Graph Kernels
CPBS
07 February 2012

Sonia Leach, PhD
Assistant Professor
Center for Genes, Environment, and Health
National Jewish Health
<table>
<thead>
<tr>
<th># Individual</th>
<th>inDBSNP?</th>
<th>chr</th>
<th>position</th>
<th>ref</th>
<th>sample</th>
<th>accession</th>
<th>geneList</th>
<th>Function</th>
<th>polyPhen GVS</th>
<th>Grantham Score</th>
<th>Score PhastCons</th>
<th>consScore GER8</th>
<th>Chimp Allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>LT000091</td>
<td>1000Genomes</td>
<td>9</td>
<td>33385641</td>
<td>C</td>
<td>A/C</td>
<td>NM_001170</td>
<td>AQP7</td>
<td>intron</td>
<td>unknown</td>
<td>NA</td>
<td>0.004</td>
<td>-4.25</td>
<td>C</td>
</tr>
<tr>
<td>LT274462</td>
<td>1000Genomes</td>
<td>9</td>
<td>33385641</td>
<td>C</td>
<td>A/C</td>
<td>NM_001170</td>
<td>AQP7</td>
<td>intron</td>
<td>unknown</td>
<td>NA</td>
<td>0.004</td>
<td>-4.25</td>
<td>C</td>
</tr>
<tr>
<td>LT000091</td>
<td>none</td>
<td>12</td>
<td>58217698</td>
<td>G</td>
<td>G/T</td>
<td>NM_005730</td>
<td>CTDSP2</td>
<td>missense</td>
<td>unknown</td>
<td>38</td>
<td>1</td>
<td>4.97</td>
<td>G</td>
</tr>
<tr>
<td>LT274462</td>
<td>none</td>
<td>12</td>
<td>58217698</td>
<td>G</td>
<td>G/T</td>
<td>NM_005730</td>
<td>CTDSP2</td>
<td>missense</td>
<td>unknown</td>
<td>38</td>
<td>1</td>
<td>4.97</td>
<td>G</td>
</tr>
<tr>
<td>LT000091</td>
<td>none</td>
<td>3</td>
<td>49724798</td>
<td>T</td>
<td>G/T</td>
<td>NM_020998</td>
<td>MST1</td>
<td>missense</td>
<td>benign</td>
<td>53</td>
<td>0.968</td>
<td>5.07</td>
<td>T</td>
</tr>
<tr>
<td>LT274462</td>
<td>none</td>
<td>3</td>
<td>49724798</td>
<td>T</td>
<td>G/T</td>
<td>NM_020998</td>
<td>MST1</td>
<td>missense</td>
<td>benign</td>
<td>53</td>
<td>0.968</td>
<td>5.07</td>
<td>T</td>
</tr>
<tr>
<td>LT000091</td>
<td>none</td>
<td>11</td>
<td>1.23E+08</td>
<td>T</td>
<td>G/T</td>
<td>NM_032873</td>
<td>UBASH3B</td>
<td>missense</td>
<td>unknown</td>
<td>205</td>
<td>1</td>
<td>NA</td>
<td>T</td>
</tr>
<tr>
<td>LT274462</td>
<td>none</td>
<td>11</td>
<td>1.23E+08</td>
<td>T</td>
<td>G/T</td>
<td>NM_032873</td>
<td>UBASH3B</td>
<td>missense</td>
<td>unknown</td>
<td>205</td>
<td>1</td>
<td>NA</td>
<td>T</td>
</tr>
<tr>
<td>LT000091</td>
<td>none</td>
<td>X</td>
<td>64986699</td>
<td>G</td>
<td>A/G</td>
<td>NM_002444</td>
<td>MSN</td>
<td>coding- synonymous</td>
<td>unknown</td>
<td>0.998</td>
<td>0.997</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>LT274462</td>
<td>none</td>
<td>X</td>
<td>64986699</td>
<td>G</td>
<td>A/G</td>
<td>NM_002444</td>
<td>MSN</td>
<td>coding- synonymous</td>
<td>unknown</td>
<td>0.998</td>
<td>0.997</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>LT000091</td>
<td>none</td>
<td>12</td>
<td>58217696</td>
<td>G</td>
<td>G/T</td>
<td>NM_005730</td>
<td>CTDSP2</td>
<td>coding- synonymous</td>
<td>unknown</td>
<td>0.903</td>
<td>-6.66</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>LT274462</td>
<td>none</td>
<td>12</td>
<td>58217696</td>
<td>G</td>
<td>G/T</td>
<td>NM_005730</td>
<td>CTDSP2</td>
<td>coding- synonymous</td>
<td>unknown</td>
<td>0.903</td>
<td>-6.66</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>LT000091</td>
<td>none</td>
<td>19</td>
<td>55284908</td>
<td>T</td>
<td>C/T</td>
<td>NM_014218</td>
<td>KIR2DL1</td>
<td>missense</td>
<td>unknown</td>
<td>81</td>
<td>0</td>
<td>-0.545</td>
<td>A</td>
</tr>
<tr>
<td>LT274462</td>
<td>none</td>
<td>19</td>
<td>55284908</td>
<td>T</td>
<td>C/T</td>
<td>NM_014218</td>
<td>KIR2DL1</td>
<td>missense</td>
<td>unknown</td>
<td>81</td>
<td>0</td>
<td>-0.545</td>
<td>A</td>
</tr>
<tr>
<td>LT000091</td>
<td>none</td>
<td>12</td>
<td>58220823</td>
<td>C</td>
<td>C/T</td>
<td>NM_005730</td>
<td>CTDSP2</td>
<td>missense</td>
<td>unknown</td>
<td>21</td>
<td>1</td>
<td>4.87</td>
<td>C</td>
</tr>
<tr>
<td>LT274462</td>
<td>none</td>
<td>12</td>
<td>58220823</td>
<td>C</td>
<td>C/T</td>
<td>NM_005730</td>
<td>CTDSP2</td>
<td>missense</td>
<td>unknown</td>
<td>21</td>
<td>1</td>
<td>4.87</td>
<td>C</td>
</tr>
<tr>
<td>LT000091</td>
<td>none</td>
<td>15</td>
<td>80215245</td>
<td>A</td>
<td>A/C</td>
<td>NM_175898</td>
<td>C10orf37,ST20</td>
<td>missense</td>
<td>unknown</td>
<td>68</td>
<td>0.012</td>
<td>0.278</td>
<td>A</td>
</tr>
<tr>
<td>LT274462</td>
<td>none</td>
<td>15</td>
<td>80215245</td>
<td>A</td>
<td>A/C</td>
<td>NM_175898</td>
<td>C10orf37,ST20</td>
<td>missense</td>
<td>unknown</td>
<td>68</td>
<td>0.012</td>
<td>0.278</td>
<td>A</td>
</tr>
</tbody>
</table>
How are they related?
No direct interactions known

