FIELDS 2019 Summer School
Modelling Complex Networks

François Théberge
theberge@ieee.org

CSE/TIMC

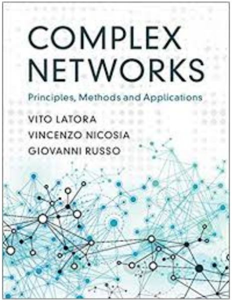
August 2019

Roadmap

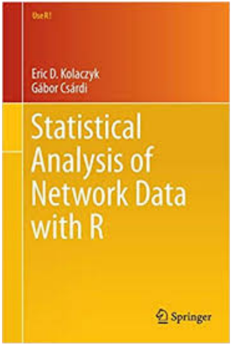
1. Relational data mining
    - measures of centrality
    - graph models
    - benchmarks
2. Community structure
    - comparing graph partitions
    - graph clustering algorithms
    - ensemble clustering on graphs (ECG)
3. Graph embedding
4. Semi-supervised learning on graphs
5. Hypergraph modularity and clustering

Some references

A few references for the background material:



Albert-László Barabási
NETWORK SCIENCE



Notebooks

I will illustrate the lectures with **Jupyter Notebooks** trough through **anaconda.com.**

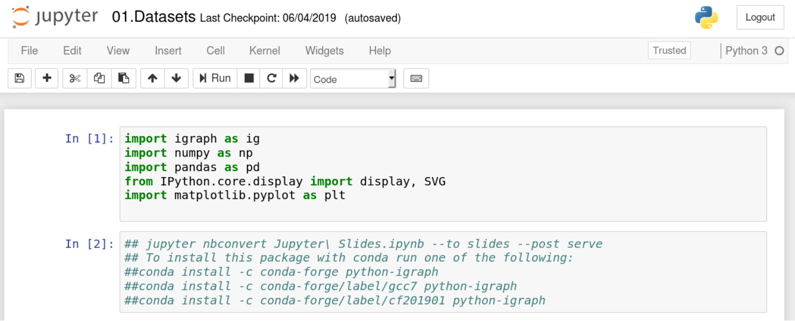
Code is in **Python 3**, using the **igraph** package (*igraph.org/python*)**.**

igraph also available in **R**: *www.r-project.org/*

Other useful software:

- *networkx* Python package
- *Gephi* to visualize larger graphs

Notebooks

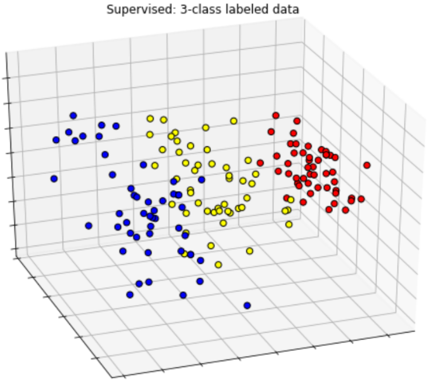

jupyter 01.Datasets Last Checkpoint: 06/04/2019 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code

In [1]:
```
import igraph as ig
import numpy as np
import pandas as pd
from IPython.core.display import display, SVG
import matplotlib.pyplot as plt
```

In [2]:
```
## jupyter nbconvert Jupyter\ Slides.ipynb --to slides --post serve
## To install this package with conda run one of the following:
##conda install -c conda-forge python-igraph
##conda install -c conda-forge/label/gcc7 python-igraph
##conda install -c conda-forge/label/cf201901 python-igraph
```

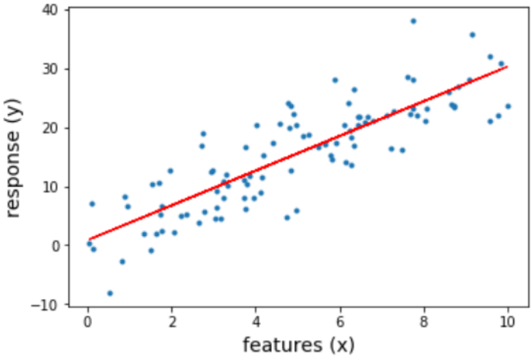
Machine Learning Terminology

- Supervised learning:
    - infer a function from labeled training data
    - ex: classification



