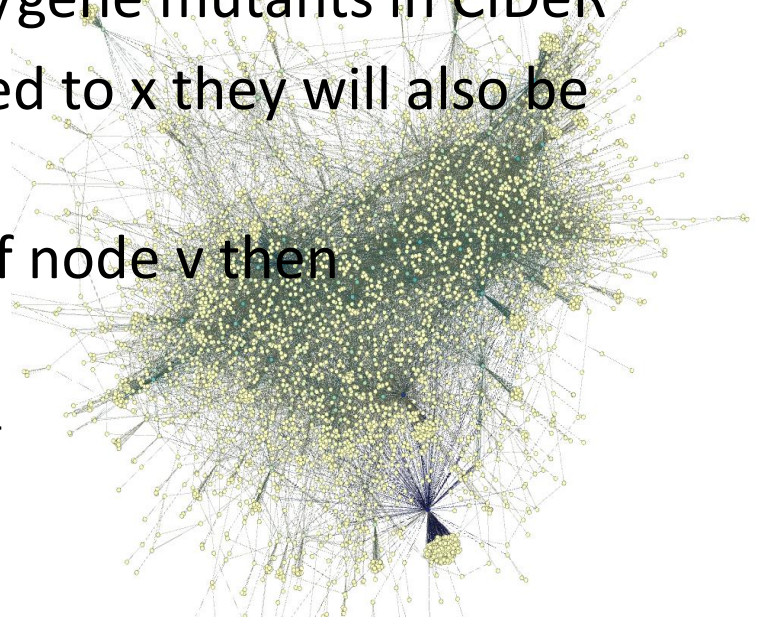


Degree distribution of Erdős–Renyi random network

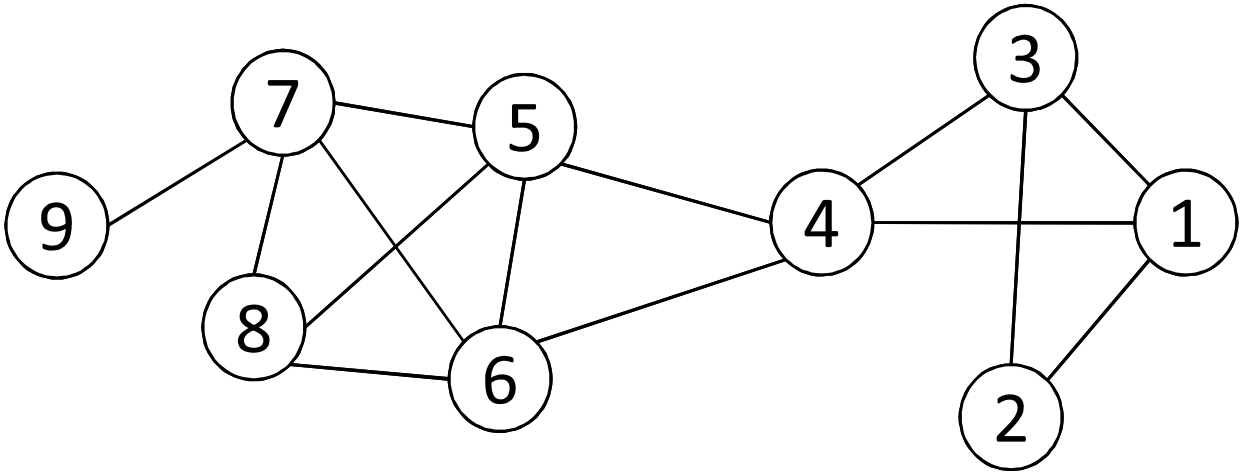
Network properties contd.

- Similarities of biological networks with Erdős-Rényi
  - Giant connected component: 4343 in CIDeR
  - Small-world effect
    - 3 in metabolic networks for 43 different species
    - 4-8 edges in protein-interaction and genetic networks
- Clustering coefficient: 1 for 15 genes/gene mutants in CIDeR
  - if node y and z are both connected to x they will also be connected to each other
  - If C_v is the clustering coefficient of node v then