This is the evidence view. Different line colors represent the types of evidence for the association.

Your Input:
- **UBASH3B**: Suppressor of T-cell receptor signaling 1 (Cts-1) (Cbl-interacting protein p70); Interferes with CBL-mediated down-regulation and degradation of receptor type tyrosine kinases. Promotes accumulation of activated tyrosine kinases, such as T-cell receptors and EGFR, on the cell surface (544 aa).
- **AQP7**: Aquaporin-7 (AQP-7) (Aquaporin-7-like) (Aquaporin 7-like) (AQ2ap); Forms a channel for water and glycerol (517 aa).
- **CTDSP2**: Carboxy-terminal domain DNA polymerase II polypeptide A small polypeptide 2 (E3.1.3.16) (Small CTD phosphatase 2) (SCP2) (Nuclear LIM interacting-interacting factor 2) (NLI-interacting factor 2) (Protein OS-4); May contribute to the development of sarcomas (382 aa).
- **KIR2DL1**: Killer cell immunoglobulin-like receptor 2DL1 precursor (MHC class I NK cell receptor) (Natural killer-associated transcript 4) (NKAT-4); (70 natural killer cell receptor domain CL-5) (CIDEBK4 antigen); Receptor on natural killer (NK) cells for HLA-C alleles (HLA-Cw4, HLA-Cw5 and HLA-Cw7); Inhibits the activity of NK cells thus preventing cell lysis (456 aa).
- **MSN**: Membranin (Membrane-organizing extension spike protein); Probably involved in connections of major cytoskeletal structures to the plasma membrane (277 aa).
- **MST1**: Hepatocyte growth factor-like protein pseudouracil (Nucleotide-stimulating protein) (MSP) (Nucleotide-stimulating protein) (Contains: Hepatocyte growth factor-like protein alpha chain; Hepatocyte growth factor-like protein beta chain); Probably has no proteolytic activity, since crucial characteristic of serine proteases catalytic sites are not conserved (711 aa) (Homo sapiens).
More of the network

Center for Genes, Environment, and Health
What do we do with that?!!?