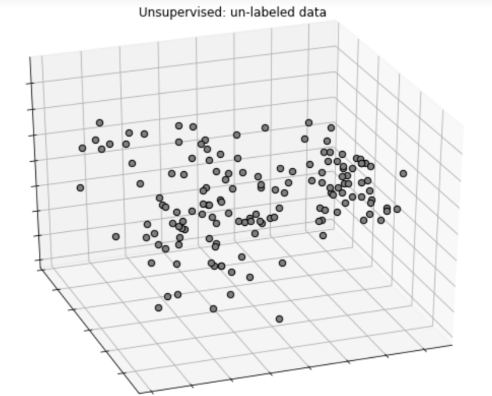
Machine Learning Terminology

- Supervised learning:
  - infer a function from labeled training data
  - ex: regression



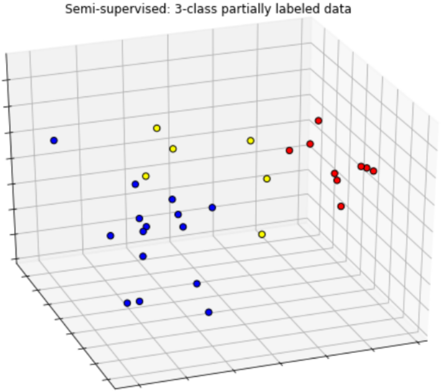
Machine Learning Terminology

- Unsupervised learning:
    - infer a function to describe structure in unlabeled data
    - ex: clustering, density estimation, dimension reduction, outlier detection.



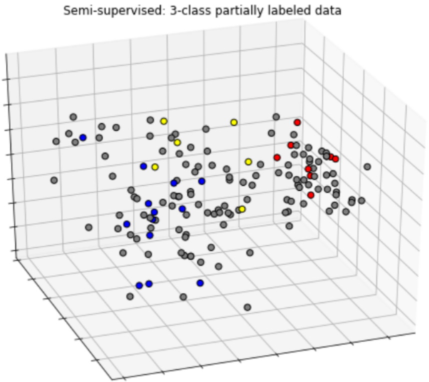
Machine Learning Terminology

- Semi-supervised learning:
  - typically labeled data is scarce



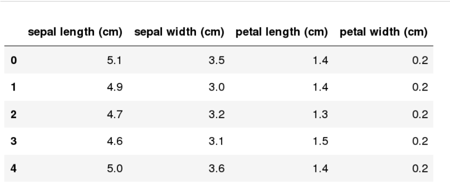
Machine Learning Terminology

- Semi-supervised learning:
  - so use both labeled and unlabeled data



Machine Learning Terminology

- previous examples assume data lives in a feature space
  - ex: vectors in $\mathbb{R}^n$
- categorical and ordinal data can also be represented
- ex: data frame in R or Python:

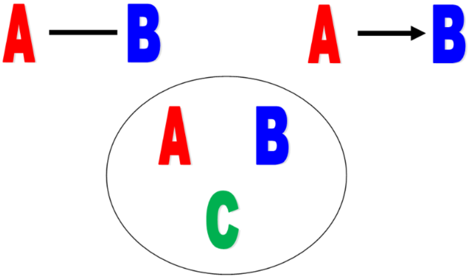


	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2

- each column is a *feature*

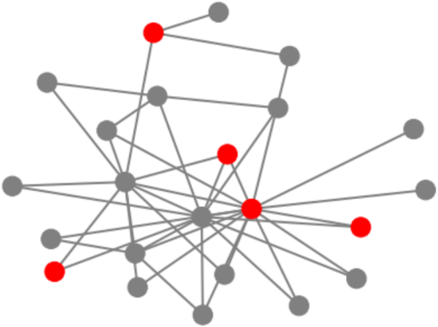
Relational Data

- not all data can be represented in a data frame
- data could be *relational*
- examples of relations between entities:
  - *A* and *B* are friends
  - *A* sends an email to *B*
  - *A*, *B* and *C* are in the same team
- the above are modelled as edges or hyperedges:



Relational Data

- relational data are often modelled as:
  - graphs: collections of edges, or
  - hypergraphs: collections of hyperedges
- entities are the nodes
- (hyper-)edges can be weighted or not, directed or not



Graphology

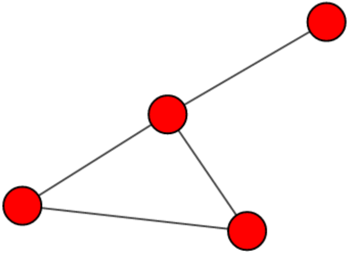
For a graph $G = (V, E)$, let $n = |V|$ and $m = |E|$

We map the vertices to integers $1 \ldots n$ for convenience

Let $A = (a_{ij})$, the adjacency matrix s.t. $a_{ij} > 0 \iff (i, j) \in E$

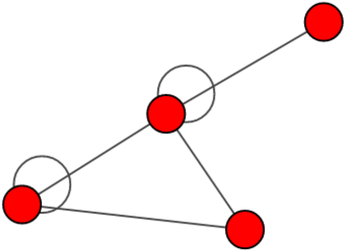
Graphology

Undirected graph: $a_{ij} = a_{ji} \in \{0, 1\}$, $a_{ii} = 0$.



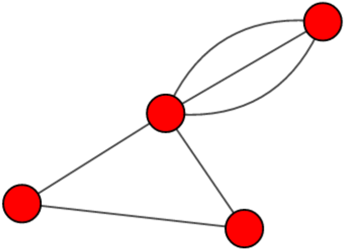
Graphology

Undirected graph with self-loops: some $a_{ii} = 1$.



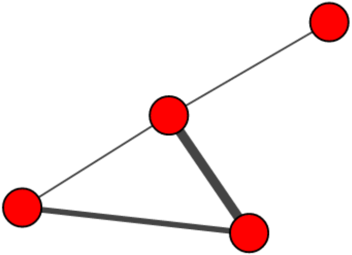
Graphology

Multigraph: $a_{ij} \in \mathbb{N}$



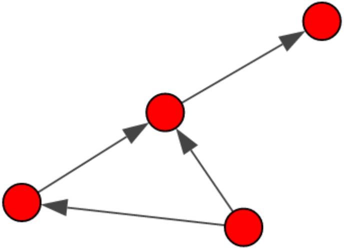
Graphology

Weighted graph: $a_{ij} \geq 0$



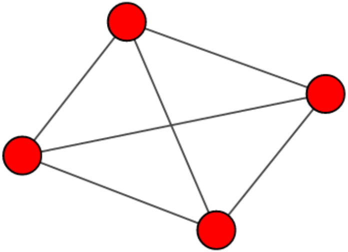
Graphology

Directed graph: can have $a_{ij} \neq a_{ji}$



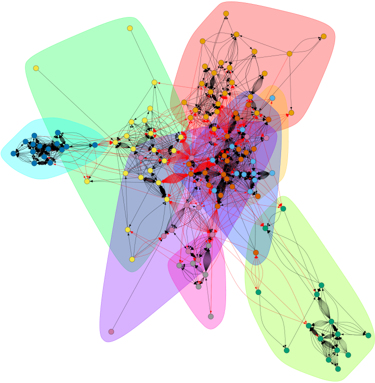
Graphology

Complete graph: all $a_{ij} = 1, \; i \neq j$ (a.k.a. clique).



