Biological Network Analysis
not independent objects, and are not built in the absence of a biological milieu. Biological devices and modules typically function within a cellular environment. When synthetic biologists engineer devices or modules, they do so using the resources and machinery of host cells, but in the process also modify the cells themselves. A major concern in this process is our present inability to fully predict the functions of even simple devices in engineered cells and construct systems that perform complex tasks with precision and reliability. The lack of predictive power stems from several sources of uncertainty, some of which signify the incompleteness of available information about inherent cellular characteristics. The effects of gene expression noise, mutation, cell death, undefined and changing extracellular environments, and interactions with cellular context currently hinder us from engineering single cells with the confidence that we can engineer computers to do specific tasks. However, most applications or tasks we set to our synthetic biological systems are generally completed by a population of cells, not any single cell. In a synthetic system, predictability and reliability may be achieved in two ways: statistically by utilizing large numbers of independent cells or by synchronizing individual cells through intercellular communication to make each cell more predictable and reliable. More importantly, intercellular communication can coordinate tasks across heterogeneous cell populations to elicit highly sophisticated behavior. Thus, it may be best to focus on multicellular systems to achieve overall reliability in performing complex tasks.

Biological devices

Biologists are familiar with manipulation of genes and proteins to probe their properties and understand biological processes. Synthetic biologists must also manipulate the material elements of the cell, but they do so for the purpose of design, to build synthetic biological systems. Synthetic biologists design complex systems by combining basic design units that represent biological functions. The notion of a device is an abstraction overlaid on physical processes that allows for decomposition of systems into basic functional parts. Biological devices process inputs to produce outputs by regulating information flow, performing metabolic and biosynthetic functions, and interfacing with other devices and their environments. Biological devices represent sets of one or more biochemical reactions including transcription, translation, protein phosphorylation, allosteric regulation, ligand/receptor binding, and enzymatic reactions. Some devices may include many diverse reactants and products (e.g. a transcriptional device includes a regulated gene, transcription factors, promoter site, and RNA polymerase), or very few (e.g. a protein phosphorylation device includes a kinase and a substrate). The diverse biochemistries underlying the different devices each provide their own advantages and limitations. Particular device types may be more suitable for specific biological activities and timescales. Although the diversity of biochemical reactions makes it difficult to interface devices, it enables the construction of complex systems with rich functionalities.
Network Biology

- Network biology provides a **better understanding of life and evolution**
- **Applications in medicine**: disease diagnosis, drug development

![Network Diagram](image)

**Cardiovascular cluster**
Disease trajectory covering 6.2 million patients


Jensen et al., Nat Commun 2014
Barabasi et al., Nat Rev Genet 2014
An Example of Precision Medicine

- Precision medicine takes biology into personal grounds

- Link a person with disease and traits
- Provide insight into ancestry of the customer

Sequence a customer’s genome

HEALTH OVERVIEW
- GENETIC RISK FACTORS
- DRUG RESPONSE
- INHERITED CONDITIONS
- TRAITS
- HEALTH TOOLS

ANCESTRY OVERVIEW
- ANCESTRY COMPOSITION
- MATERNAL LINE
- PATERNAL LINE
- NEANDERTHAL ANCESTRY
- ANCESTRY TOOLS

An Example of Precision Medicine

- Precision medicine takes biology into personal grounds

  - Link a person with disease and traits
  - Provide insight into ancestry of the customer

  Networks

  HEALTH OVERVIEW
  - GENETIC RISK FACTORS
  - DRUG RESPONSE
  - INHERITED CONDITIONS
  - TRAITS
  - HEALTH TOOLS

  ANCESTRY OVERVIEW
  - ANCESTRY COMPOSITION
  - MATERNAL LINE
  - PATERNAL LINE
  - NEANDERTHAL ANCESTRY
  - ANCESTRY TOOLS
Plan For Today

1) Very basic biology
2) Protein-protein interaction networks
3) Finding disease modules in networks
   ▪ It is a community detection task!
4) Predicting biological attributes, such as protein functions
   ▪ Guilt-by-association principle
   ▪ Gene recommender systems
Plan For Today

