Finding representative nodes in probabilistic graphs

Laura Langohr and Hannu Toivonen

Department of Computer Science and
Helsinki Institute for Information Technology HIIT,
University of Helsinki, Finland

September 11, 2009

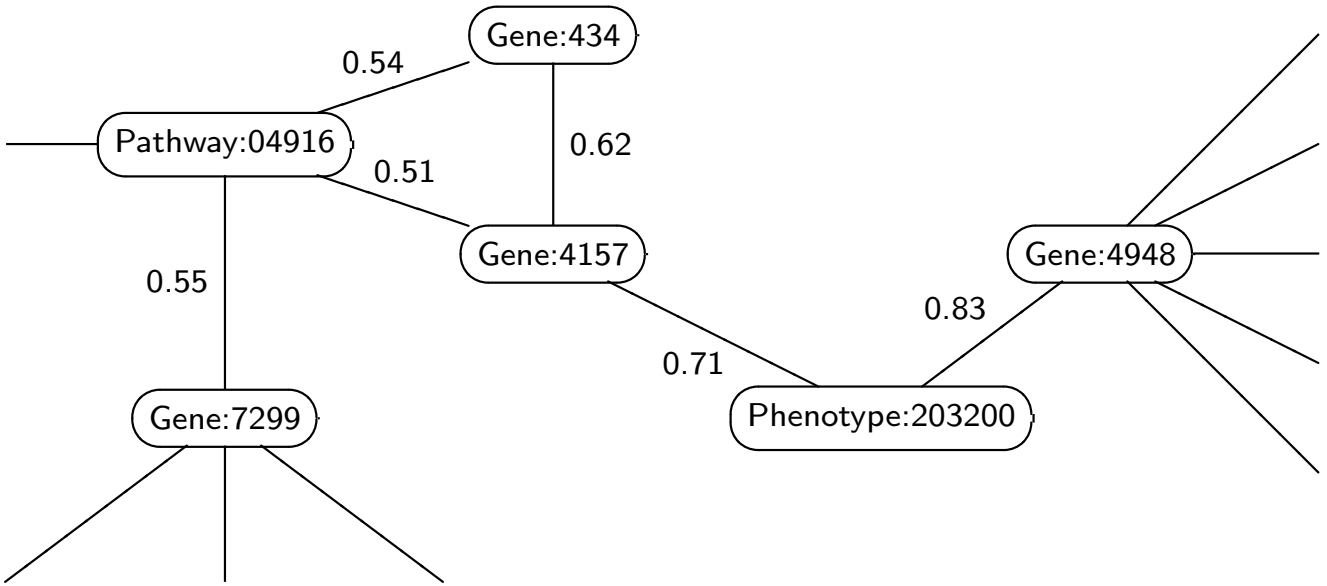
Workshop on Explorative Analytics of Information Networks
ECML PKDD 2009

Motivation

- Abstract a large set of nodes
- Remove redundancy
- Find groups of nodes

Biomine

- 12 biological databases integrated into a single large graph
- About 1 million nodes
- About 10 million edges



http://biomine.cs.helsinki.fi

Related work

- Clustering to find representatives
  - Clustering approximation

    (Yan et al., 2009; Kaufman and Rousseeuw, 1990; Ester et al., 1995)
  - Reducing the number of datapoints in databases

    (Riquelme et al., 2003; Rozsypal and Kubat, 2003; Pan et al., 2005)
- Representatives in graphs
  - Reducing the number of receivers sending feedback