Relational Data

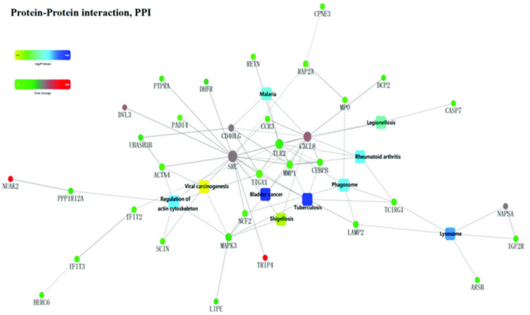
Relational data occurs in widely different contexts such as email exchanges (Enron email graph):



REF: https://lucidmanager.org/analyse-enron-corpus/

Relational Data

biology (protein-protein interaction graph):

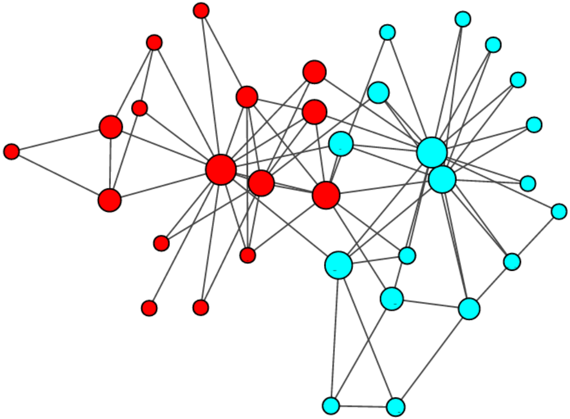


Protein-protein interaction graph.

REF: Yang, Minghui et. al., Inter. J. of Molecular Sci. 19. 2406. 10.3390/ijms19082406.

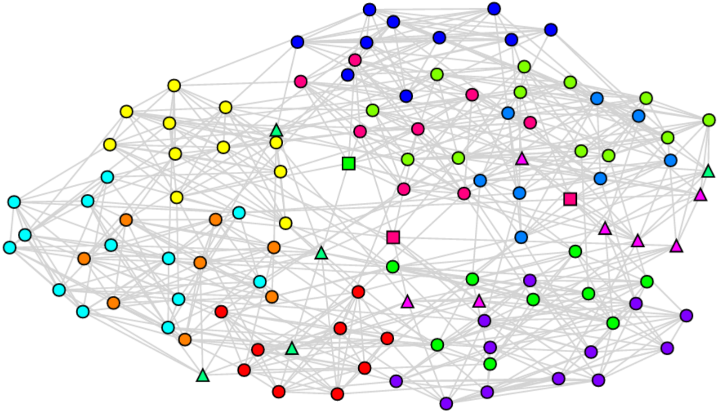
Relational Data

social ties (Zachary Karate Club graph):



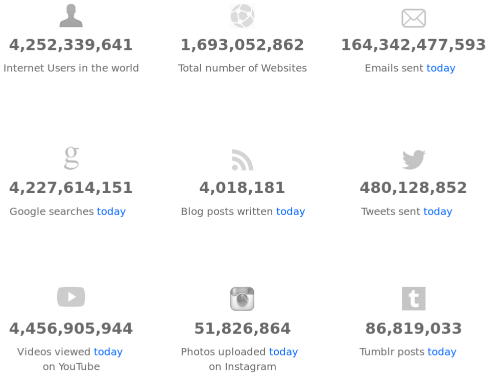
Relational Data

events (games between college football teams):



Relational Data

And there can be a lot of data to consider:



4,252,339,641
Internet Users in the world

1,693,052,862
Total number of Websites

164,342,477,593
Emails sent today

4,227,614,151
Google searches today

4,018,181
Blog posts written today

480,128,852
Tweets sent today

4,456,905,944
Videos viewed today
on YouTube

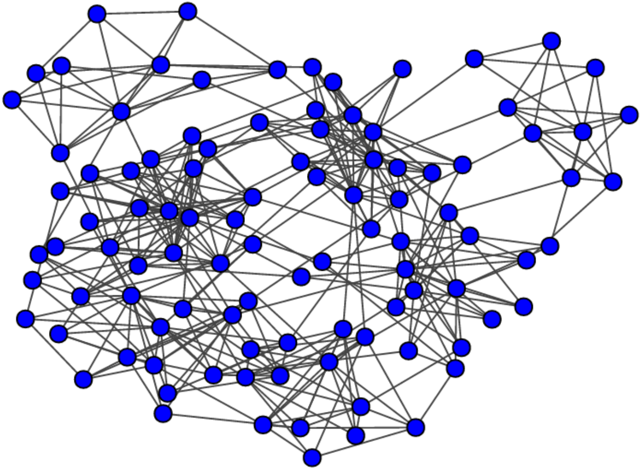
51,826,864
Photos uploaded today
on Instagram

