Finding representatives in a heterogeneous network

Laura Langohr

Department of Computer Science
University of Helsinki

May 19, 2009

Introduction

K-medoids

Experiments

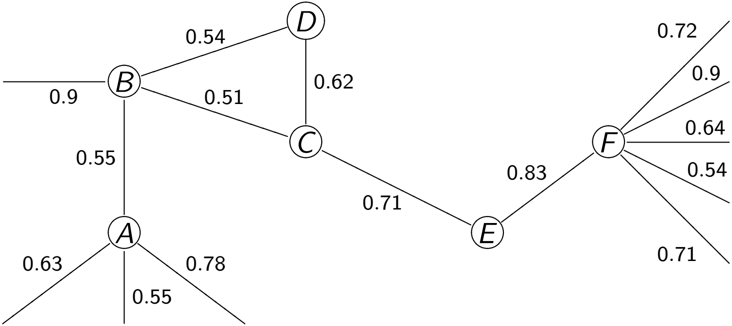
Future Work

Conclusion

Motivation

- **Finding representative vertices**
- Given a list of 100 vertices
- But only resources to study 10 vertices
- Cluster 100 vertices in 10 clusters
- For each cluster suggest a vertex as representative

Example graph

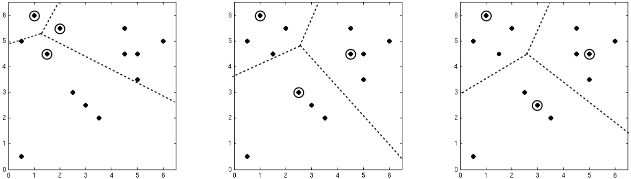


K-medoids

- Clustering method
- Objects are partitioned into k clusters
- First, an initial partitioning is created
- The partition is then iteratively improved
- Cluster centers are objects → medoids

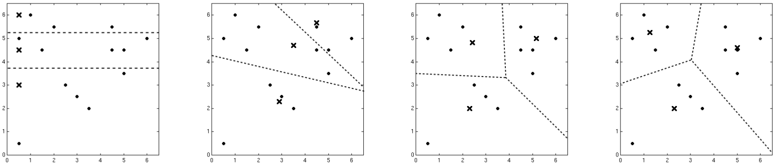
Algorithm

1. *K* objects are randomly chosen as medoids
2. Assign remaining objects to the medoid that is the nearest
3. Calculate new medoid for each cluster

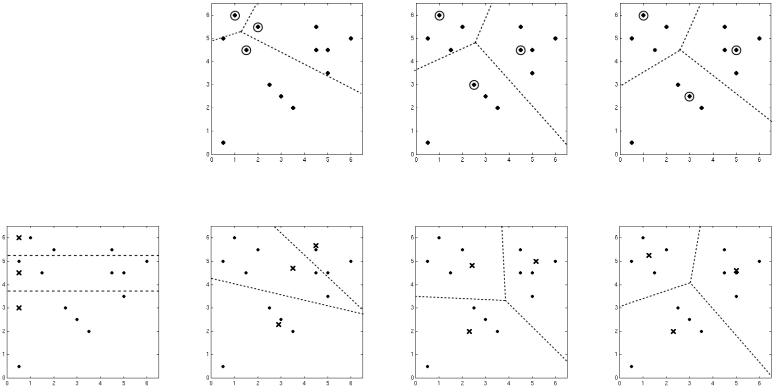


K-means

- *K*-medoids is similar to *k*-means
- *K*-means uses mean value as cluster center



K-medoids vs *k*-means



K-medoids in a heterogeneous network

- Select few representatives from a large set of vertices
- Representatives should be independent of each other
- Relations between two vertices in a graph → link
- Including undiscovered relations
- Undiscovered relations are manifested as path(s)

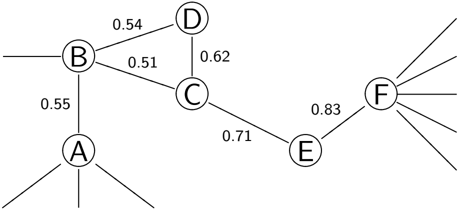
Measure for link strength

- Probability of a path is the product of the probabilities of the edges along the path

  $g(\mathbf{p}) = \prod_{i=1}^{k} w(e_i)$

- Probability of the best path between two vertices

  $$P_{bp} = \max_{p \in Pa(G,o,o')} g(\mathbf{p})$$



Algorithm

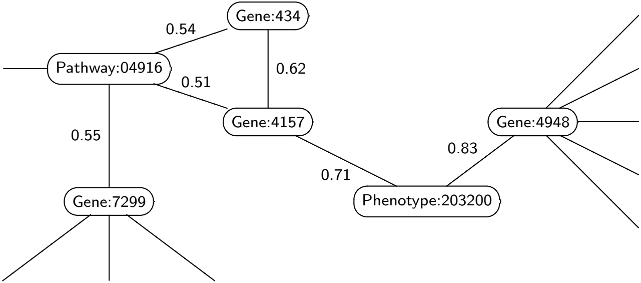
1. Calculate similarity matrix
2. Choose k objects randomly as initial medoids
3. Assign each remaining object to the most similar medoid
4. Calculate new medoid for each cluster

$$medoid(C_j) = \operatorname*{argmax}_{o \in C_j} \prod_{\substack{o' \in C_j \\ o' \neq o}} P_{bp}(G, o, o')$$

