Clustering Uncertain Graphs

Haoqiu Wu
Joyce Epp
Junyuan Leng
Yuwei Guo

2018.4.4

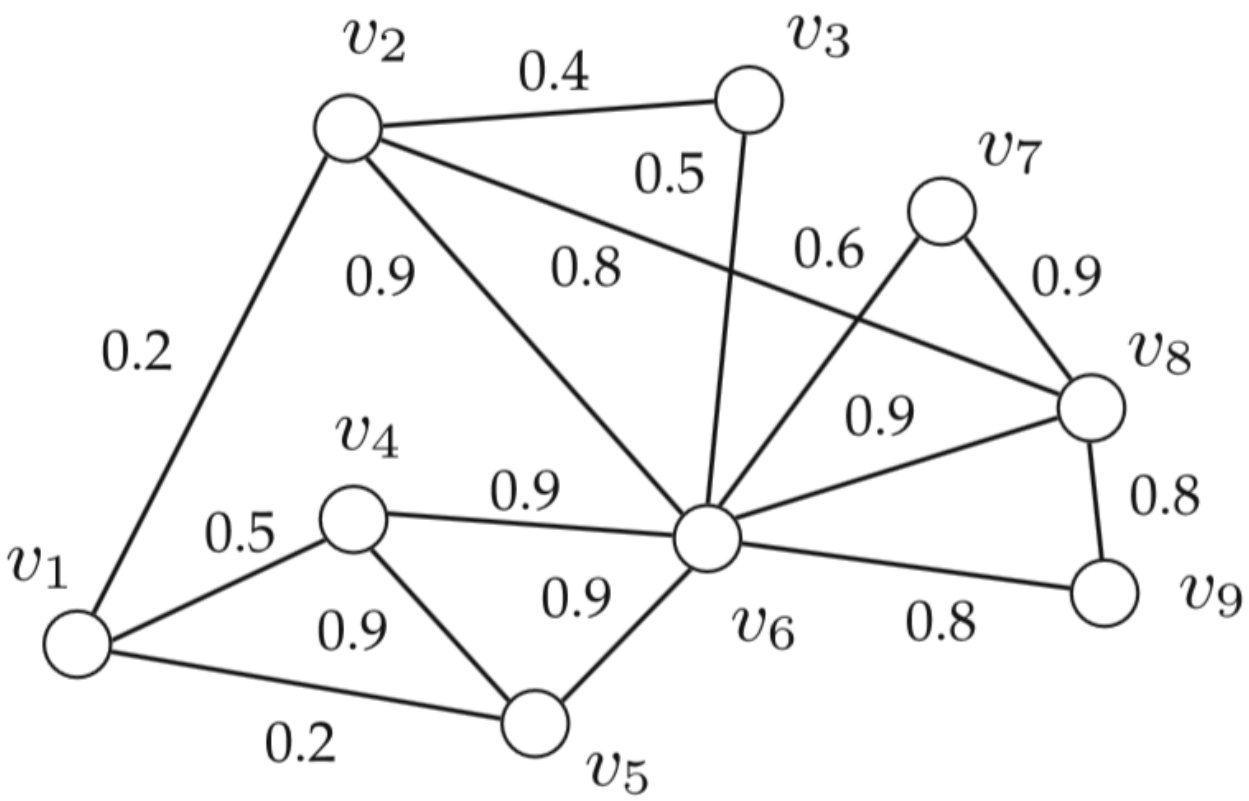
Outline

- Literature Review
- Datasets
- Algorithms
- Conclusion

Literature Review

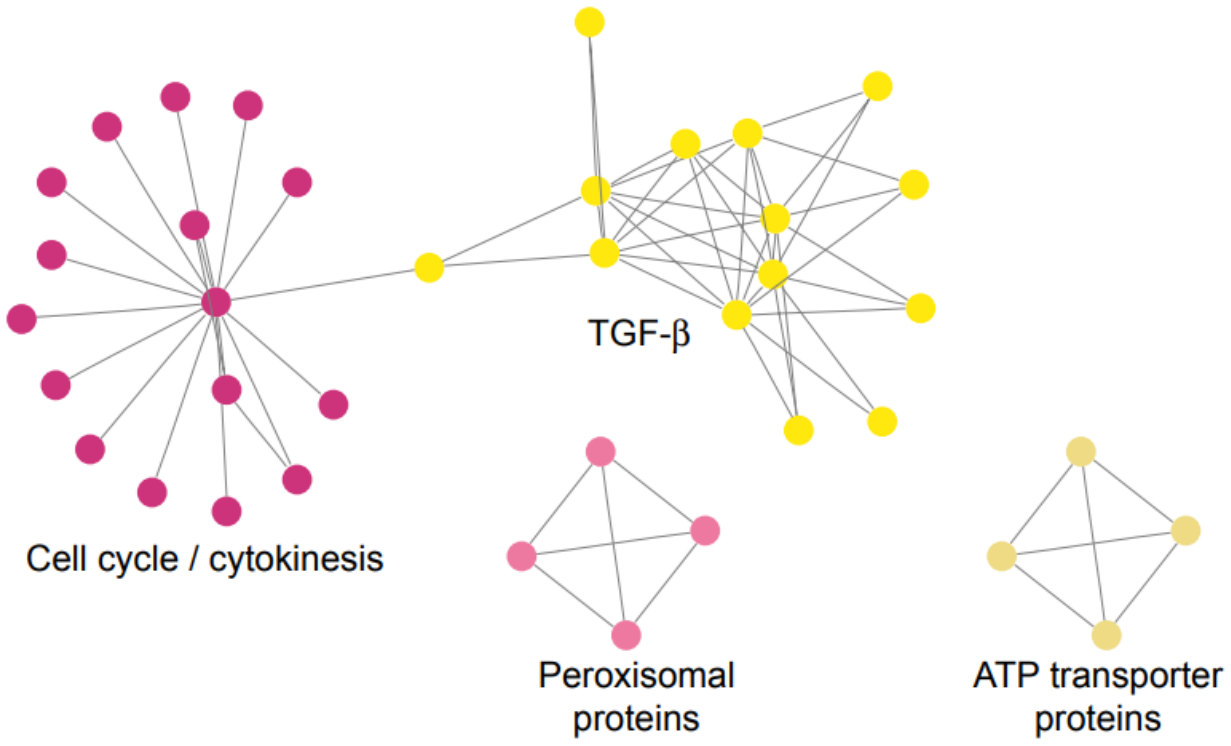
- Uncertain graphs are a special type of graph where there is uncertainty with respect to the existence of an edge

- Defined as $G(V, E, p)$ where V is the set of nodes, E is the set of edges and p is the probability of an edge $p : E \rightarrow (0, 1]$