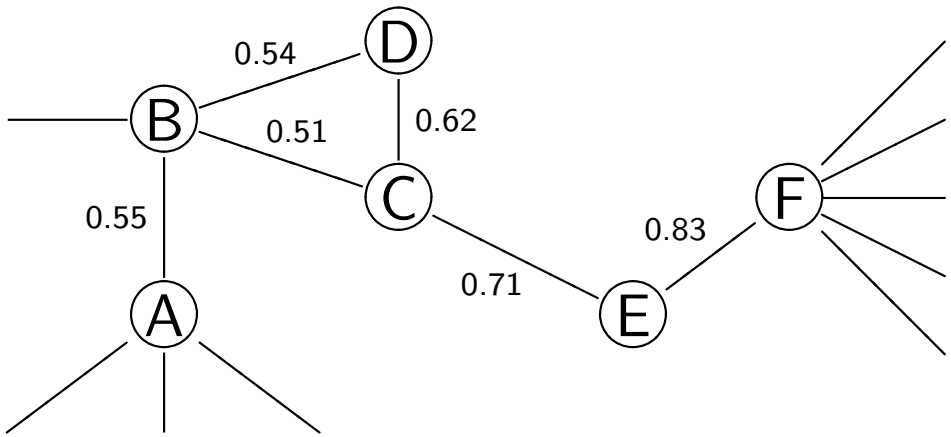
    (DeLucia and Obraczka, 1997; Liang et al., 2001)
- Searching for special node(s), but not representatives
  - Viral marketing (Domingos and Richardson, 2001)
  - Center-piece subgraphs (Tong and Faloutsos, 2006)
  - PageRank (Page et al., 1999)

Clustering and representatives in graphs

- Given a set S of nodes
- Calculate pairwise similarities in S
- Cluster nodes
  - k-medoids
  - Hierarchical clustering
- Output the most central node for each cluster

Similarities in probabilistic graphs

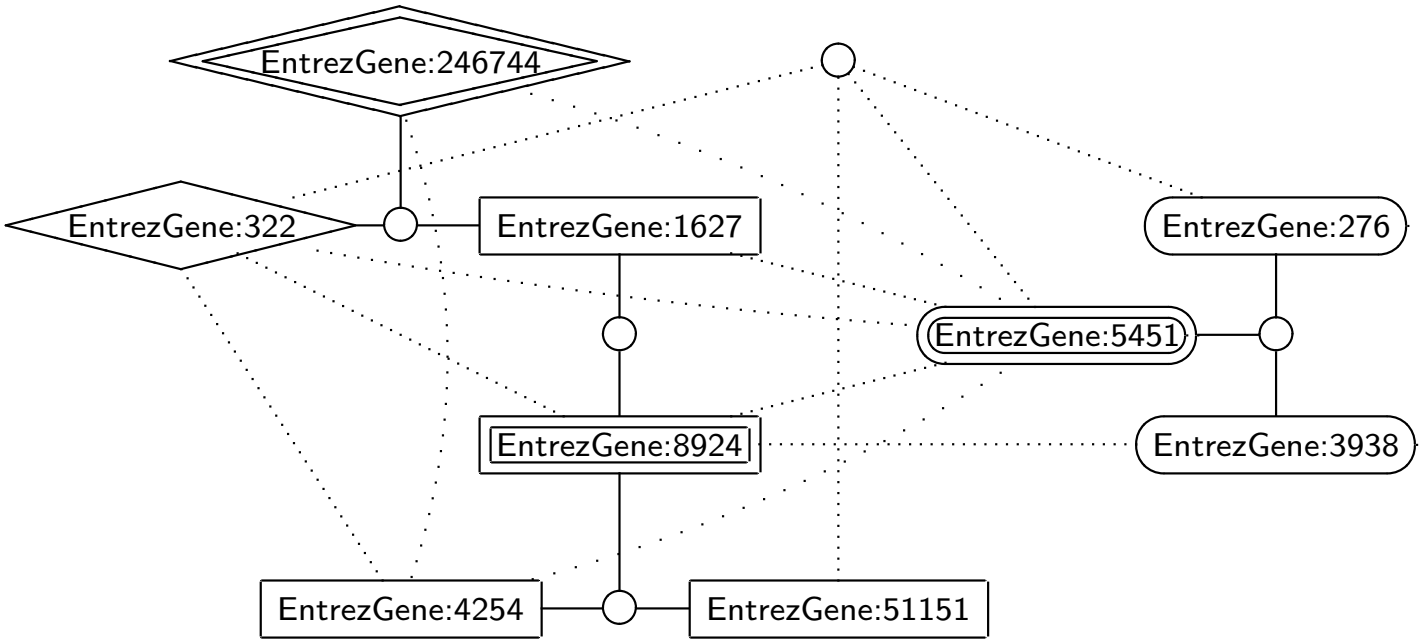
- Probabilistic (a.k.a. Bernoulli random) graph $G = (V, E)$
- Probability $p(e)$ is the edge weight of edge $e \in E$
- Path P consists of edges $e_1, \ldots, e_k$
- Probability of a path: $p(P) = p(e_1) \cdot \ldots \cdot p(e_k)$.
- Probability of the best path, given two nodes $u, v \in V$:
  $s(u, v) = \max_{P \text{ is a path from } u \text{ to } v} p(P)$