1) Very basic biology
2) Protein-protein interaction networks
3) Finding disease modules in networks
   - It is a community detection task!
4) Predicting biological attributes, such as protein functions
   - Guilt-by-association principle
   - Gene recommender systems
Gene is a basic unit of heredity

Genes are **segments of DNA** that determine properties of an organism as a whole and functions of cells within it

Genes encode a functional unit called **protein**

**Central dogma** describes a two-step process, **transcription** and **translation**, by which the information in DNA flows into proteins
Central Dogma of Biology (2)

- **Transcription**: Producing RNA sequence from DNA template in the **nucleus**
- **Translation**: The synthesis of a protein from RNA template in the **cytoplasm**
- Transformation of a gene into a protein is called **expression**
The Human Genome

- Human Genome Project: 1990-2003, $3 billion
- Genome consists of 23 pairs of chromosomes and has a total of 3.2G bp
- Average gene length is 8k bp, there are ~25k genes
- Only 2-3% of the human DNA are genes, the rest of the DNA does not encode genes but has important regulatory roles

Lander et al, Nature 2001

bp = base pair
Plan For Today

1) Very basic biology
2) Protein-protein interaction networks
3) Finding disease modules in networks
   - It is a community detection task!
4) Predicting biological attributes, such as protein functions
   - Guilt-by-association principle
   - Gene recommender systems
Protein Interaction Networks

- A very common type of biological networks
- Undirected, binary/weighted network
- **Nodes**: proteins
- **Edges**: interactions

Yeast protein-protein interaction (PPI) network

Color signifies the phenotypic effect of removing a protein:
- **red**: lethal
- **green**: non-lethal
- **orange**: slow growth
- **yellow**: unknown

Jeong et al., Nature 2001
Protein-Protein Interactions

- How do we know that a pair of proteins interact?
  - A complex containing these two proteins has been crystallized

- High throughput screening methods enable rapid, parallel acquisition of experimental data
  - Yeast two-hybrid system

- Problems with high throughput methods:
  - False positive and false negative edges
  - Networks are incomplete and noisy

Yeast Two-Hybrid Screening (Y2H)

- Classical screening technology for the study of PPI

Checking for interaction between two proteins, called here Bait and Prey

No PPI interaction!

PPI interaction!

Interaction of bait and prey: LexA and B42 are tethered to one another and transactivate the reporter gene
Protein Interaction Network

Is there a relation between network structure and biological function and disease?
Protein Interaction Networks

Data:

- Three yeast protein-protein interaction (PPI) networks
- List of essential yeast proteins, these proteins form a minimal protein set required for a living cell
- Mapping of proteins to phenotypes (i.e., observable traits, such as diseases) associated with deletion of each protein

![Network images](Binary (Y2H-union), Co-complex (Combined-AP/MS), Literature (LC-multiple))
Hub Proteins

- **Hub proteins**: 20% nodes in the network with the highest degree
- **Observations:**
  - **Hub proteins** associate with **essential proteins**, confirmed in many but not all networks
  - **Hub proteins** associate with **larger numbers of phenotypes** than non-hub proteins

![Graphs showing the distribution of essential genes and phenotypes across different datasets](image)

- Y2H-union
- Combined-AP/MS
- LC-multiple
- **Equations**:
  - \( E = 0.02k + 0.5 \)
  - \( R^2 = 0.84 \)
  - \( P < 10^{-9} \)
  - \( E = -0.004k + 0.6 \)
  - \( R^2 = 0.57 \)
  - \( P = 10^{-15} \)
  - \( E = 0.0002k + 0.2 \)
  - \( R^2 = 0.0025 \)
  - \( P = 0.81 \)

**Reported by**: Yu et al., Science 2008
For a protein $p_1$, take the fraction of essential proteins among all proteins whose distance to protein $p_1$ is equal to $d$:

$$Q(p_1, d) = \sum_{p \in S_d(p_1)} \frac{I(p \text{ is essential})}{|S_d(p_1)|}$$

Note: $I(X) = \begin{cases} 1 & \text{if } X \text{ is true} \\ 0 & \text{otherwise} \end{cases}$

![Graph showing the distribution of giant component size among essential proteins](image1)

![Graph showing the fraction of essential proteins](image2)
Disease Proteins

Nodes: proteins
Edges: PPI interactions
Node colors: protein-disease associations
Given **disease proteins**, compute **shortest path distance** $d_s$ of each disease protein to the closest disease protein.