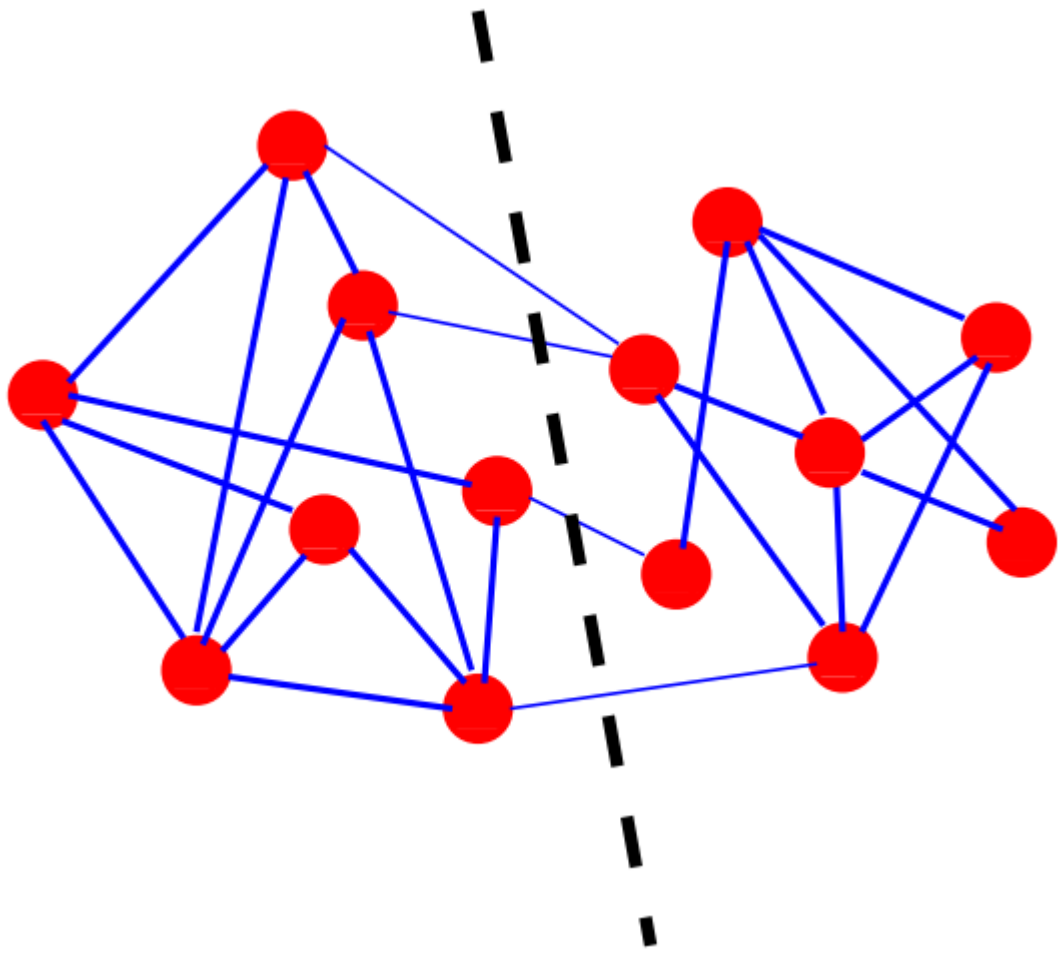
Literature Review

- Assign probabilities to edges in a social network, such as the probability that user *v* sends a message to user *u*
- Assign a probability of interaction in protein-protein interactions
- In general, it allows us to incorporate uncertainty into the network



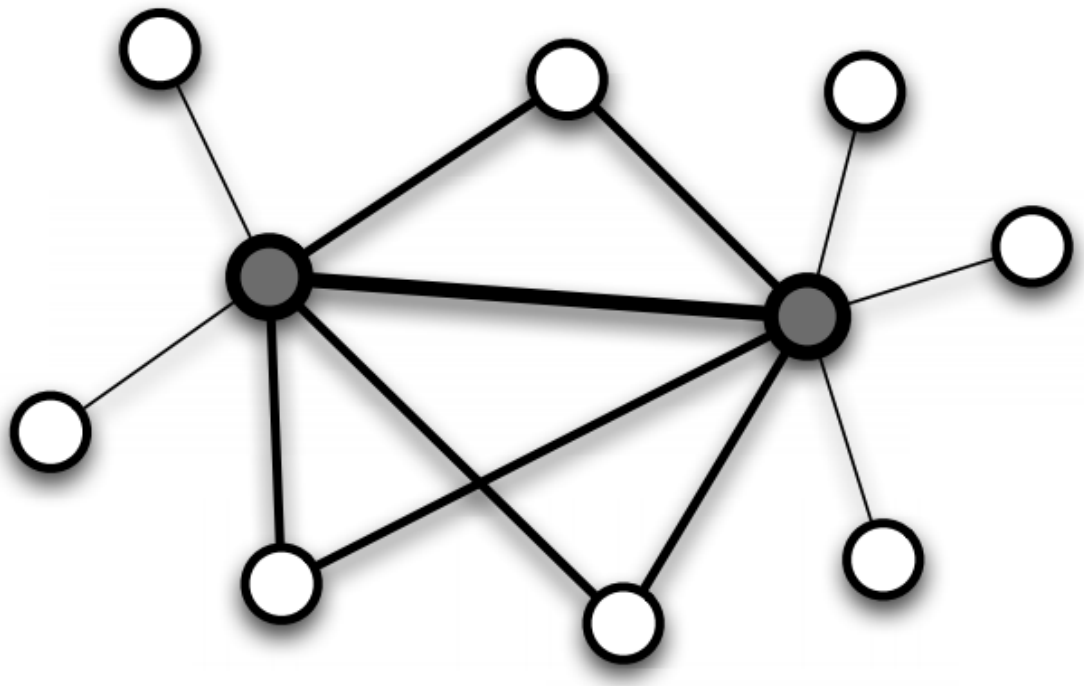
Literature Review

- Clustering puts objects that are similar into a cluster and puts dissimilar objects into different clusters
- Comes from the notion of community detection
- Traditional methods include graph partitioning, hierarchical clustering, partitional clustering and spectral clustering