86,819,033
Tumblr posts today

REF: snapshot from livestats.com

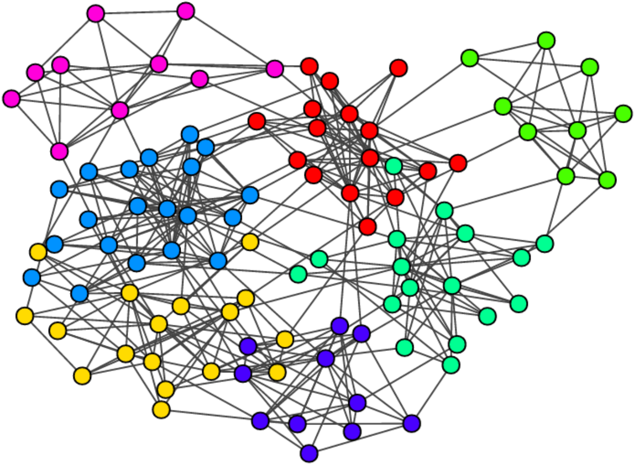
Issues with Relational Data

Consider some graph:



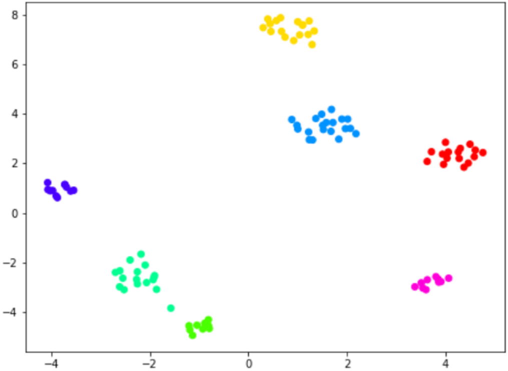
Issues with Relational Data

With some communities:



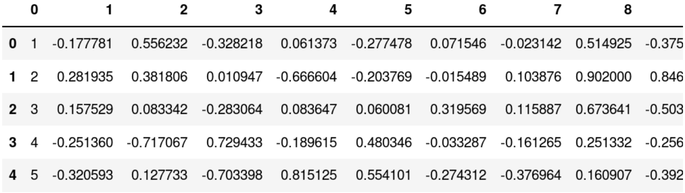
Issues with Relational Data

And an embedding with 2-dimensional view in vector space:



Issues with Relational Data

Working in vector space (data frames):



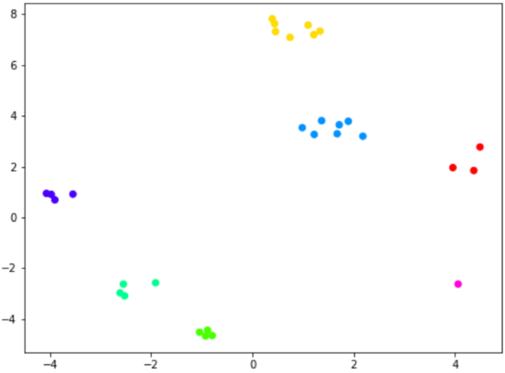
	0	1	2	3	4	5	6	7	8	
0	1	-0.177781	0.556232	-0.328218	0.061373	-0.277478	0.071546	-0.023142	0.514925	-0.375
1	2	0.281935	0.381806	0.010947	-0.666604	-0.203769	-0.015489	0.103876	0.902000	0.846
2	3	0.157529	0.083342	-0.283064	0.083647	0.060081	0.319569	0.115887	0.673641	-0.503
3	4	-0.251360	-0.717067	0.729433	-0.189615	0.480346	-0.033287	-0.161265	0.251332	-0.256
4	5	-0.320593	0.127733	-0.703398	0.815125	0.554101	-0.274312	-0.376964	0.160907	-0.392