- $P(d_s)$ is **shifted towards smaller** $d_s$ compared to the random expectation $P^{\text{rand}}(d_s)$
  - $\Rightarrow$ Disease proteins **agglomerate** in one network neighborhood
Disease Proteins

- **Disease module principle:** Disease proteins tend to cluster in one network neighborhood.

- **Local interaction principle:** Disease proteins tend to interact with each other.

- Mutations in interacting proteins tend to lead to diseases with similar phenotypes (i.e., symptoms).
Disease Proteins

- **Disease module principle:** Disease proteins tend to cluster in one network neighborhood
- **Local interaction principle:** Disease proteins tend to interact with each other
- Mutations in interacting proteins tend to lead to diseases with similar phenotypes (i.e., symptoms)

Can we use these principles to detect disease modules in biological networks?
Plan For Today

1) Very basic biology

2) Protein-protein interaction networks

3) Finding disease modules in networks
   - It is a community detection task!

4) Predicting biological attributes, such as protein functions
   - Guilt-by-association principle
   - Gene recommender systems
By disease module principle, disease proteins are localized in network neighborhoods

**Disease module** $D$:

- **Set of proteins** involved in disease $D$
- **Abnormalities/mutations** in these proteins cause a disease to develop
Goal: Find disease modules in a PPI network

This is a community detection problem

Many community detection methods:

- Girvan-Newman method
- Clique percolation method
- Louvain method
- Spectral clustering
- Link clustering

See CS224W lectures on Community Detection
Finding Disease Modules

- Three basic stages:
  1. Construct a PPI network
  2. Apply a community detection method
  3. Evaluate the quality of detected communities

- Questions:
  - How to evaluate which detected communities are "good" disease modules?
  - How to assign a detected community to a disease?
Evaluating Detected Communities

- A typical method detects **many communities** in a PPI network
- Some detected communities might have a **biological meaning**, some might represent spurious effects
- **Task:** Evaluate the quality of each detected community

Is there a significant association between proteins in a detected community and a disease?
Evaluating Detected Communities

Is there a significant association between proteins in a detected community and a disease?

- This means:
  - “Are unusually many (or: unusually few) proteins in a community actually disease proteins?”

- More precise:
  - “If I picked \( n \) proteins at random (with \( n \) being the size of a community), how probable is it that among these proteins, there are at least as many disease proteins as there are in the community?”
Let \( C = \{g_1, g_2, ..., g_n\} \) be a detected community.

Let \( D = \{d_1, d_2, ..., d_K\} \) be disease proteins involved in disease \( D \).

Let \( k = |C \cap D| \) be the size of the overlap between \( C \) and \( D \).

If I picked \( n \) proteins at random, how probable is it that among these proteins there are at least \( k \) disease proteins?

What is the probability of observing association at least this extreme due to chance?
Construct a 2 x 2 contingency table:

<table>
<thead>
<tr>
<th></th>
<th>Associated with disease $D$</th>
<th>Not associated with disease $D$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within community $C$</td>
<td>$k$</td>
<td>$n - k$</td>
<td>$n$</td>
</tr>
<tr>
<td>Outside community $C$</td>
<td>$K - k$</td>
<td>$N - n - K + k$</td>
<td>$N - n$</td>
</tr>
<tr>
<td>Total</td>
<td>$K$</td>
<td>$N - K$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

$k = |C \cap D|$

$D = \{d_1, d_2, ..., d_K\}$

$C = \{g_1, g_2, ..., g_n\}$

Probability to get this **contingency table** if there is no association between $C$ and $D$:

<table>
<thead>
<tr>
<th></th>
<th>Associated with disease $D$</th>
<th>Not associated with disease $D$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Within community $C$</strong></td>
<td>$k$</td>
<td>$n - k$</td>
<td>$n$</td>
</tr>
<tr>
<td><strong>Outside community $C$</strong></td>
<td>$K - k$</td>
<td>$N - n - K + k$</td>
<td>$N - n$</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>$K$</td>
<td>$N - K$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

$$P(|C \cap D| = k) = \frac{(K)\binom{N-k}{n-k}}{\binom{N}{n}}$$

This is our null model!
Fisher’s Exact Test

- **Exact hypergeometric probability** of observing this particular contingency table, assuming the given marginal totals:

\[
P(|C \cap D| = k) = \frac{\binom{K}{k} \binom{N-k}{n-k}}{\binom{N}{n}}
\]

- **Goal**: Probability of observing association between \(C\) and \(D\) at least this extreme due to chance

- Consider **all possible overlaps** between \(C\) and \(D\) that are equal or larger than \(k\):

\[
P(|C \cap D| \geq k) = \sum_{r=k}^{\min(K,n)} P(|C \cap D| = r)
\]

**Statistical enrichment of community** $C$ **in disease** $D$ : $P(|C \cap D| \geq k)$

**One-tailed Fisher’s exact test:** Probability of observing the overlap as extreme or more extreme under the null hypothesis of no association:

$$P(|C \cap D| \geq k) = \sum_{r=k}^{\min(K,n)} P(|C \cap D| = r)$$
Data:

- Human protein-protein interaction network
  - 13,460 nodes, 150,000 edges
- Human diseases
  - 70 diseases, each with at least 20 disease proteins

Community detection methods:

- Link clustering [Ahn et al., Nature 2010]
- Louvain method [Blondel et al., TE 2008]
- Markov clustering method (MCL) [Van Dongen, SIAM 2008]
Setup:

1. Use **community detection method** to find communities in the PPI network

2. Use **Fisher’s exact test** to determine, for each community-disease pair, if community is **significantly enriched** with disease proteins

3. Use **Bonferroni correction** to counteract the problem of **multiple statistical comparisons**
   - If testing $m$ hypotheses at a desired significance level $\alpha = 0.05$, then the Bonferroni correction would test each individual hypothesis at $\alpha = 0.05/m$
Protein Communities

- **Link clustering**
- **Louvain method**
- **MCL**

Note:
\[ \text{Jaccard}(C, D) = \frac{|C \cap D|}{|C \cup D|} \]

- Community-disease pairs with significant overlap versus their Jaccard similarity

---

Protein Communities

- No detected community coincides with a **full set** of disease proteins
- 36% of MCL communities are **significantly enriched** in at least one disease
- **Proteins in an enriched community** that are not yet associated with a disease are disease protein candidates
Other Statistical Issues

- Other **tests for enrichment:**
  - Binomial, Chi-squared, Z-test, Kolmogorov-Smirnov, permutation
  - Gene Set Enrichment Analysis (GSEA) uses a variation of Kolmogorov-Smirnov statistic to get p-values [http://software.broadinstitute.org/gsea]

- All tests look for **over-enrichment**; some look for **under-enrichment**

- Correction for **multiple hypothesis testing**

- Some diseases may be **subsets** of other diseases
Proteins in detected communities should have something in common, e.g., they are:
- part of the same biological pathway/cellular component
- co-expressed under certain conditions
- putative targets of the same regulatory factor

Use enrichment tests to check whether communities are enriched in biological pathways, components, etc.