$$C_v = \frac{2.n_v}{d.(d-1)}$$

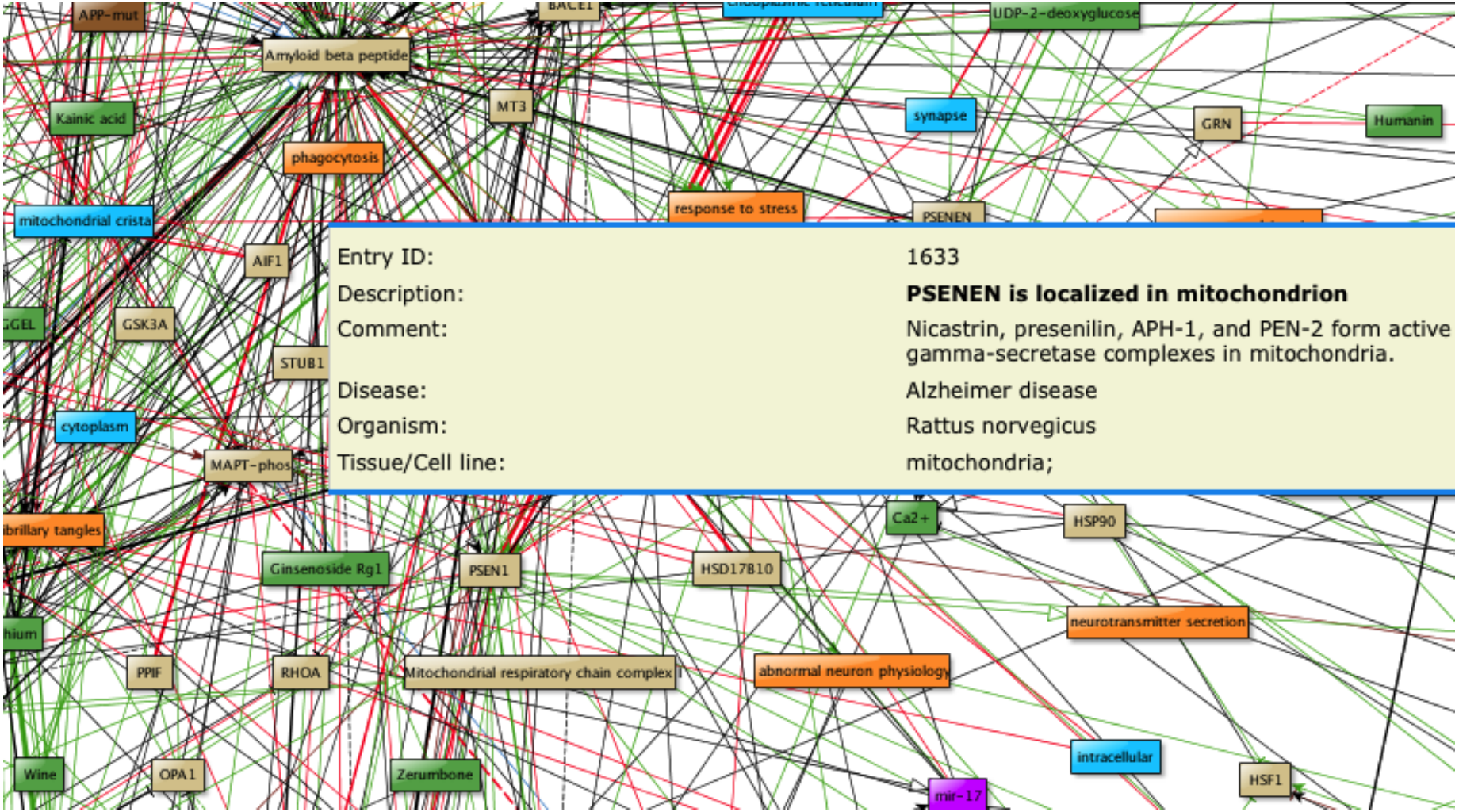


Importance of nodes

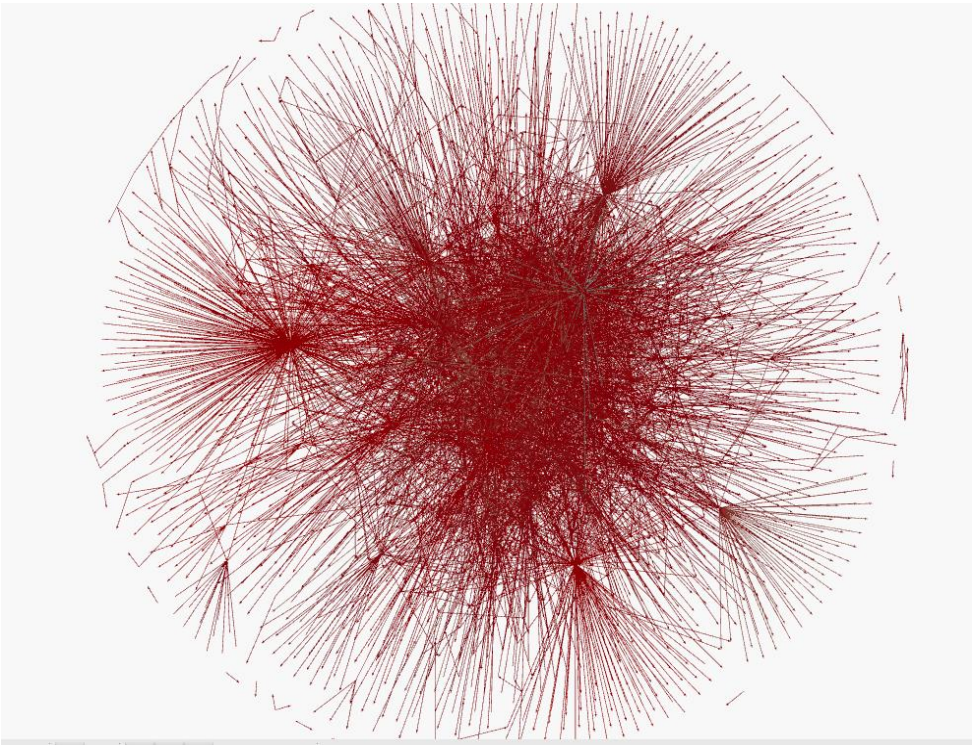


- **Closeness centrality:** a measure of how close a node is to all other nodes (average distance)
- **Betweenness Centrality:** the percentage of shortest paths that pass through the node
- **Eigenvector centrality:** value of the eigenvector of the adjacency matrix

Part of Alzheimer's disease network: CIDeR



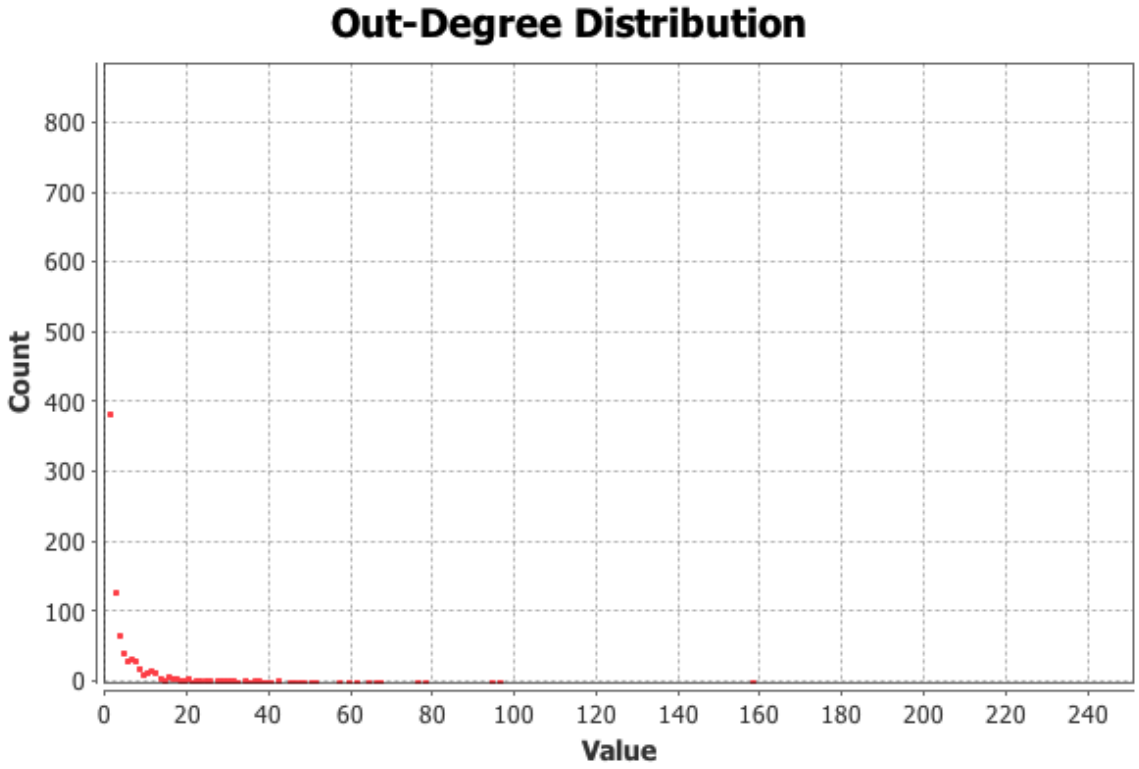
CIDeR Diabetes network



Topological properties:

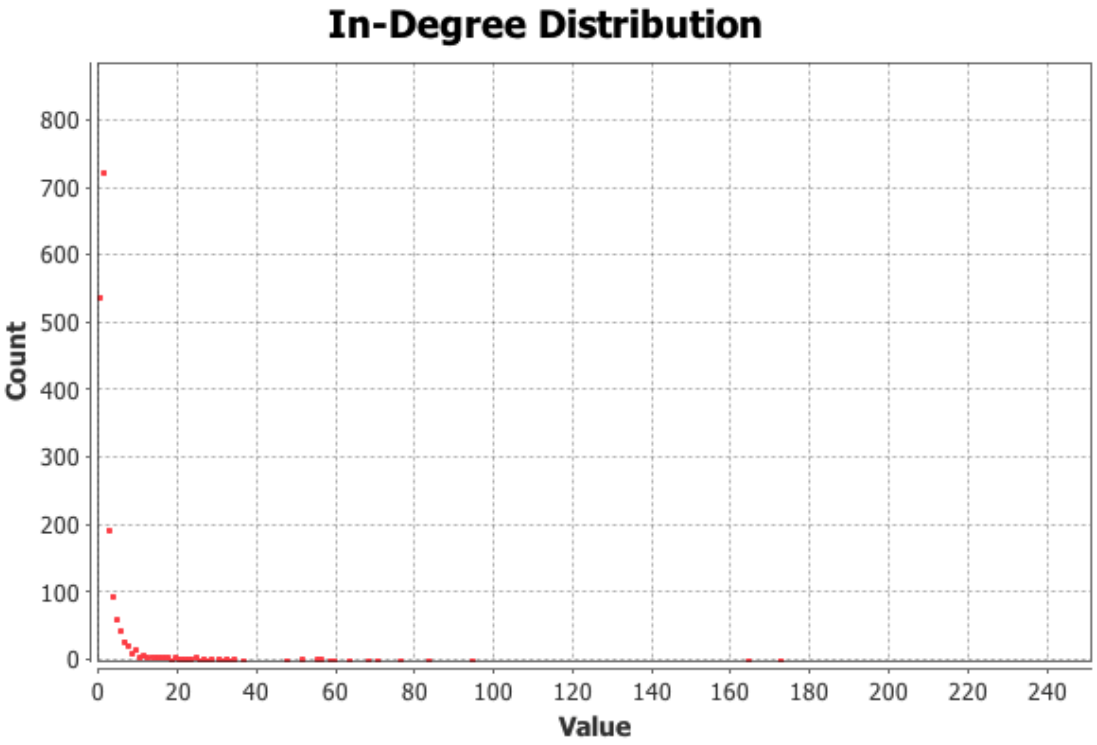
- ✓ 1846 nodes, 5396 edges
- ✓ Connected components: 25
- ✓ GCC: 992 nodes
- ✓ Diameter: 11
- ✓ Average path length: 4
- ✓ average degree: 2.9
- ✓ Graph density: 0.0002
- ✓ Average clustering coefficient: 0.057