• Can we tell how ‘close’ two nodes are in the network?
  – One link, two link, three links…..?
    • Known as ‘shortest path distance’

• Fact: the diameter of the STRING network for all human proteins is 5
  – >80% of the network is within 2 links from each other

• By this measure, everything is ‘close’ to each other
• Shortest path = 2 in all graphs: A) connects x and y by hub while B) does not and C) best situation because likely true interaction

• Need measure that looks beyond a single path - rather need one that looks at GLOBAL topology

from Kohler et al., 2008
Distance measures on a graph

- ‘Random Walker’ – begin walk in any node, and with probability proportional to degree of node, move to adjacent node, ad infinitum

\[
\begin{bmatrix}
1 & 1 & 1 \\
1 & 1 & 0 \\
0 & 1 & 0 \\
\end{bmatrix}
\]

Adjacency Matrix A

\[
\begin{bmatrix}
1/3 & 1/3 & 1/3 \\
1/2 & 1/2 & 0 \\
0 & 1 & 0 \\
\end{bmatrix}
\]

Probability Matrix P

- ‘Closeness’ of nodes i,j is proportion of time spent in node j when started from node i
What is probability that start in state 1, go to another state, and end up in state 1?

1→1→1 OR 1→2→1 OR 1→3→1

\[
\frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times 0 =
\]

0.11 + 0.1667 + 0 = 0.2778
Distance measures on a graph

What is probability that start in state 2, go to another state, and end up in state 1?

2→1 → 1 OR 2→2→1 OR 2→3→1

\[
\begin{array}{c|ccc}
\text{State} & 1 & 2 & 3 \\
\hline
1 & 1/3 & 1/3 & 1/3 \\
2 & 1/2 & 1/2 & 0 \\
3 & 0 & 1 & 0 \\
\end{array}
\]

Probability Matrix P

\[
\begin{align*}
\frac{1}{2} \times \frac{1}{3} & + \frac{1}{2} \times \frac{1}{2} + 0 \times 0 = \\
0.1667 & + 0.25 + 0 = \\
0.4167
\end{align*}
\]
**Distance measures on a graph**

What is probability that start in state \( 1 \), go to another state, and end up in state \( 1 \)?

1→1→1 OR 1→2→1 OR 1→3→1

\[
\frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times 0 = \]

\[
\begin{bmatrix}
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\
1 & 2 & 0 \\
0 & 1 & 0
\end{bmatrix}
\]

**Probability Matrix P**

```
State 1 2 3
1 1/3 1/3 1/3
2 1/2 1/2 0
3 0 1 0
```
What is probability that start in state 1, go to another state, and end up in state 1?

\[ \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/2 & 1/2 & 0 \\ 0 & 1 & 0 \end{bmatrix} \]

\[ \begin{bmatrix} 1/3 \\ 1/2 \\ 0 \end{bmatrix} \]

Center for Genes, Environment, and Health
Distance measures on a graph

What is probability that start in state 2, go to another state, and end up in state 1?

$$2 \rightarrow 1 \rightarrow 1 \text{ OR } 2 \rightarrow 2 \rightarrow 1 \text{ OR } 2 \rightarrow 3 \rightarrow 1$$

$$\frac{1}{2} \times \frac{1}{3} + \frac{1}{2} \times \frac{1}{2} + 0 \times 0 = \frac{3}{14}$$

Center for Genes, Environment, and Health
Distance measures on a graph

What is probability that start in state $i$, go to another state, and end up in state $j$?

$$
\begin{bmatrix}
1 \sim 1 & ? & ? \\
2 \sim 1 & ? & ? \\
? & ? & ?
\end{bmatrix}
\begin{bmatrix}
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\
\frac{1}{2} & \frac{1}{2} & 0 \\
0 & 1 & 0
\end{bmatrix}
= 
\begin{bmatrix}
0.2778 & ? & ? \\
0.4167 & ? & ? \\
? & ? & ?
\end{bmatrix}
$$

State | 1 | 2 | 3
--- | --- | --- | ---
1 | 1/3 | 1/3 | 1/3
2 | 1/2 | 1/2 | 0
3 | 0 | 1 | 0

Probability Matrix $P$
Distance measures on a graph

What is probability that start in state $i$, go to another state, and end up in state $j$?