k-medoids

1. Choose k nodes randomly from S as initial medoids
2. Cluster remaining nodes in S to the most similar medoid;
   If similarities to all medoids are zero $\longrightarrow$ outlier
3. Find a new medoid for each cluster
4. Repeat steps 2. and 3. until convergence

Example


EntrezGene:246744
EntrezGene:322
EntrezGene:1627
EntrezGene:276
EntrezGene:5451
EntrezGene:8924
EntrezGene:3938
EntrezGene:4254
EntrezGene:51151

Hierarchical clustering

1. Each node in a cluster of its own
2. Merge those two clusters that give the best merged cluster
3. Repeat 2. step until exactly k clusters remain
4. Find a medoid for each cluster

Test data

- Data published by Köhler et al. (2008)
- 110 disease-gene families based on OMIM
- Each family related to one disease
- Three to 41 genes per family

Test setting

- k-medoids vs. hierarchical clustering vs. random selection of representatives
- $k = 3$ and $k = 10$
- 100 test runs

Measures

- Measures of representativeness

Measures

- Measures of representativeness
  - Average similarity of objects to their closest representative: $ASR = \frac{1}{|S|-K} \sum_{x \in S, x \neq m(x)} s(x, m(x))$

Measures

- Measures of representativeness
  - Average similarity of objects to their closest representative:
    $ASR = \frac{1}{|S|-K} \sum_{x \in S, x \neq m(x)} s(x, m(x))$
  - Fraction of non-represented classes:
    $NRC = \frac{1}{K} |\{k \mid \nexists j : m_j \in H_k,\ j = 1..K\}|$

Measures

- Measures of representativeness
  - Average similarity of objects to their closest representative:
    $ASR = \frac{1}{|S|-K} \sum_{x \in S, x \neq m(x)} s(x, m(x))$
  - Fraction of non-represented classes:
    $NRC = \frac{1}{K} |\{k \mid \nexists j : m_j \in H_k,\ j = 1..K\}|$
- Measures of underlying clustering