CIDeR Diabetes network: highest out-degree nodes

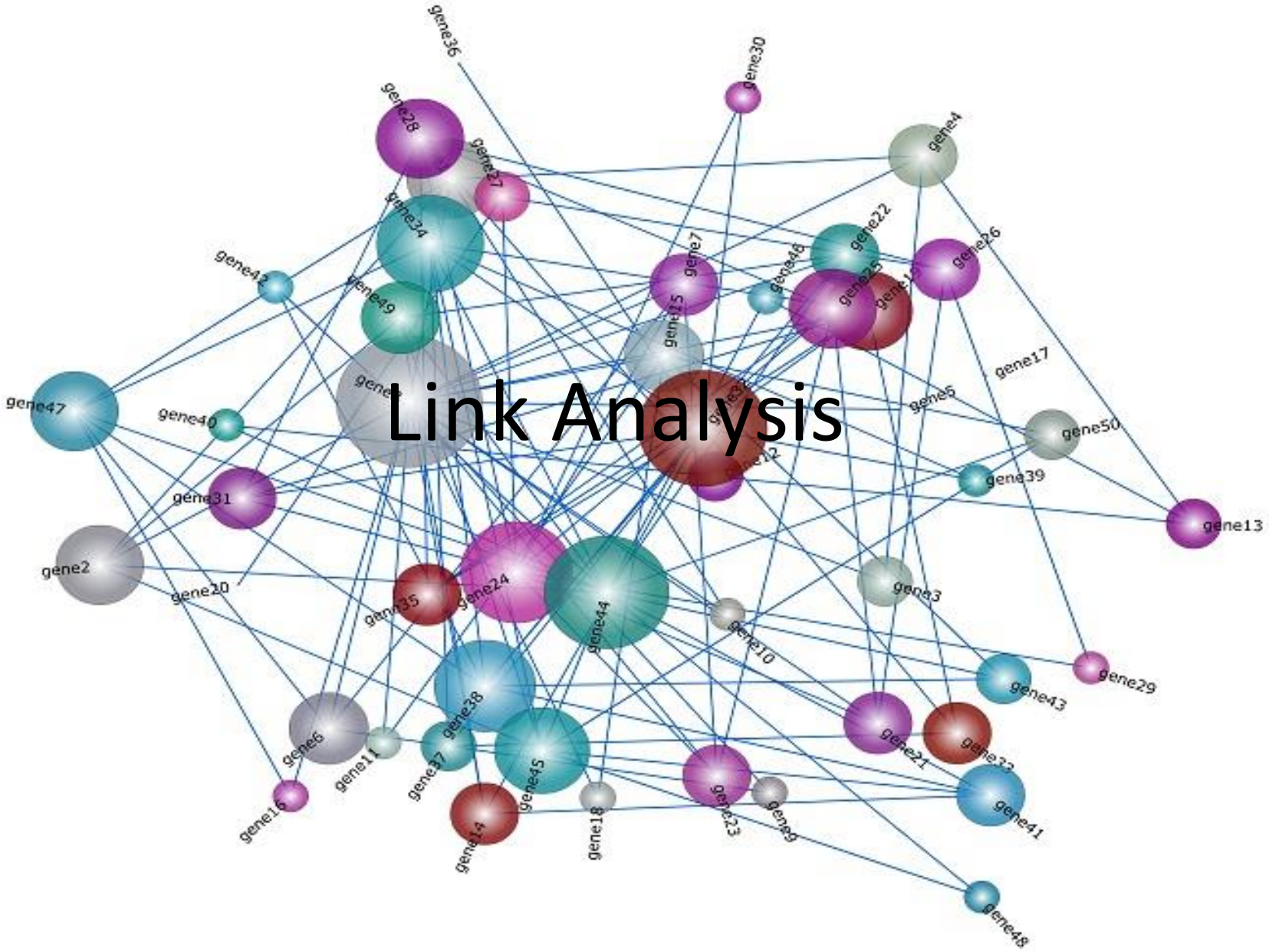


- Environment: High-fat diet (158)
- Cellular component: LPIN1 (96)
- Genes: SRT1 (94), Angiotensin II (67), IL6 (61), GPAM (59), TNF (48)
- Disease: Diabetes mellitus type II (78), Obesity (66),
- Complex PPI: Insulin (76)
- Phenotype: Hyperglycemia (64)
- Drugs: Valsartin (57), Pioglitazone (51), Metformin (50)

CIDeR Diabetes network: Highest in-degree nodes

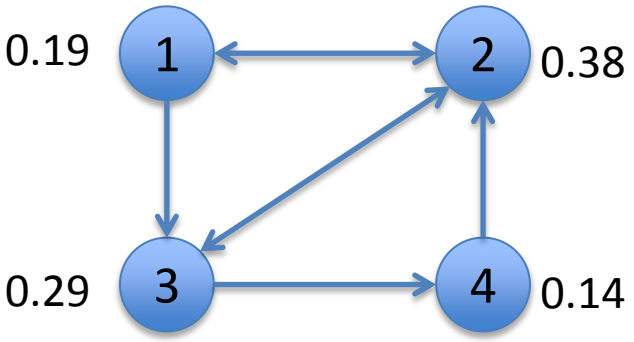


- Disease: D M II (172), Insulin resistance (164), Obesity (96)
- Process: Inflammatory response (83), Insulin receptor signaling pathway (55)
- Genes: TNF (76), IRS1 (70), IL6 (68), CCL2 (59), AKT1 (56), ADIPOQ (55), FASN (34), IL1B (34)
- Phenotype: Insulin sensitivity (63), Hypertension (47)
- Complex PPI: AMPK (58), NF-kappaB complex (51), Insulin (36), Glycated hemoglobin (33)
- Drug/chemical compound: Triacylglycerol (51)


Link Analysis
gene36
gene30
gene28
gene4
gene27
gene24
gene22
gene26
gene42
gene7
gene46
gene49
gene25
gene17
gene47
gene6
gene40
gene50
gene39
gene31
gene13
gene2
gene3
gene20
gene24
gene44
gene10
gene29
gene43
gene38
gene6
gene21
gene32
gene11
gene37
gene45
gene41
gene16
gene14
gene18
gene23
gene9
gene48

PageRank algorithm [BP98]

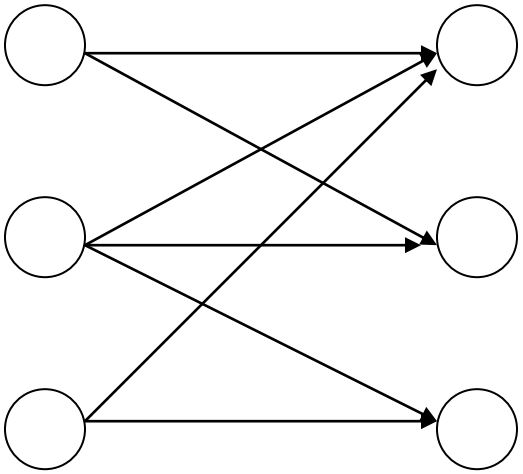
- Important nodes should be pointed to by other important nodes
- Each node has 1 vote that gets equally divided into all nodes it points to
- Random walk on the web graph
  - pick a page at random
  - with probability 1- α jump to a random page
  - with probability α follow a random outgoing link
- Rank according to the stationary distribution



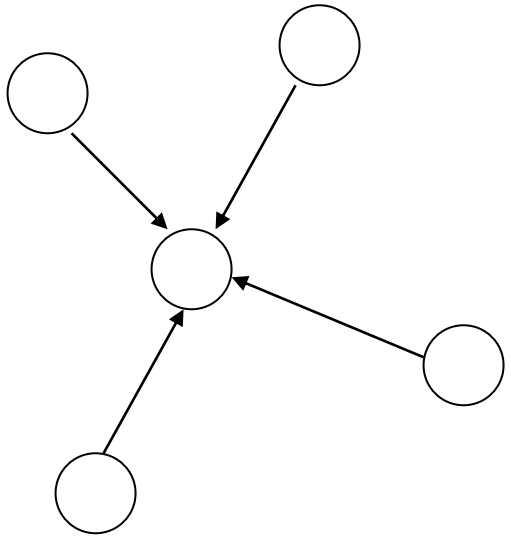
$$\begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 0.5 & 0 & 0.5 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