Literature Review

- Probabilities make the algorithms more complex
- Treating probabilities as weights or ignoring small probabilities does not work
- Cannot treat uncertain graphs as deterministic, so we need to adapt or create new algorithms specific for uncertain graphs



Literature Review

- Edit distance between deterministic graph G and Q is defined as the number of edges that need to be added or removed from graph G to get graph Q.
  - Want to minimize the edit distance
  - ClusterEdit problem

$$D(G,Q) = |E_G \setminus E_Q| + |E_Q \setminus E_G|$$

$$D(G,Q) = \sum_{\substack{u=1,\\ v<u}}^{n} |\mathbf{G}(u,v) - \mathbf{Q}(u,v)|$$

G, Q is the 0-1 adjacency matrix

- Edit distance between probabilistic graph $\mathcal{G}$ and deterministic graph Q is defined as the expected edit distance between every Gϵ $\mathcal{G}$ and Q.
  - Want to minimize the expected edit distance
  - pClusterEdit problem

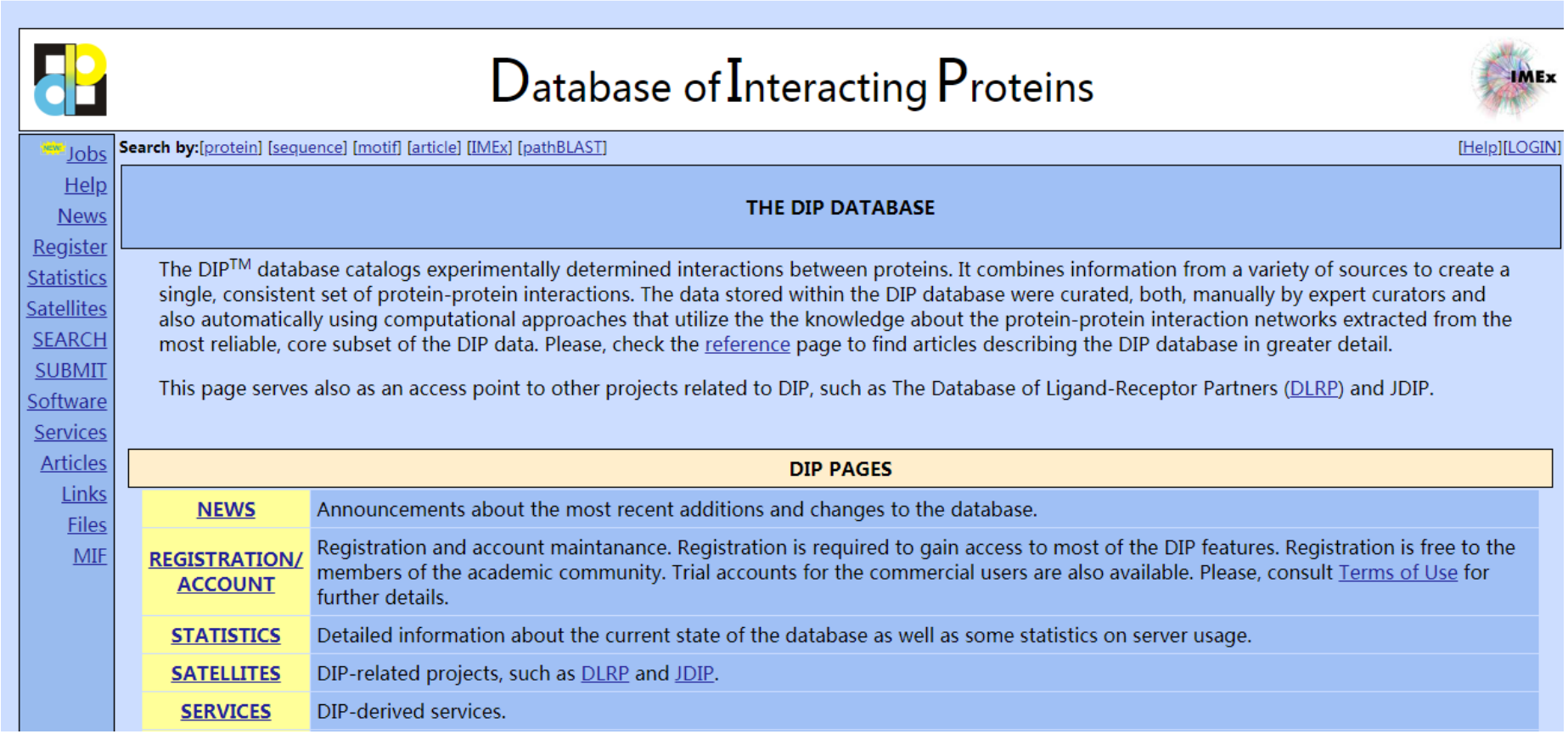
This ensures nodes in the same cluster should have large interaction: P-> 1