\[
\begin{bmatrix}
1 \sim 1 & 1 \sim 2 & 1 \sim 3 \\
2 \sim 1 & 2 \sim 2 & 2 \sim 3 \\
3 \sim 1 & 3 \sim 2 & 3 \sim 3 \\
\end{bmatrix} =
\begin{bmatrix}
0.2778 & 0.6111 & 0.1111 \\
0.4167 & 0.4167 & 0.1667 \\
0.5 & 0.5 & 0
\end{bmatrix}
\]
Distance measures on a graph

What is probability that start in state \( i \), go to another state, and end up in state \( j \)?

\[
\begin{bmatrix}
1 \sim 1 & ? & ? \\
2 \sim 1 & ? & ? \\
? & ? & ?
\end{bmatrix} = \begin{bmatrix}
0.2778 & ? & ? \\
0.4167 & ? & ? \\
? & ? & ?
\end{bmatrix}
\]

State	1	2	3
\begin{bmatrix}
1/3 & 1/3 & 1/3 \\
1/2 & 1/2 & 0 \\
0 & 1 & 0
\end{bmatrix}

Probability Matrix \( P \)

\[
\begin{bmatrix}
1 \sim 1 & ? & ? \\
2 \sim 1 & ? & ? \\
? & ? & ?
\end{bmatrix} = \begin{bmatrix}
p_{11}p_{11} + p_{12}p_{21} + p_{13}p_{31} & ? & ? \\
p_{21}p_{11} + p_{22}p_{21} + p_{23}p_{31} & ? & ? \\
? & ? & ?
\end{bmatrix} = P \times P
\]
Distance measures on a graph

1-step probabilities = $P = P^1$
2-step probabilities = $P \times P = P^2$
3-step probabilities = $P \times P \times P = P^2 \times P = P^3$

$n$-step probabilities = $P^{n-1} \times P = P^n$

$\infty$-step probabilities = $P^\infty$
Distance measures on a graph

What is probability that start in state i, go to another state, and end up in state j?

What is probability that start in state i and end up in state j?

Answer: prob start in i AND end up in j in 1 step, OR end up in j in 2 steps OR end up in j in 3 steps OR .... end up in j in infinite steps
Computing Measures

\[ \sum_{k=0}^{\infty} P^k = \frac{1}{(I - P)} = (I - P)^{-1} \]

is a generalization of the geometric series for matrices (versus scalar version below)

\[ \sum_{k=1}^{\infty} ar^{i-1} = \frac{a}{(1 - r)} \text{ converges if } |r| < 1 \]
Computing Measures

\[
\sum_{k=0}^{\infty} P^k = \frac{1}{(I - P)} = (I - P)^{-1}
\]

Identity matrix = matrix of all 1s on diagonal
Computing Measures

\[ \sum_{k=0}^{\infty} P^k = (I - P)^{-1} \]

(Neumann Series (also stationary distribution of Markov Chain))
Computing Measures

\[
\sum_{k=0}^{\infty} P^k = (I - P)^{-1}
\]
(Neumann Series)

\[
\sum_{k=0}^{\infty} \alpha^k P^k = (I - \alpha P)^{-1}
\]
Computing Measures

\[
\sum_{k=0}^{\infty} P^k = (I - P)^{-1}
\]

(Neumann Series)

\[
\sum_{k=0}^{\infty} \alpha^k P^k = (I - \alpha P)^{-1}
\]

(von Neumann diffusion kernel)
Computing Measures

\[ \sum_{k=0}^{\infty} P^k = (I - P)^{-1} \]  
(Neumann Series)

\[ \sum_{k=0}^{\infty} \alpha^k P^k = (I - \alpha P)^{-1} \]  
(von Neumann diffusion kernel)

\[ \sum_{k=0}^{\infty} \frac{\alpha^k P^k}{k!} = ??? \]
Computing Measures

\[ \sum_{k=0}^{\infty} P^k = (I - P)^{-1} \]  
(Neumann Series)

\[ \sum_{k=0}^{\infty} \alpha^k P^k = (I - \alpha P)^{-1} \]  
(von Neumann diffusion kernel)

\[ \sum_{k=0}^{\infty} \frac{\alpha^k P^k}{k!} = 1 + \frac{\alpha P}{1!} + \frac{\alpha^2 P^2}{2!} + \frac{\alpha^3 P^3}{3!} + \ldots = \exp(\alpha P) \]  
(exponential diffusion kernel by Taylor series expansion)
Center for Genes, Environment, and Health
Center for Genes, Environment, and Health
Center for Genes, Environment, and Health
Center for Genes, Environment, and Health
Center for Genes, Environment, and Health
But... ummmm... why do we care again?
Do I have to care about ALL of these edges and nodes? Can’t I just keep those that are relevant to my initial input nodes?
Relevant Subgraph Extraction