Effectors and Receptors



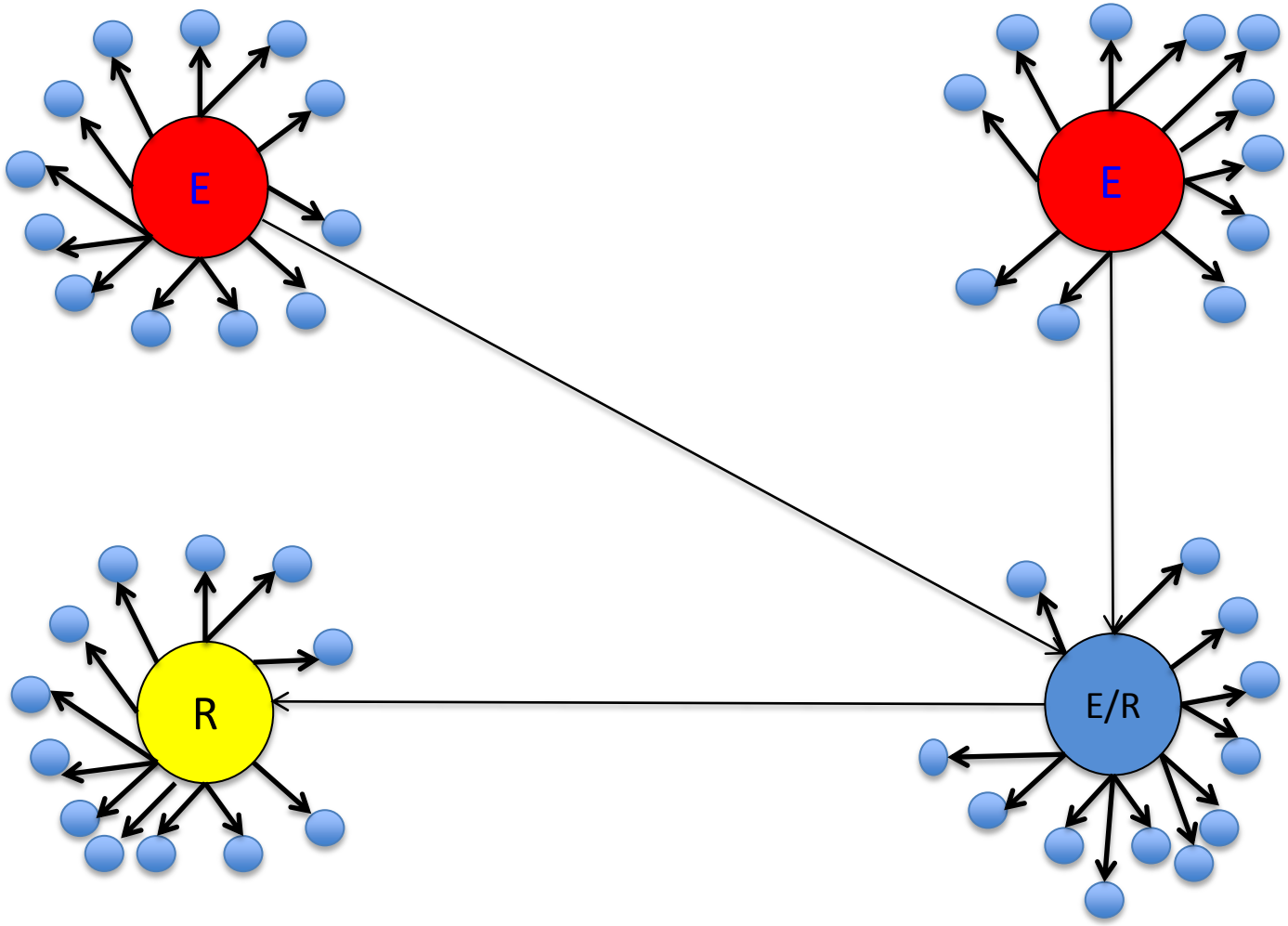
Hubs/effectors **Authorities/receptors**



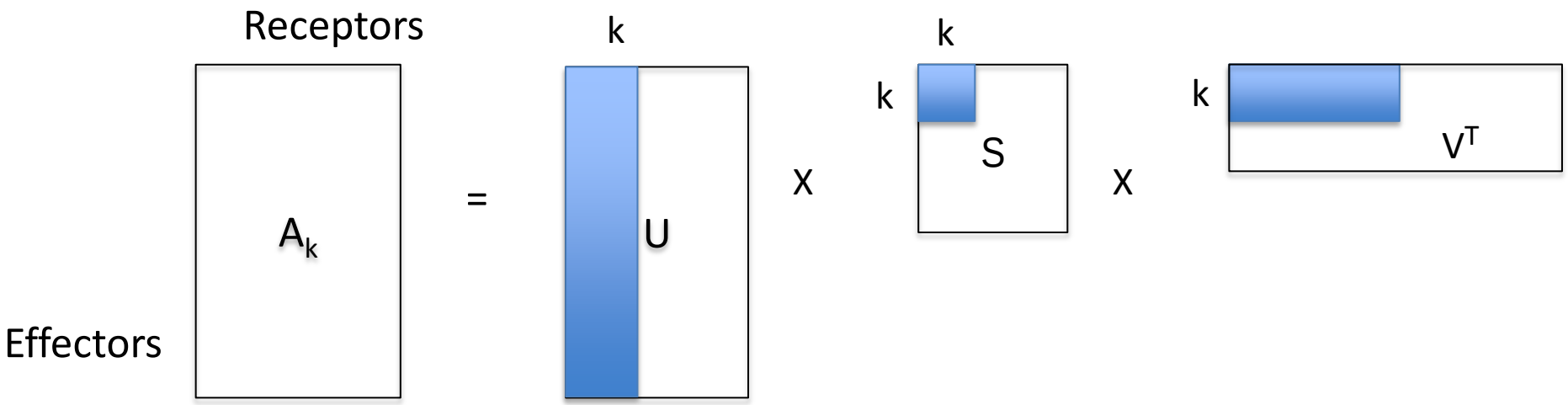
Page with large in-degree

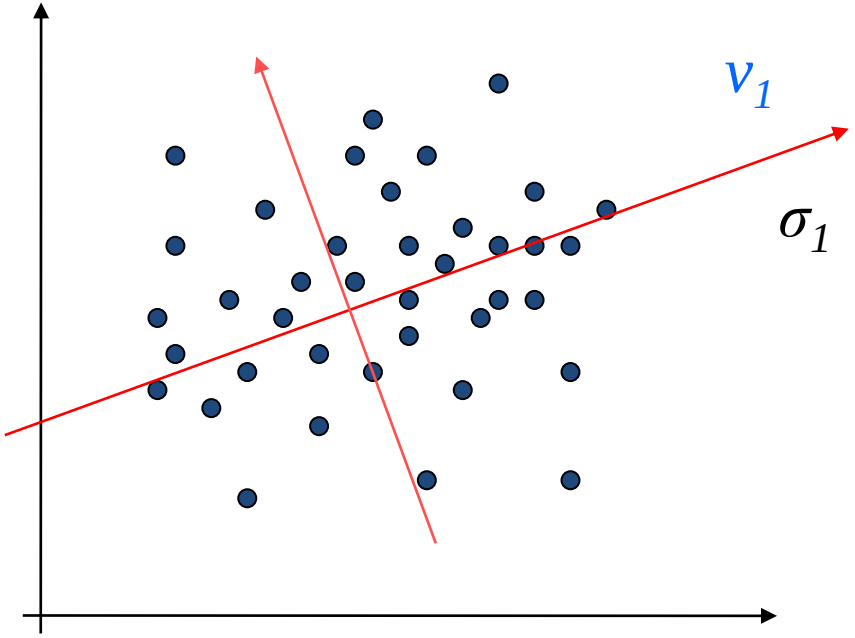
- Preferential attachment
- Important effectors point to important receptors
- Important receptors are pointed to by important effectors
- The seemingly circular description is solvable by SVD