This ensures nodes in different clusters should have small interaction: P-> 0

$$D(\mathcal{G},Q) = \mathop{\mathbb{E}}_{G \sqsubseteq \mathcal{G}} \left[\sum_{\substack{u=1\\ v<u}}^{n} X_{uv} \right] = \sum_{\{u,v\} \in E_Q} (1 - P_{uv}) + \sum_{\{u,v\} \notin E_Q} P_{uv}$$

Datasets

Core PPI network: http://dip.doe-mbi.ucla.edu/dip/Main.cgi



Database of Interacting Proteins

Jobs
Help
News
Register
Statistics
Satellites
SEARCH
SUBMIT
Software
Services
Articles
Links
Files
MIF

Search by:[protein] [sequence] [motif] [article] [IMEx] [pathBLAST] [Help][LOGIN]

THE DIP DATABASE

The DIP™ database catalogs experimentally determined interactions between proteins. It combines information from a variety of sources to create a single, consistent set of protein-protein interactions. The data stored within the DIP database were curated, both, manually by expert curators and also automatically using computational approaches that utilize the the knowledge about the protein-protein interaction networks extracted from the most reliable, core subset of the DIP data. Please, check the reference page to find articles describing the DIP database in greater detail.

This page serves also as an access point to other projects related to DIP, such as The Database of Ligand-Receptor Partners (DLRP) and JDIP.

DIP PAGES

NEWS	Announcements about the most recent additions and changes to the database.
REGISTRATION/ ACCOUNT	Registration and account maintanance. Registration is required to gain access to most of the DIP features. Registration is free to the members of the academic community. Trial accounts for the commercial users are also available. Please, consult Terms of Use for further details.
STATISTICS	Detailed information about the current state of the database as well as some statistics on server usage.
SATELLITES	DIP-related projects, such as DLRP and JDIP.
SERVICES	DIP-derived services.

Datasets

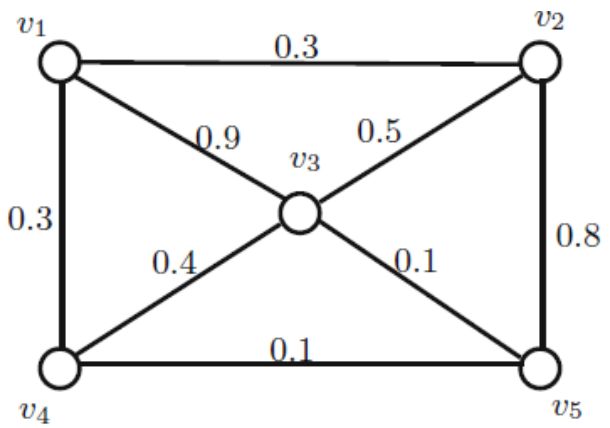
Core PPI network: http://dip.doe-mbi.ucla.edu/dip/Main.cgi

- contains 2708 nodes that represents proteins and 7123 edges.
- The edge probabilities show how likely it is that the interaction actually happens between two proteins.
- About 20% of the edges have probability over 0.98.
- The remaining edge probabilities are uniformly distributed in the remaining range [0.27, 0.98].
- characterized by power-law degree distribution, short paths and high clustering coefficient.