• Input: weighted (directed) graph and set of ‘seed nodes’

• Output: subset of graph connecting relevant nodes
Relevant Subgraph Extraction
Relevant Subgraph Extraction
Another example
Relevant Subgraph Extraction:

Absorbing Markov Chain

\[ P = \begin{pmatrix}
.00 & .25 & .25 & .25 & .25 \\
.25 & .00 & .25 & .25 & .25 \\
.25 & .25 & .00 & .25 & .25 \\
.25 & .25 & .25 & .00 & .25 \\
.25 & .25 & .25 & .25 & .00 \\
\end{pmatrix} \]
Relevant Subgraph Extraction:
Absorbing Markov Chain

\[ x P = \begin{pmatrix}
0.00 & 0.25 & 0.25 & 0.25 & 0.25 \\
0.25 & 0.00 & 0.25 & 0.25 & 0.25 \\
0.25 & 0.25 & 0.00 & 0.25 & 0.25 \\
0.25 & 0.25 & 0.25 & 0.00 & 0.25 \\
0.25 & 0.25 & 0.25 & 0.25 & 0.00
\end{pmatrix} \]

All query nodes but \( x \) are ‘absorbing’ states
Relevant Subgraph Extraction:
Absorbing Markov Chain

All query nodes but $x$ are 'absorbing' states

$xP = \begin{pmatrix}
0 & 0.25 & 0.25 & 0.25 & 0.25 \\
0.25 & 0 & 0.25 & 0.25 & 0.25 \\
0.25 & 0.25 & 0 & 0.25 & 0.25 \\
0.25 & 0.25 & 0.00 & 0.25 & 0.25 \\
0.25 & 0.25 & 0.25 & 0.25 & 0.00
\end{pmatrix}$

reorder

$xP = \begin{pmatrix}
0.00 & 0.25 & 0.25 & 0.25 & 0.25 \\
0.25 & 0.00 & 0.25 & 0.25 & 0.25 \\
0.00 & 0.00 & 1 & 0 & 0 \\
0.00 & 0.00 & 0 & 1 & 0 \\
0.25 & 0.25 & 0.25 & 0.25 & 0.00
\end{pmatrix}$
Relevant Subgraph Extraction:
Absorbing Markov Chain

$P = \begin{pmatrix}
1 & 2 & 3 & 4 & 5 \\
.00 & .25 & .25 & .25 & .25 \\
.25 & .00 & .25 & .25 & .25 \\
.25 & .25 & .00 & .25 & .25 \\
.25 & .25 & .25 & .00 & .25 \\
.25 & .25 & .25 & .25 & .00 \\
\end{pmatrix}$

$xP = \begin{pmatrix}
\begin{pmatrix} xQ & xR \end{pmatrix} \\
0 & I \\
\end{pmatrix}$

$xP = \begin{pmatrix}
1 & 2 & 3 & 4 & 5 \\
.00 & .25 & .25 & .25 & .25 \\
.25 & .00 & .25 & .25 & .25 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
.25 & .25 & .25 & .25 & .00 \\
\end{pmatrix}$

All query nodes but $x$ are ‘absorbing’ states

reorder
Create $^{3}\text{P}$

$xP = \begin{bmatrix} xQ & xR \\ 0 & I \end{bmatrix}$

All query nodes but $x$ are ‘absorbing’ states

$P = \begin{bmatrix} 1 & 2 & \ldots & 11 \\ \? & \? & \ldots & \? \end{bmatrix}$

$xP = \begin{bmatrix} 1 & 2 & 3 & \ldots & 11 \\ 2 & 3 & \ldots & \? & \? \end{bmatrix}$

reorder

$xP = \begin{bmatrix} 1 & \ldots & 11 & 3 & 7 & 10 \\ \? & \? & \ldots & \? & \? & \? \end{bmatrix}$
Create Transition Probability Matrix

\[ P = \]

\[
\begin{array}{ccccccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\
1 & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} \\
2 & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} \\
3 & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} \\
4 & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} \\
5 & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} \\
6 & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} \\
7 & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} \\
8 & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} \\
9 & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} \\
10 & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} \\
11 & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} \\
\end{array}
\]
Create Absorbing Markov Chain