Get data from biomedical databases:
- Processes, components: Gene Ontology
- Pathways: KEGG, Reactome, MSigDB
1) Very basic biology
2) Protein-protein interaction networks
3) Finding disease modules in networks
   - It is a community detection task!
4) Predicting biological attributes, such as protein functions
   - Guilt-by-association principle
   - Gene recommender systems
Functional Interaction Networks

Gene co-expression network

Cell cycle

Protein degradation

Genes

Conditions

Fraser et al., Nat Genet 2004
Mostafavi, Morris, Proteomics 2012

UNK1

CDC3

UNK2

CDC16

UNK1

RPT1

RPT1

RPT6

RPN3

CLB4

“What does my gene do?”

**Goal:** Determine a gene’s function based on who it interacts with – “guilt-by-association” principle

“Give me more genes like these”

- E.g., Find more multiple sclerosis genes, find new ciliary genes, find more members of a protein complex
“What Does My Gene Do?”

**Input**

- Network data

**Query gene**

- TP53

**Output**

- Community detection, then enrichment analysis

- “Guilt-by-association” principles

---

“Give Me More Genes Like These”

Input

Network data

Query list

Gene recommender system

Output

Networks
- Co-expression
- Shared protein domains
- Physical interactions
- Pathway
- Co-localization
- Genetic interactions

Functions
- muscle system process
- muscle contraction
- regulation of system process
- regulation of muscle system process
- heart contraction
Finding “Guilty Associates”

- Predict gene functions by guilt-by-association:
  - Query list: “positive examples”

- Question: Which of the unlabeled nodes are likely involved in this gene function/biological process?
  - Two main approaches:
    - Direct/Indirect neighbor scoring
    - Label propagation

Red: Genes involved in a gene function/biological process
White: Unlabeled genes
Let $W$ be a $n \times n$ (weighted) adjacency matrix over $n$ genes in a genome.

Let $y = \{-1, 0, 1\}^n$ be a vector of labels:

- 1: **positive** gene, known to be involved in a gene function/biological process
- -1: **negative** gene
- 0: **unlabeled** gene

**Goal:** Predict which of the unlabeled genes are likely positive.
“Guilty Associates” Problem

- **Goal:** Predict which of the unlabeled genes are likely positive
- Learn a vector of discriminant scores $f$, where $f_i$ is the likelihood that node $i$ is positive
- **Example:**

$$y = [1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$$

$W = (\text{weighted})$ adjacency matrix

$$f = ?$$
Direct Neighbor Scoring

**Approach #1:** Node score $f_i$ is the weighted sum of the labels of $i$’s direct neighbors:

$$f_i = \sum_{j=1}^{n} W_{ij}y_j$$

**Example:**

- $f_{GA} = W_{GA,MCA1} \cdot y_{MCA1}$
- $f_{GB} = W_{GB,CDC48} \cdot y_{CDC48} + W_{GB,TDH2} \cdot y_{CDC48}$
- $f_{GC} = W_{GC,TDH2} \cdot y_{TDH2}$

Red: Positive genes  
White: $f_i = 0$
Direct Neighbor Scoring

- **Approach #1:** Node score $f_i$ is a weighted sum of the labels of $i$’s direct neighbors:

\[
f_i = \sum_{j=1}^{n} W_{ij} y_j
\]

- **Example:**

- $f_{GA} = W_{GA,MCA1} \cdot y_{MCA1}$
- $f_{GB} = W_{GB,CDC48} \cdot y_{CDC48} + W_{GB,TDH2} \cdot y_{CDC48}$
- $f_{GC} = W_{GC,TDH2} \cdot y_{TDH2}$

- **One half** of GC’s neighbors are positives
- **One third** of GA’s neighbors are positives
- **But:** $f_{GC} = f_{GA}$ (if $W$ is binary)
Direct Neighbor Scoring

**Approach #2:** Normalize matrix $W$ using the weighted node degrees:

$$f_i = \frac{1}{d_i} \sum_{j=1}^{n} W_{ij} y_j, \quad d_i = \sum_j W_{ij}$$

Matrix notation:

$$f_i = D^{-1}Wy$$
$$D = \text{diag}(d)$$

**Example:**

- $f_{GA} = \frac{1}{3}W_{GA,MCA1} \cdot y_{MCA1}$
- $f_{GB} = \frac{1}{3}(W_{GB,CDC48} \cdot y_{CDC48} + W_{GB,TDH2} \cdot y_{TDH2})$
- $f_{GC} = \frac{1}{2}W_{GC,TDH2} \cdot y_{TDH2}$