Algorithms

Algorithm 1. pKwikCluster algorithm for probabilistic graph clustering.

repeat
  Choose $u \in V$ **randomly**
  $C(u) \leftarrow u$
  for all $v \in V$ such that $p(u, v) \geq 0.5$ **do**
    $C(u) \leftarrow C(u) \cup v$
  end for
  $V \leftarrow V - C(u)$
until $V = \emptyset$



. A simple probabilistic graph $\mathcal{G}$ with 5 nodes and 8 edges

- Implemented using Java
- Randomized algorithm for the pClusterEdit problem

Furthest

- Top-down iterative algorithm
- All nodes start in the same cluster, add a new center at each iteration and assign remaining nodes to the nearest center

Algorithm 2. FURTHEST algorithm for probabilistic graph clustering.

repeat
  $\mathcal{C} \leftarrow \emptyset$, $\mathcal{C} \subset V$ is the set of nodes acting as cluster centers
  for all $u \in V$ **do**
    $C(u) \leftarrow u$
  end for
  First iteration:
   $\mathcal{C} \leftarrow \mathcal{C} \cup \{c_1, c_2\}$, such that $c_1, c_2 \in V - \mathcal{C}$ and $p(c_1, c_2)$ is minimum
  i-th iteration:
   $\mathcal{C} \leftarrow \mathcal{C} \cup c_i$, such that $c_i \in V - \mathcal{C}$ and the probability between c_i and members of $\mathcal{C}$ is minimum
  for all $u \in V - \mathcal{C}$ **do**
    Assign u to the cluster with which it is more probable to share an edge
    $V \leftarrow V - \{u\}$
  end for
until $V \leftarrow \emptyset$

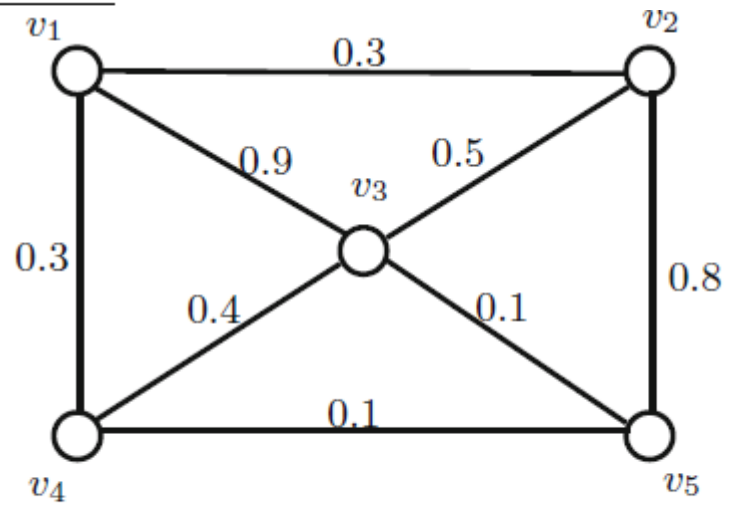
Algorithms

Bottom-up Iterative algorithm

Algorithm 3. Agglomerative algorithm for probabilistic graph clustering.

```
repeat
    for all u ∈ V do
        u forms a singleton cluster
    end for
    for all pairs of clusters do
        Find pair with maximum average edge probability p_ae
        if p_ae ≥ 0.5 then
            Merge pair of clusters into one and continue
        else
            Stop and display current clustering
        end if
    end for
until p_ae < 0.5
```

$$\frac{1}{|Vi||Vj|} \sum_{u \in V_i, v \in V_j} P_{uv}$$



. A simple probabilistic graph $\mathcal{G}$ with 5 nodes and 8 edges

*Details on board

Results

Algorithm	# of cluster (all)	Edit distance (all)	# of cluster (nonsingleton)	Edit distance (nonsingleton)	# of nodes in biggest cluster	Complexity
pKwikCluster	1276	5188	575	4465	23	O(n)
Furthest	1386	5139.02	610	4729.92	26	$O(km^2)$/ O(km)
Agglomerative	1775	3428.17	542	2277.84	38	O(km2)/ O(kmlogm)

- Efficiency/Scalability
- Cluster quality

Conclusion

- Give a graph edit distance based definition of clustering in probabilistic graphs
- Implement three efficient algorithms for clustering large probabilistic graphs
- Test our algorithms on real probabilistic protein-protein interaction network
- Our algorithm discover clusters and identify interaction relationship among proteins

The End

Thank You!