Many tools exist to work over such data

Statistical techniques such as sampling can be used to handle large datasets

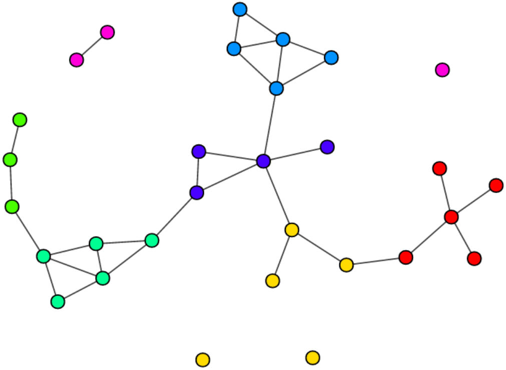
Issues with Relational Data

Sampling preserves key properties (clusters, average distances, etc):



Issues with Relational Data

But all of this is quickly destroyed in graph space:

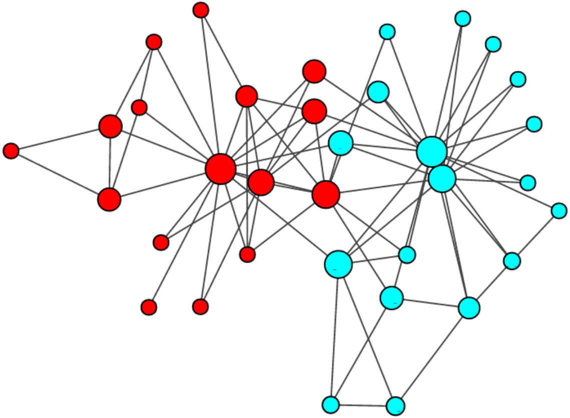


Relational Data

Problems in Graph Mining

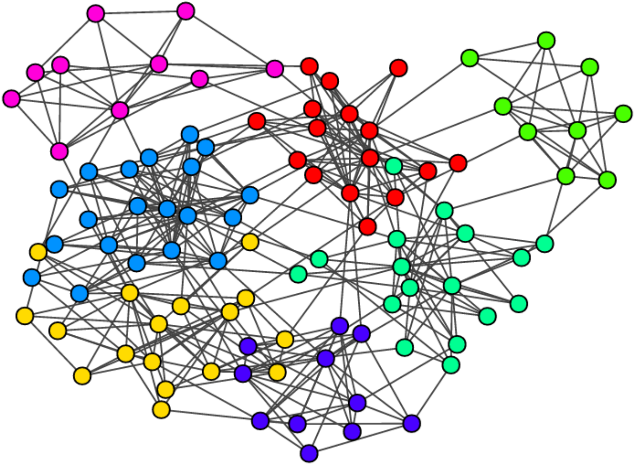
Problems in Graph Mining

Measures of centrality:



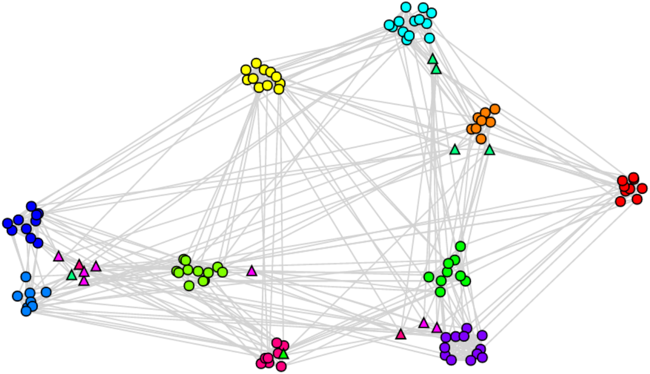
Problems in Graph Mining

Finding communities:



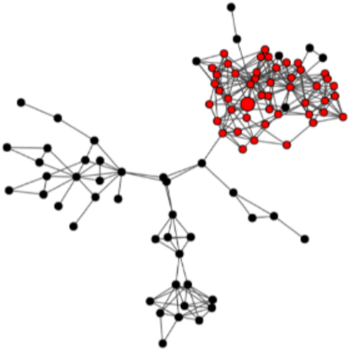
Problems in Graph Mining

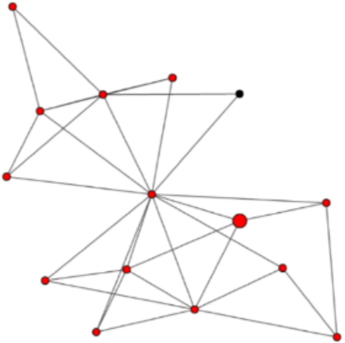
Anomaly detection:



Problems in Graph Mining

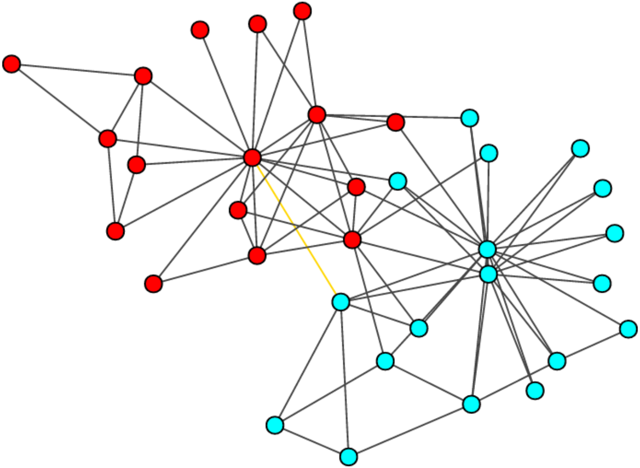
seed set expansion (local sampling)





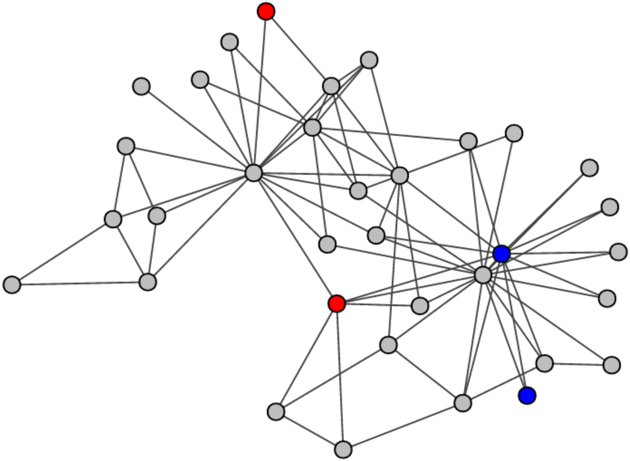
Problems in Graph Mining

link (edge) prediction:



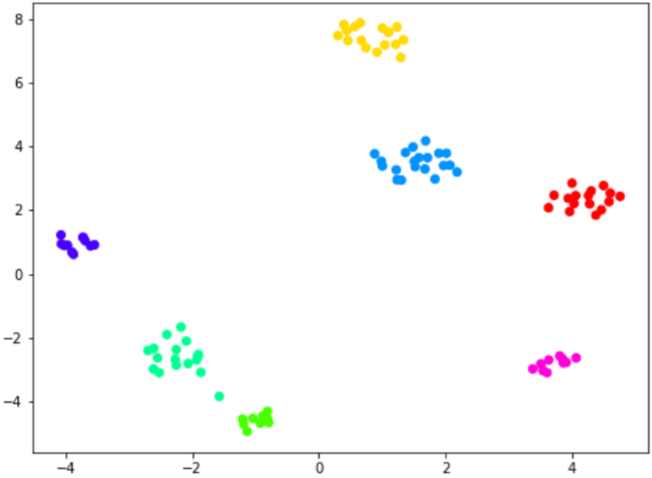
Problems in Graph Mining

semi-supervised learning:



Problems in Graph Mining

vector space embedding:


8
6
4
2
0
−2
−4
−4
−2
0
2
4