Red: Positive genes
White: $f_i = 0$
Towards **Indirect Neighbor Scoring**

- **Matrix** $P = D^{-1}W$ is known as **Markov transition matrix**
  - $D$ is a diagonal matrix with diagonal elements $d_i$
  - **$P$ is a row stochastic matrix**, $\sum_j P_{ij} = 1$

- Row $i$ is a probability distribution over **random walks** starting at node $i$

- $P_{ij}$ is probability of a **random walker** following a link from node $i$ to node $j$
Random Walk Interpretation

- **Random walk** interpretation extends a direct neighbor approach to include **indirect neighbors**

- **Idea:** Extend the formula $f = D^{-1}Wy = Py$ to include **second-degree neighbors**

- Probability of a random walk of length **two** between node $i$ and node $j$ is:

$$[P^2]_{ij} = \sum_{k=1}^{n} P_{ik}P_{kj}$$
Approach #3: Consider second-degree neighbors when calculating node score $f_i$ as:

$$f_i = \sum_{j=1}^{n} P_{ij} y_j + \sum_{j=1}^{n} [P^2]_{ij} y_j$$

Direct neighbors

Second-degree neighbors
Example of Indirect Neighbor Scoring

\[ P = D^{-1}W \]

\[ f_i = \sum_{j=1}^{n} P_{ij}y_j + \sum_{j=1}^{n} [P^2]_{ij} y_j \]

Direct neighbors

Second-degree neighbors

- Direct neighbor of a positive gene
- Second-order neighbor of a positive gene
- Red: Positive genes
- White: \( f_i = 0 \)
- \( [P^2]_{ij} > 0 \) if there is a walk of length 2 between \( i \) and \( j \)
Approach #3 can be extended to include other nodes at a distance of length $r$ (usually $r < 4$).

Increasing $r$ beyond 2 often results in degradation of prediction performance [Chua et al., Bioinformatics 2006, Myers et al., Genome Biology 2005].

**Note:** Probability of a random walk from $i$ to $j$ in $r$ steps is given by $[P^r]_{ij}$.

**Next:** Use random walks to derive label propagation.
Label Propagation Approach

- Label propagation *generalizes* local neighborhood-based approaches by *considering random walks of all lengths* between nodes

- The algorithm can be derived as:
  1. Iterative diffusion process [Zhou et al., NIPS 2004]
  2. Solution to a specific convex optimization task [Zhou et al., NIPS 2004, Zhu et al., ICML 2003]
  3. Maximum a posteriori (MAP) estimation in Gaussian Markov Random Fields [Rue and Held, Chapman & Hall, 2005]

- **Next:** Derivation using an *iterative formulation*
Intuition: **Diffuse labels through edges of the network**

(a) Initial Labels

- Red: positive nodes
- White: unlabeled nodes

(b) First Iteration

- Red: positive nodes
- Pink: $f_i > 0$
- White: $f_i = 0$
The diffusion process is defined as an iterative process [Zhou et al., NIPS 2004]

Diffusion of labels through edges:

- Start with initial label information, $f_i^{(0)} = y_i$
- In each iteration, each node receives label information from its neighbors, and also retains its initial label
- $\lambda$ specifies relative amount of label information from its neighbors and its initial label
- Finally, the label of each unlabeled node is set to be the label of which it has received most information
The diffusion process is defined as the following iteration [Zhou et al., NIPS 2004]

At iteration $r = 0$, define $f_i^{(0)} \leftarrow y_i$

At iteration $r + 1$, the score of node $i$ is the weighted average of the scores of $i$’s neighbors in iteration $r$, and $i$’s initial label:

$$f_i^{(r+1)} \leftarrow (1 - \lambda)y_i + \lambda \sum_{j=1}^{n} W_{ij}f_j^{(r)}$$