Example

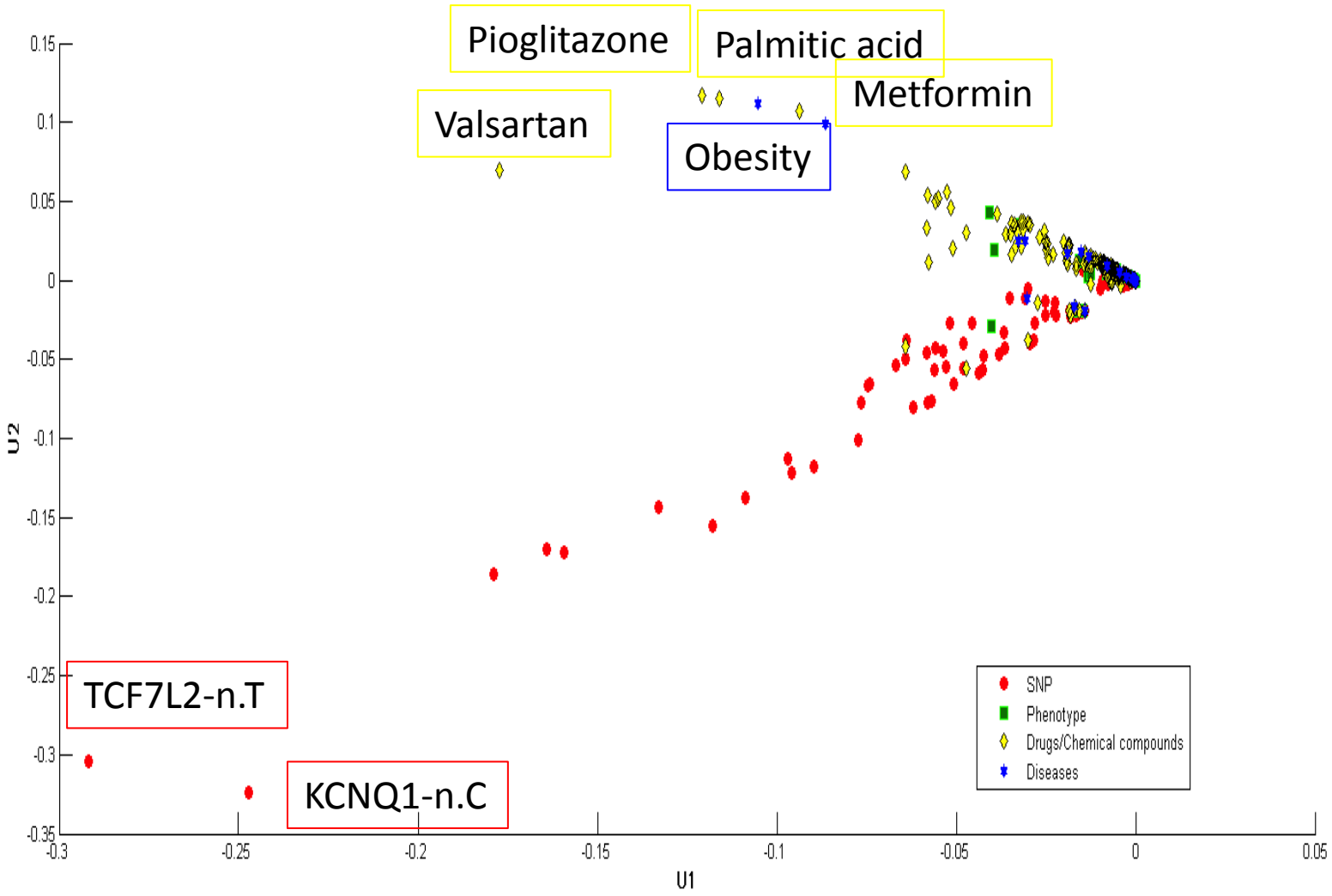


Singular Value Decomposition

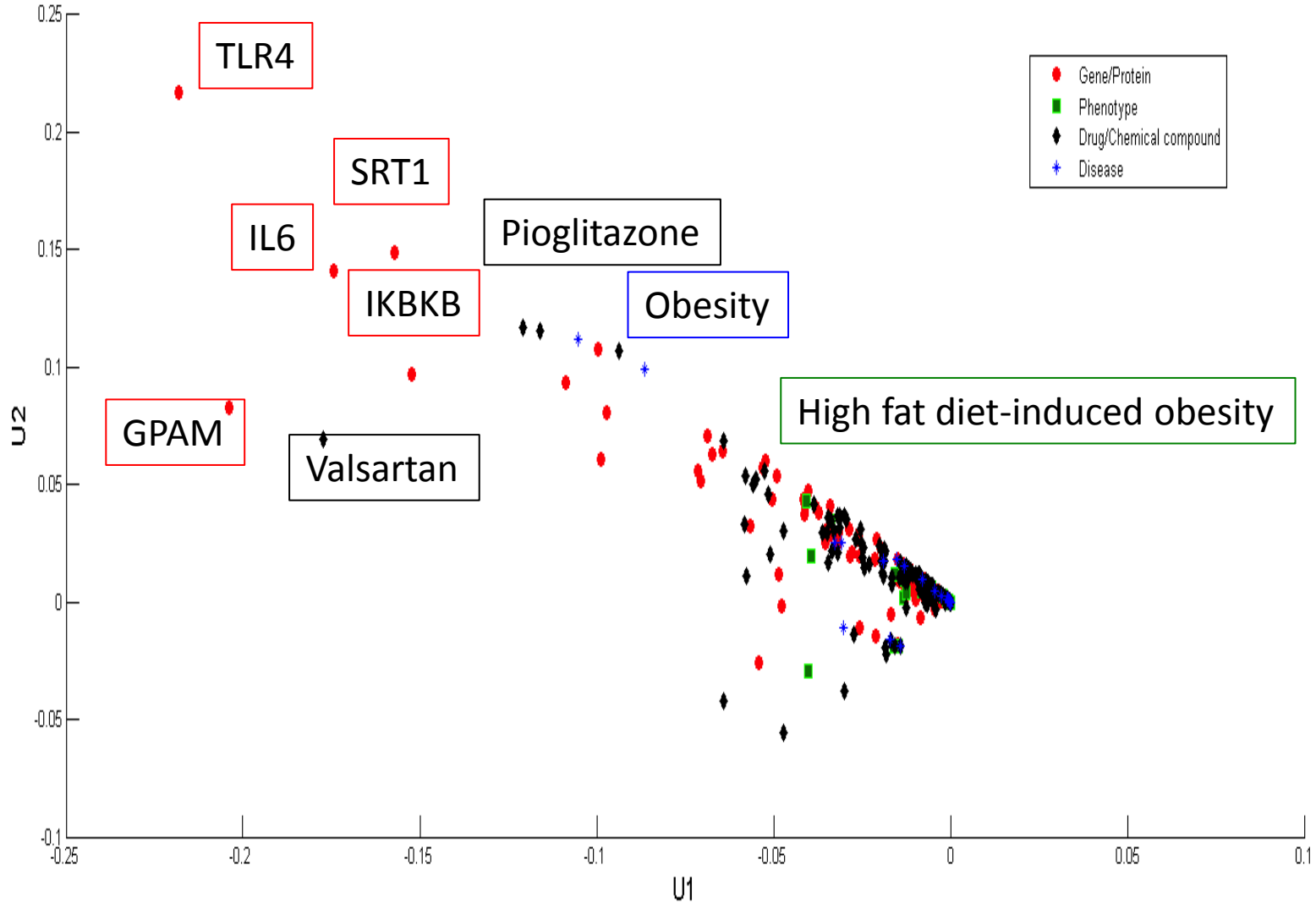




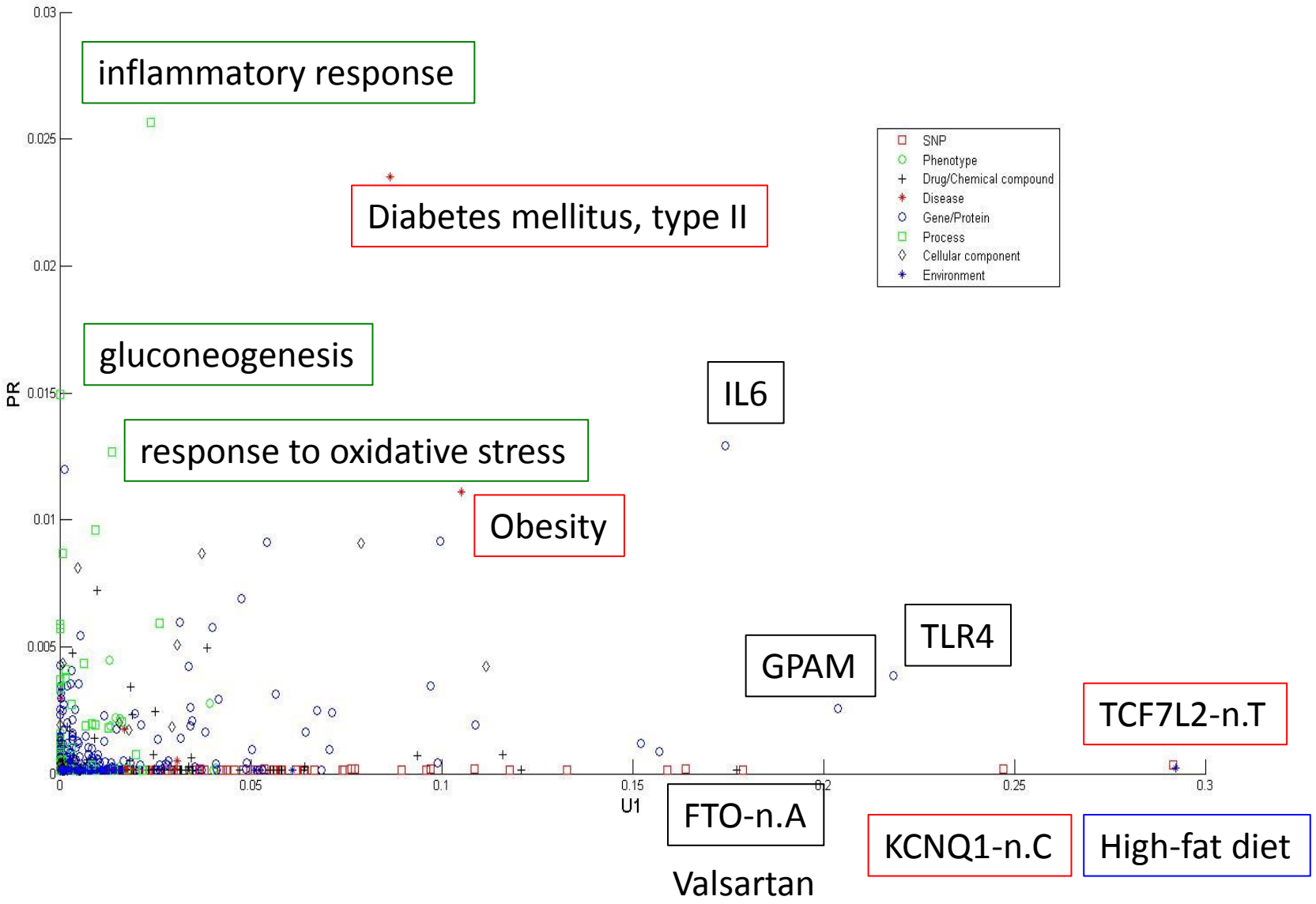
CIDeR Diabetes Network (SNPs)



CIDeR Diabetes Network (genes/proteins)



SVD Results: U1 vs PR

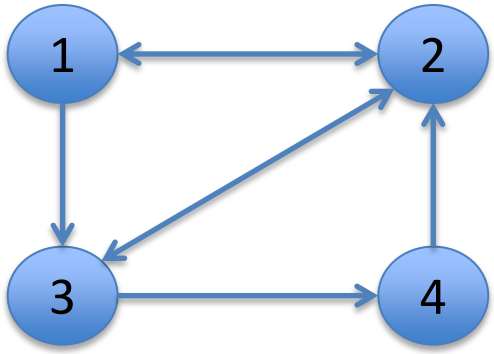


Diabetes-Significant Effectors (SVD)

Index	Element	Type	U1	Rank	In-deg.	Out-deg.
202	TCF7L2-n.T	SNP	1	454	3	22
129	KCNQ1-n.C	SNP	2	800	1	7
1383	TLR4	gene/protein	3	44	20	37
966	GPAM	gene/protein	4	70	14	59
69	FTO-n.A	SNP	5	1366	0	13
633	Valsartan	drug/chemical compound	6	1654	0	57
1040	IL6	gene/protein	7	5	69	62
188	SLC30A8-n.C	SNP	8	802	1	9
96	HHEX-n.C	SNP	9	1393	0	6
1311	SIRT1	gene/protein	10	209	7	94

HITS [Kleinberg98]

- Every node has an authority score and a hub score
- Start with initial hub and authority vectors



$$x^{(0)} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad y^{(0)} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad L = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

- For the adjacency matrix L: $x^{(1)} = L^T . y^{(0)} = \begin{bmatrix} 1 & 3 & 2 & 1 \end{bmatrix}^T$
- Calculate $y^{(1)}$: $y^{(1)} = L . x^{(1)} = \begin{bmatrix} 5 & 3 & 4 & 3 \end{bmatrix}^T$
- Normalize $x^{(1)}$ and $y^{(1)}$ to L1-norm and repeat until convergence