Repeat steps 3. and 4. until clustering converges

Biomine

- 12 biological databases are integrated
- Over 1 million vertices
- Over 9 million edges

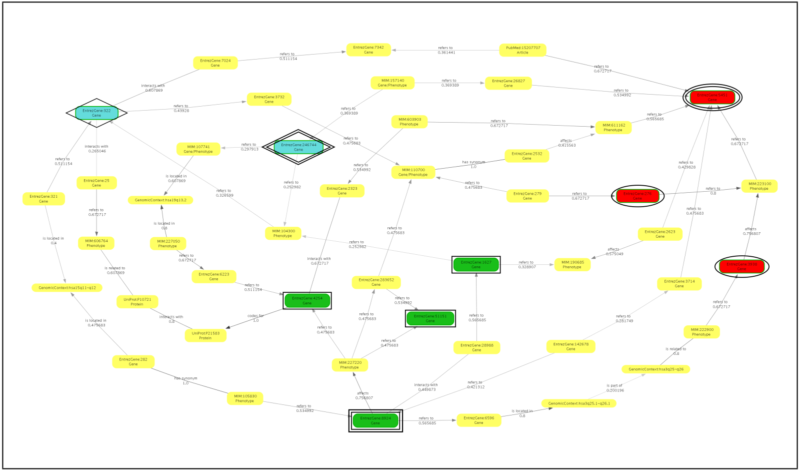


http://biomine.cs.helsinki.fi

Artificial example

- Three phenotypes, for each three genes
- k-medoids with nine genes, and $k = 3$

Result


EntrezGene:7342
Gene
refers to
0.361441
PubMed:15207707
Article
refers to
0.511154
EntrezGene:7024
Gene
refers to
0.672717
MIM:157140
Gene/Phenotype
refers to
0.369389
EntrezGene:26827
Gene
interacts with
0.607869
refers to
0.534992
EntrezGene:5451
Gene
EntrezGene:3732
Gene
refers to
0.43928
refers to
0.369389
EntrezGene:322
Gene
refers to
0.565685
MIM:603903
Phenotype
refers to
0.672717
MIM:611162
Phenotype
refers to
0.475683
refers to
0.672717
affects
0.415563
MIM:107741
Gene/Phenotype
refers to
0.297913
EntrezGene:246744
Gene
interacts with
0.265046
EntrezGene:2532
Gene
refers to
0.511154
refers to
0.429828
has synonym
1.0
refers to
0.534992
MIM:110700
Gene/Phenotype
is located in
0.607869
EntrezGene:25
Gene
refers to
0.252982
refers to
0.475683
MIM:223100
Phenotype
EntrezGene:2323
Gene
refers to
0.8
refers to
0.326599
EntrezGene:279
Gene
refers to
0.672717
EntrezGene:276
Gene
EntrezGene:321
Gene
GenomicContext:hsa19q13.2
refers to
0.475683
refers to
0.672717
is located in
0.8
refers to
0.475683
affects
0.756807
MIM:104300
Phenotype
EntrezGene:2623
Gene
is located in
0.4
MIM:606764
Phenotype
MIM:227050
Phenotype
refers to
0.252982
affects
0.575049
refers to
0.672717
interacts with
0.672717
EntrezGene:1627
Gene
refers to
0.328907
MIM:190685
Phenotype
EntrezGene:3938
Gene
is related to
0.607869
EntrezGene:6223
Gene
EntrezGene:283652
Gene
EntrezGene:3714
Gene
refers to
0.511154
GenomicContext:hsa15q11-q12
refers to
0.534992
EntrezGene:4254
Gene
UniProt:P10721
Protein
refers to
0.672717
refers to
0.475683
codes for
1.0
EntrezGene:51151
Gene
refers to
0.565685
refers to
0.281749
interacts with
0.8
is located in
0.475683
refers to
0.475683
MIM:222900
Phenotype
UniProt:P21583
Protein
refers to
0.475683
EntrezGene:28988
Gene
EntrezGene:142678
Gene
is related to
0.8
EntrezGene:282
Gene
MIM:227220
Phenotype
GenomicContext:hsa3q25-q26
has synonym
1.0
interacts with
0.449873
refers to
0.421312
is part of
0.200196
affects
0.756807
MIM:105830
Phenotype
GenomicContext:hsa3q25.1-q26.1
refers to
0.534992
is located in
0.8
EntrezGene:8924
Gene
refers to
0.565685
EntrezGene:6596
Gene

Future Work

- Hierarchical clustering
- Statistical evaluation
- Comparison to an existing method

Conclusion

- Finding representative vertices, e.g. genes
- *K*-medoids on Biomine
- Example with nine genes is promising