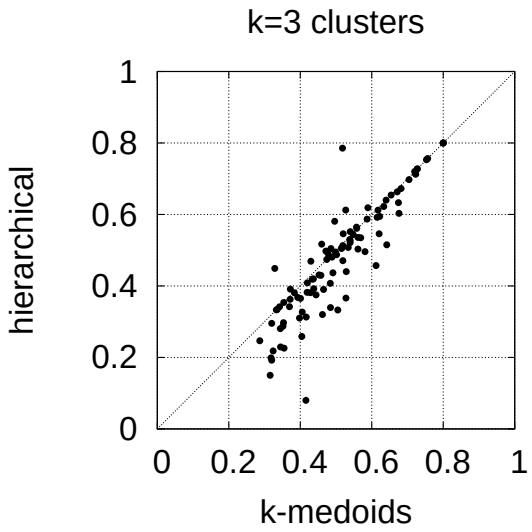
Measures

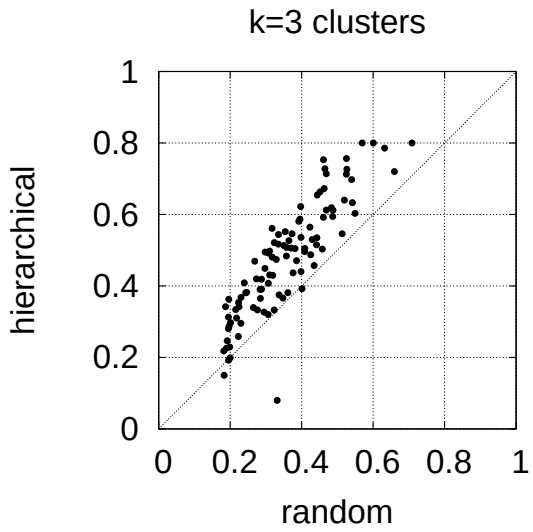
- Measures of representativeness
  - Average similarity of objects to their closest representative:
    $ASR = \frac{1}{|S|-K} \sum_{x \in S, x \neq m(x)} s(x, m(x))$
  - Fraction of non-represented classes:
    $NRC = \frac{1}{K} |\{k \mid \nexists j : m_j \in H_k,\ j = 1..K\}|$
- Measures of underlying clustering
  - Average Compactness of Clusters:
    $ACC = \frac{1}{k'} \sum_{k=1}^{K} \min_{x,y \in C_k} s(x, y)$

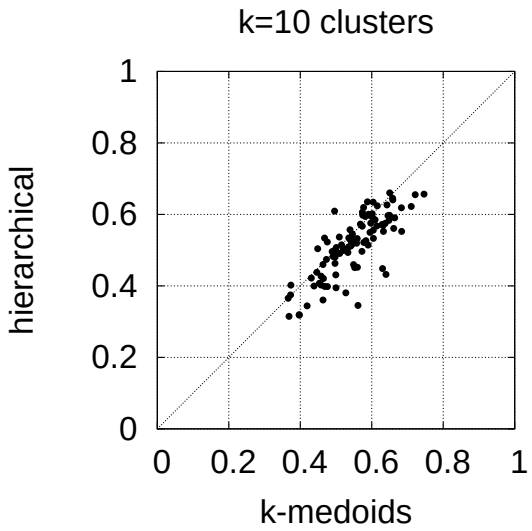
Measures

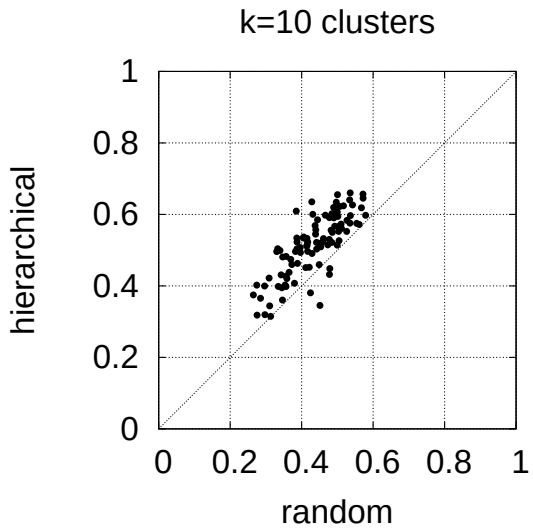
- Measures of representativeness
  - Average similarity of objects to their closest representative:
    $ASR = \frac{1}{|S|-K} \sum_{x \in S, x \neq m(x)} s(x, m(x))$
  - Fraction of non-represented classes:
    $NRC = \frac{1}{K} |\{k \mid \nexists j : m_j \in H_k,\ j = 1..K\}|$
- Measures of underlying clustering
  - Average Compactness of Clusters:
    $ACC = \frac{1}{k'} \sum_{k=1}^{K} \min_{x,y \in C_k} s(x, y)$
  - "Wrongly assigned" objects:
    $WAO = \frac{1}{|S|} \sum_{k=1}^{K} \min_{k'=1..K} |C_k \setminus H_{k'}|$

Average similarity to nearest representative (ASR)