Left and right singular vectors

- HITS converges to the principal eigenvectors of LL^T and L^TL
- Proof sketch

$$L = USV^T \quad L^T = (USV^T)^T = VSU^T$$

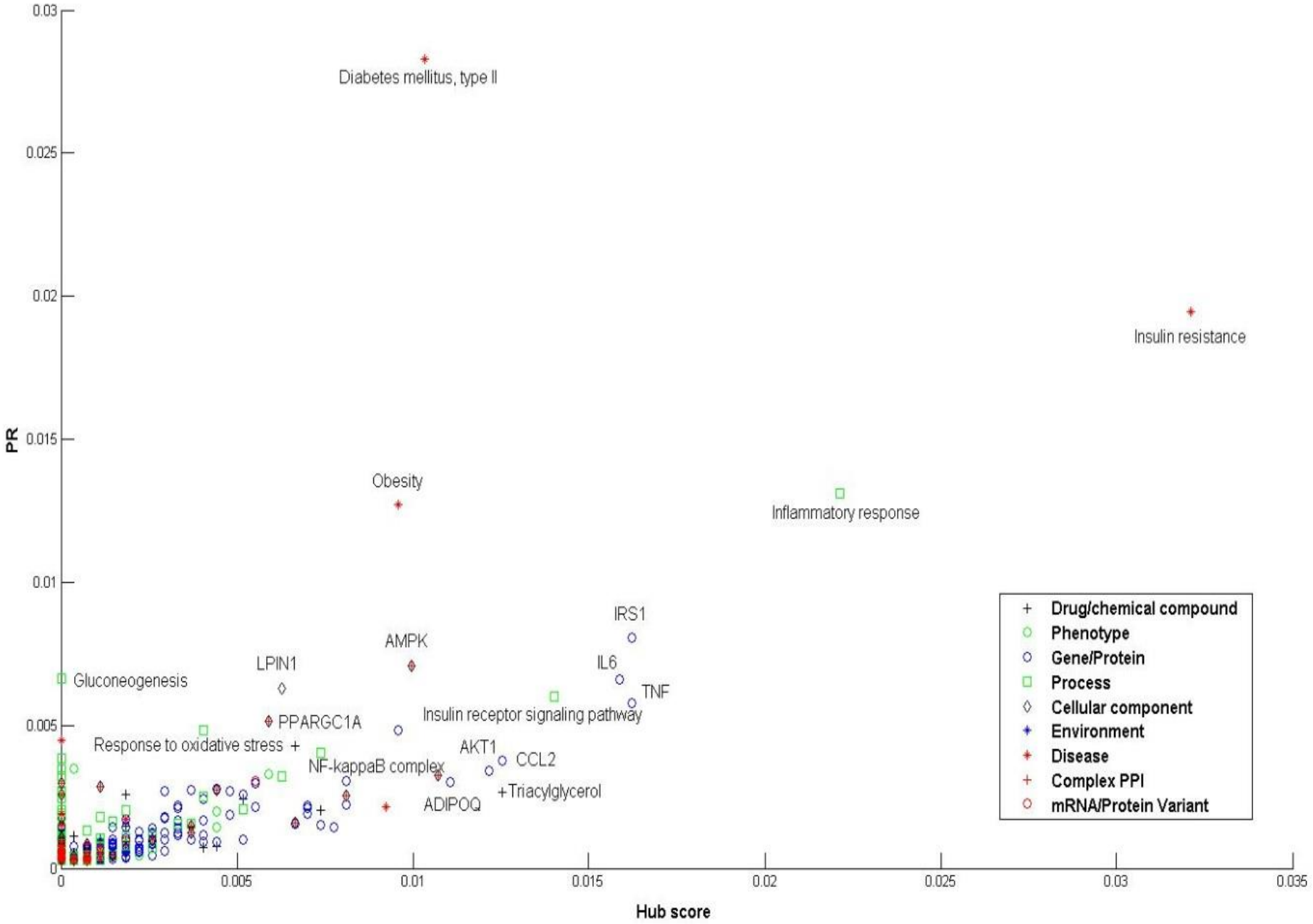
$$LL^T = USV^TVSU^T = US^2U^T$$

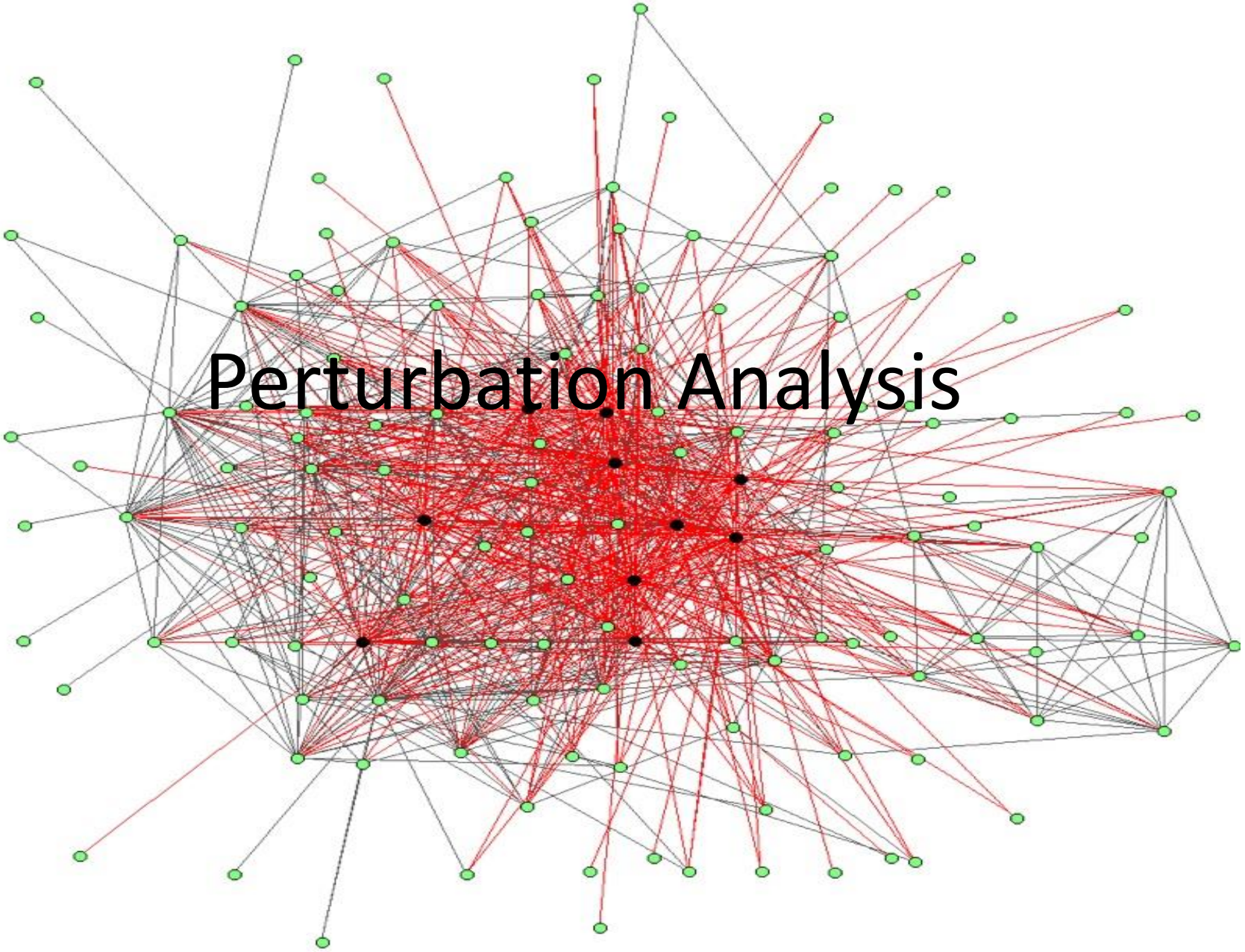
$$(LL^T)^k = US^{2k}U^T$$

$$L^TL = VSU^TUSV^T = VS^2V^T \quad (L^TL)^k = VS^{2k}V^T$$

- Hence, U and V and $S^{1/2}$ are the eigenvector and eigenvalue matrices of LL^T and L^TL respectively