$0 < \lambda < 1$ is a model parameter
Example of Label Propagation

(a) Initial Labels

\[ f^{(0)} = y \]

(b) First Iteration

\[ f^{(1)} = \lambda W y + (1 - \lambda) y \]

(c) Second Iteration

All nodes reachable with a walk of length 2 are assigned a non-zero value

\[ f^{(2)} = \lambda W f^{(1)} + (1 - \lambda) y \]

Convergence Condition

- If all eigenvalues of $W$ are in range $[-1, 1]$, then the sequence $f^{(r)}$ converges to:

\[
f = (1 - \lambda) \sum_{r=0}^{\infty} (\lambda W)^r y
\]

- $[W^r]_{ij} > 0$ if a walk of length $r$ between $i$ and $j$
- Weight $\lambda^r$ decreases with increasing distance

$\Rightarrow$ Discriminant scores $f$ are weighted sum of walks of all lengths between the nodes

$\Rightarrow$ High score $f_i$ is assigned to $i$ if $i$ is connected to positively labeled nodes with many short walks
Example of Label Propagation

(a) Initial Labels
\[ f^{(0)} = y \]

(b) First Iteration
\[ f^{(1)} = \lambda W y + (1 - \lambda) y \]

(c) Second Iteration
\[ f^{(2)} = \lambda W f^{(1)} + (1 - \lambda) y \]

(d) Final Scores
\[ f = (1 - \lambda) \sum_{r=0}^{\infty} (\lambda W)^r y \]

All nodes reachable with a walk of length 2 are assigned a non-zero value.

Recall: The infinite sum converges only if all eigenvalues of $W$ are in range $[-1, 1]$.

To satisfy this condition, normalize $W$ before diffusion:

- Symmetric normalization:
  $$ S = D^{-1/2} W D^{-1/2} $$

- Asymmetric normalization:
  $$ P = D^{-1} W $$

Note: Avoid self-reinforcement by setting diagonal elements of $W$ to 0.

Note: Label information is spread symmetrically since $S$ is a symmetric matrix.
Exact Solution of Label Propagation

- Given that $\rho(W) \leq 1$, use Taylor expansion to compute the exact solution for label propagation:

$$f = (1 - \lambda) \sum_{r=0}^{\infty} (\lambda S)^r y$$

- Note: The diffusion result $f$ does not depend on the initial value $f^{(0)}$
“Guilty Associates”: Recap

- **Direct neighbor scoring** depends on:
  - Strength of links to query genes
  - # of query gene neighbors

- **Label propagation scoring** depends on:
  - Iteratively propagating “direct neighbor score” allowing indirect links to impact scores
  - Whether or not a gene is in a connected component of genes with query genes
  - Example algorithm: GeneMANIA [Mostafavi et al., Genome Biology 2008]
Gene function prediction is a **multi-label node classification task**

Every node (gene) is assigned one or more labels (cellular functions)

**Setup:**

1. For each gene function we use a **guilt-by-association based approach** to learn a discriminative score $f_i$ for each node $i$
2. During the training phase, we observe only a certain fraction of genes and all their functions
3. The task is then to predict functions for the remaining genes

Determine the optimal value of $\lambda$ parameter using cross-validation
Gene Function Prediction: Results

Label propagation-based approaches outperform local neighborhood-based approaches

![Graph showing AUROC for different evaluation categories: GeneMANIA (15 networks), Label Propagation on BioPIXIE, Local Neighborhood on BioPIXIE. The categories include BP 3, BP 11, BP 30, and BP 101. The graph indicates that GeneMANIA generally has a higher AUROC compared to the other approaches.]
Comparison of label propagation with three normalization methods on the protein-interaction (PI) and genetic-interaction (GI) networks
GeneMANIA Tool

Query list:
- MRE11A
- RAD51
- MLH1
- MSH2
- DMC1
- RAD51AP1
- RAD50
- MSH6
- XRCC3
- PCNA
- XRCC2

Zuberi et al., NAR 2013