All query nodes but $x$ are ‘absorbing’ states

$$xP = \begin{array}{ccccccc}
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots 
\end{array}$$
Reorder Matrix

\[ xP = \begin{bmatrix} xQ & xR \\ 0 & I \end{bmatrix} \]
But what is $\times P$?

$$\times P = \begin{bmatrix} \times Q & \times R \\ 0 & I \end{bmatrix}$$

$\times R$ = transition probability matrix to move from $Q$ set to remaining query nodes

$I$ = transition probability matrix to move around among remaining query nodes

$\times Q$ = transition probability matrix to move around among non-query nodes (and $\times$)
Relevant Subgraph Extraction

- $E[n(x, i)] = \text{expected number of times start walk in state } x, \text{ and visit state } i$

$$P_{ij} = \Pr(x_{t+1} = j \mid x_t = i)$$

$$E[n(x, i)] = \Pr(x_0 = i \mid x_0 = x) + \Pr(x_1 = i \mid x_0 = x)$$

$$+ \Pr(x_2 = i \mid x_0 = x) \ldots$$

$$E[n(x, i)] = \sum_{l=0}^{\infty} \Pr(x_l = i \mid x_0 = x) = \sum_{l=0}^{\infty} \left[ (xQ)^l \right]_{xi} = \left[ I - xQ \right]^{-1}_{xi} = xN_{xi}$$
Relevant Subgraph Extraction

- $E[n(x, i)] = \text{expected number of times start walk in state } x, \text{ and visit state } i$
- $x E(i, j) = \text{expected number of times start walk in state } x \text{ and visit edge } i, j$

\[
P_{ij} = \Pr(x_{t+1} = j \mid x_t = i) \quad x P = \begin{bmatrix} xQ & xR \\ 0 & I \end{bmatrix}
\]

\[
E[n(x, i)] = \sum_{l=0}^{\infty} \Pr(x_l = i \mid x_0 = x) = \sum_{l=0}^{\infty} (xQ)^l = I^{-xQ} x_i = xN_{xi}
\]

\[
x E(i, j) = x N_{xi} x P_{ij}
\]
Relevant Subgraph Extraction

- \( E[n(x, i)] \) = expected number of times start walk in state \( x \), and visit state \( i \)
- \( ^x E(i, j) \) = expected number of times start walk in state \( x \) and visit edge \( i, j \)

\[
P_{ij} = \Pr(x_{t+1} = j \mid x_t = i)
\]

\[
E[n(x, i)] = \sum_{l=0}^{\infty} \Pr(x_l = i \mid x_0 = x) = \sum_{l=0}^{\infty} \left( ^x Q \right)^l = \left[ I - ^x Q \right]^{-1} = ^x N_{xi}
\]

\[
^x E(i, j) = ^x N_{xi} \times P_{ij}
\]
Relevant Subgraph Extraction

- **Random Walk-based**
  
  \[ P_{ij} = \Pr(x_{t+1} = j \mid x_t = i) \quad \quad xP = \text{rows and columns of } P \text{ pertaining to } x \]

  \[ E[n(x,i)] = \sum_{l=0}^{\infty} \Pr(x_l = i \mid x_0 = x) = \sum_{l=0}^{\infty} (xP)^l = [1-xP]_x^{-1} = xN_{xi} \]

  \[ xE(i,j) = xN_{xi} xP_{ij} \]

- **Graph Kernel-based**

  von Neumann \[ \sum_{k=0}^{\infty} \alpha^k A^k = [I - \alpha A]^{-1} \]

  diffusion kernel \[ \sum_{k=0}^{\infty} \frac{\alpha^k A^k}{k!} = e^{\alpha A} \]

  or \[ e^{-\beta L} \quad \text{where Laplacian } L = D - A \]
Measures and Topology

- Shortest path = 2 in all graphs: A) connects x and y by hub while B) does not and C) best situation because likely true interaction.
- Diffusion kernel and Random-walk based measures correct for GLOBAL topology.
- For sufficiently small $\beta$, the DK can be seen as lazy RW where go to neighbor with prob $\beta$, or remain in current node with prob $1-d_i \beta$.

from Kohler et al., 2008
References

• Random k-walks
  – http://bioinformatics.oxfordjournals.org/content/26/9/1211.full Faust et al

• Graph Kernels