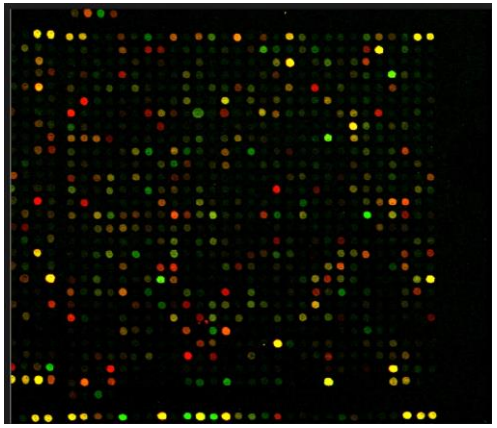
PR vs Hub scores



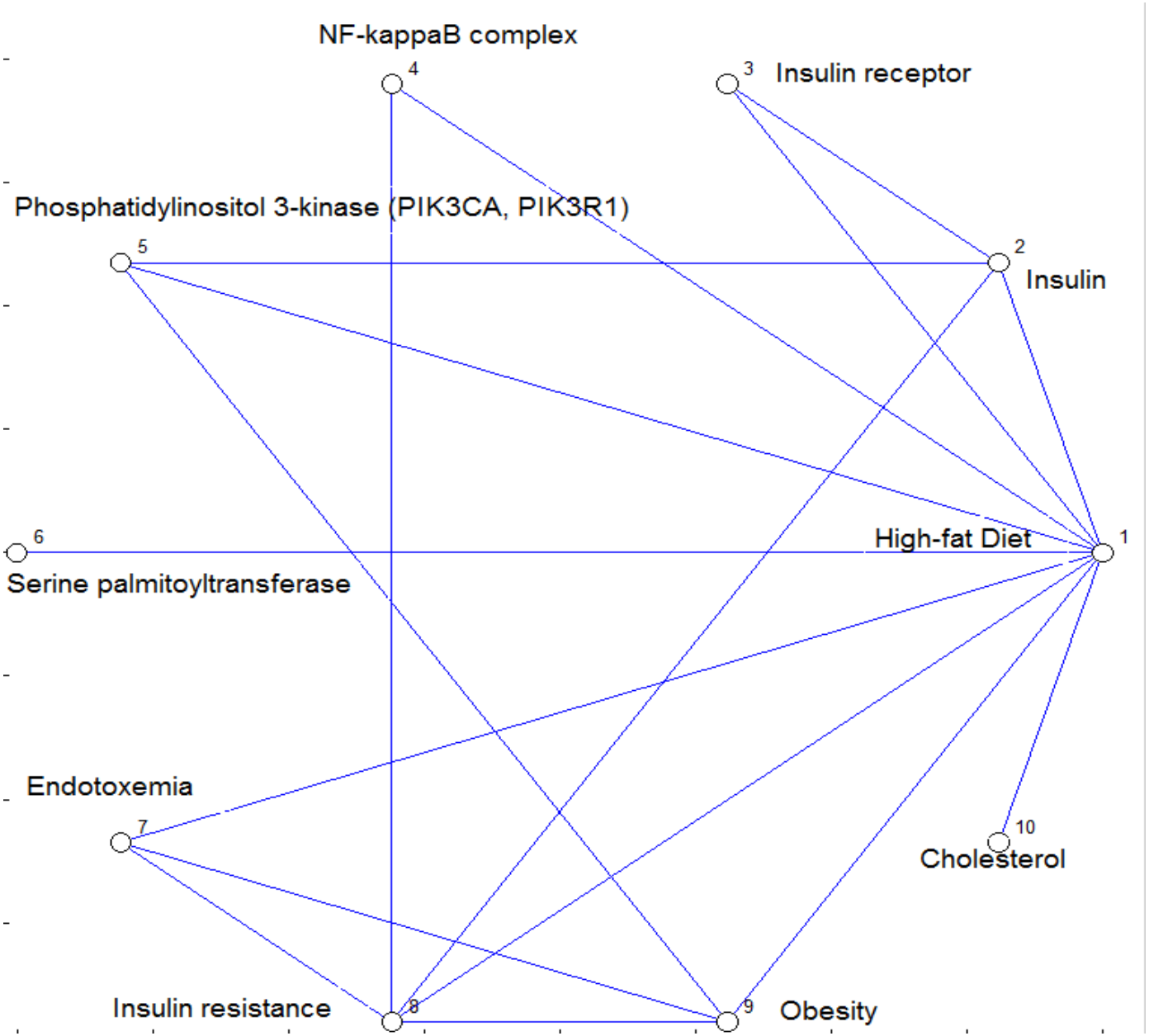

Perturbation Analysis

Diseases as network perturbations

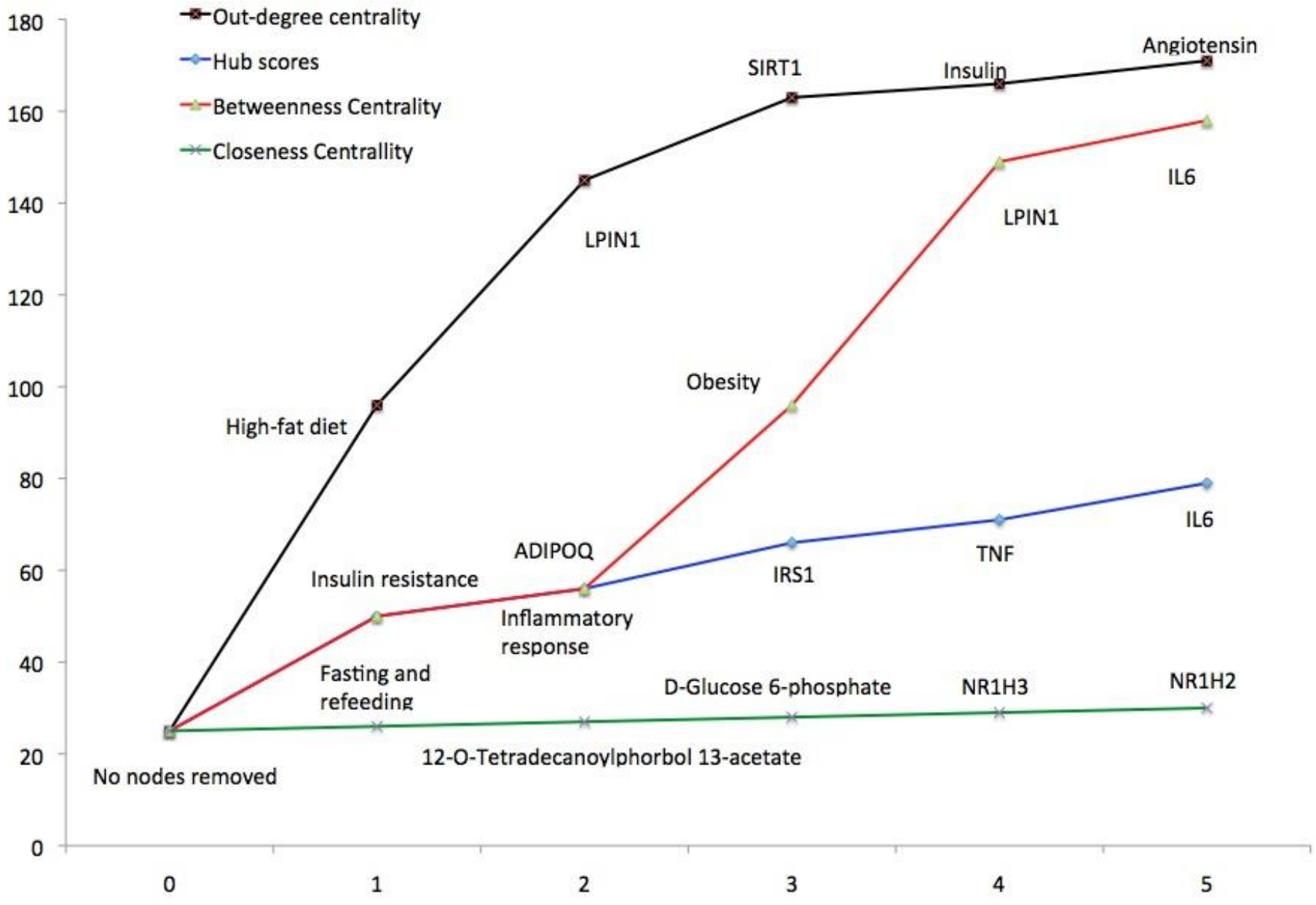
- Diseases as specific types of network perturbation (del Sol et al 2010, Hwuang et al 2009)
- Genetic mutations, environmental factors, epigenetic perturbations
- Some gene regulatory network states may correspond to disease states
- Study the topological properties of disease-perturbed networks
- Study the effect of network perturbations
  - Perturbation of nodes
  - Perturbation of edges



Hub and spoke communities (Kang et al 2011)

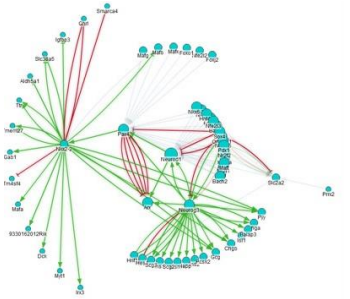


Effect of Perturbation on CCs

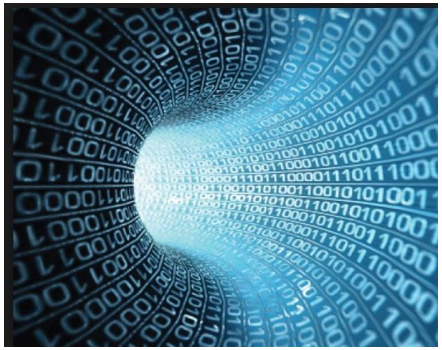


Conclusion

- Limitations
  - Separate analysis for each type of edge
  - Perturbation in gene regulatory networks



- Challenges
  - Are all biological networks scale-free? Przuli et al (2004) show otherwise for PPI of S. Cerevisiae
  - Incomplete and tissue specific data
  - Capturing dynamism
  - Analyzing the entire interactome as opposed to individual networks
  - Big data



- Future work:
  - Detecting community structure
  - Other diseases: Alzheimer's, Parkinson's, Cancer…


QUESTIONS?