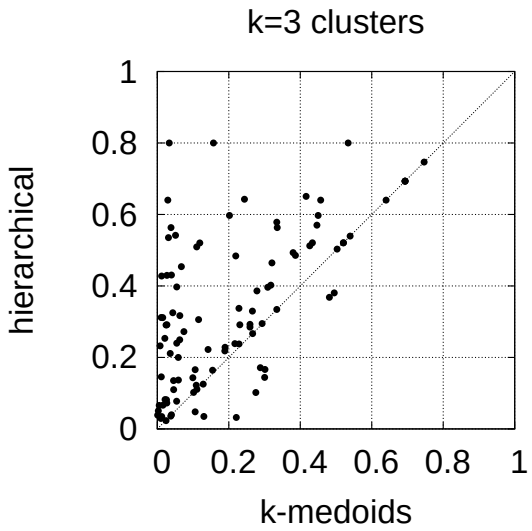


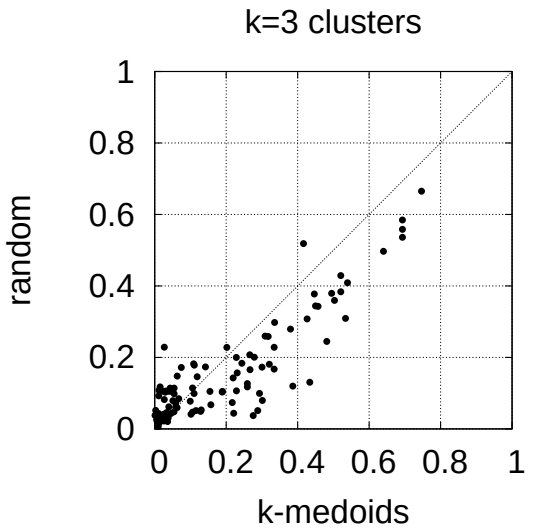


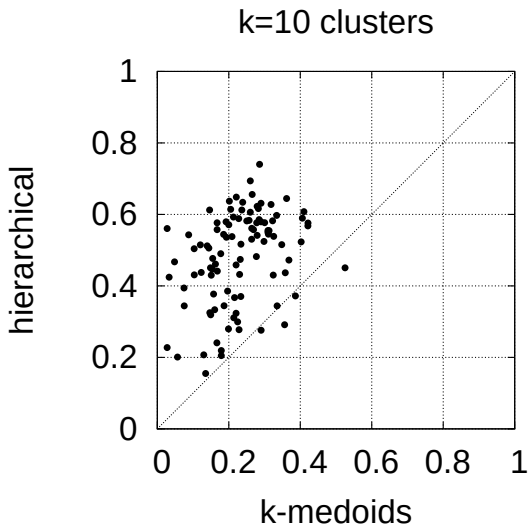
Fraction of non-represented classes (NRC)

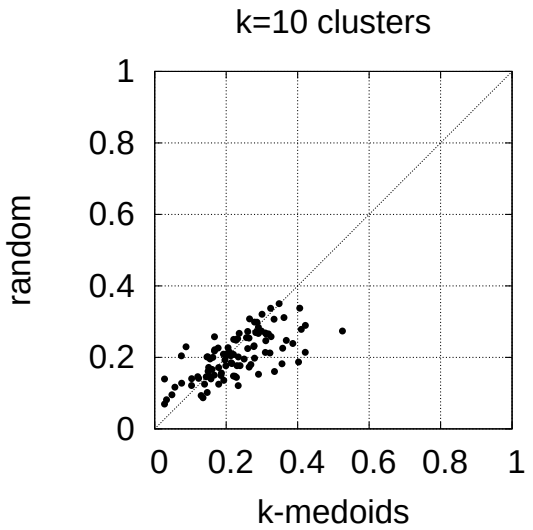
	k=3	k=10
k-medoids	14 %	29 %
hierarchical	16 %	21 %
random	34 %	39 %

Average compactness of clusters (ACC)









"Wrongly assigned" objects (WAO)

	k=3	k=10
k-medoids	18 %	44 %
hierarchical	15 %	25 %
random	27 %	46 %

Conclusion

- Probabilistic graph
- Identifying representatives
- k-medoids and hierarchical clustering are promising approaches to identify representatives
- Hierarchical method is more robust