IBM Research Dublin
School of Computer Science
Carnegie Mellon University
Dept. of Computer Science
Rutgers

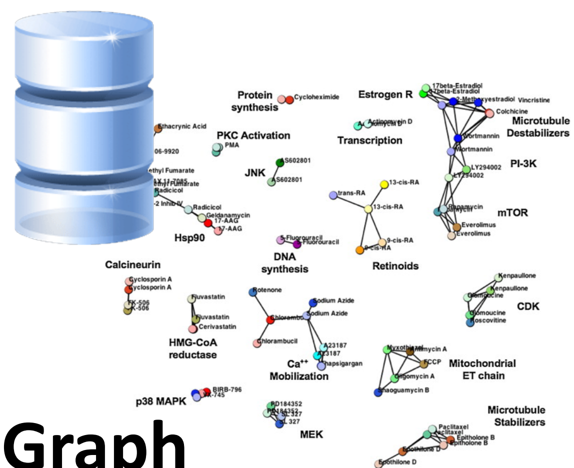
A scalable Approach to Size-independent Network Similarity

Michele Berlingerio
Tina Eliassi-Rad

Danai Koutra
Christos Faloutsos

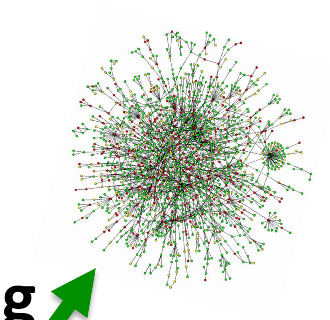
WIN, September 28th-29th 2012, NYU Stern School of Business

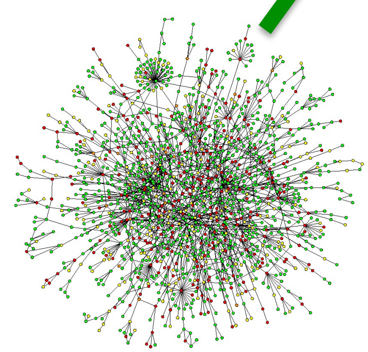
Why network similarity? (1)



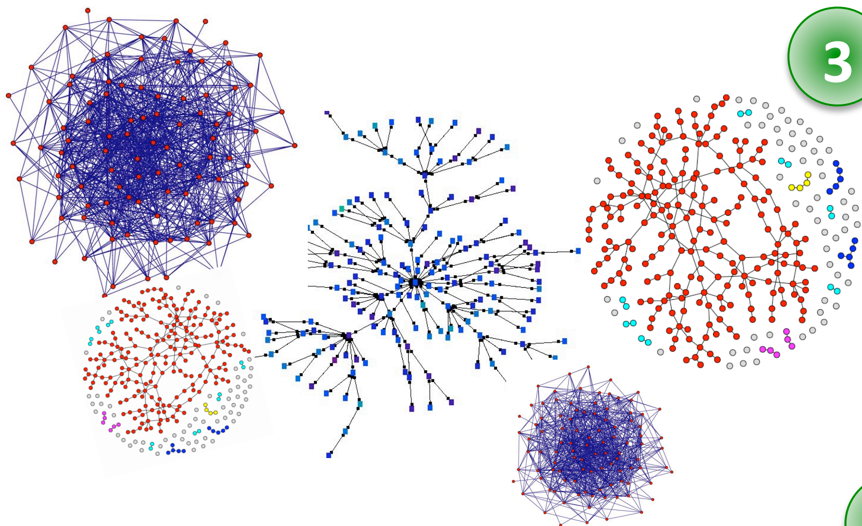
1 Graph Database: clustering

2 x-fer learning



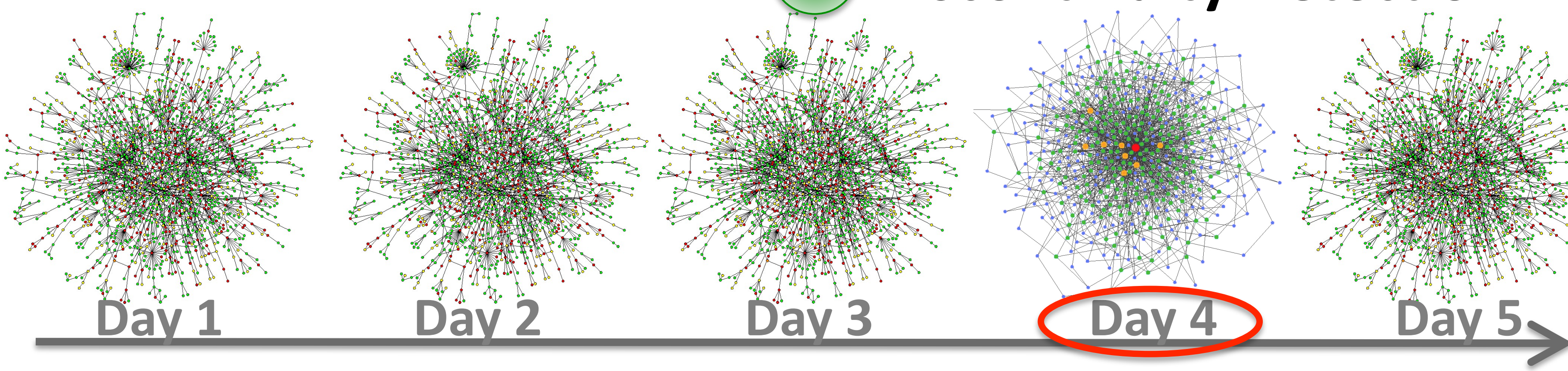


Why network similarity? (2)



3 **Anomaly detection: different models?**

4 **Discontinuity Detection**

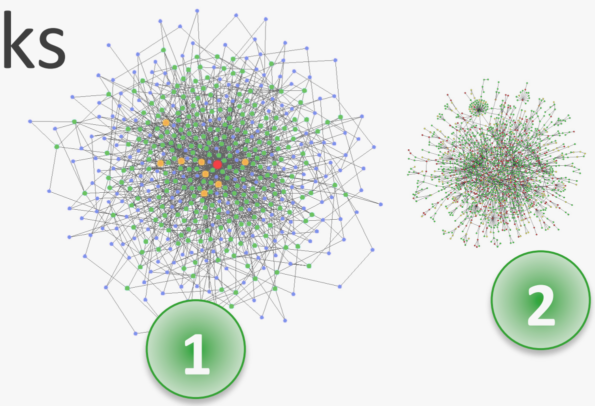


RoadMap

- **Problem Definition**
- *NetSimile*
- Experiments
  – Applications
- Conclusions

Network Similarity: Definition

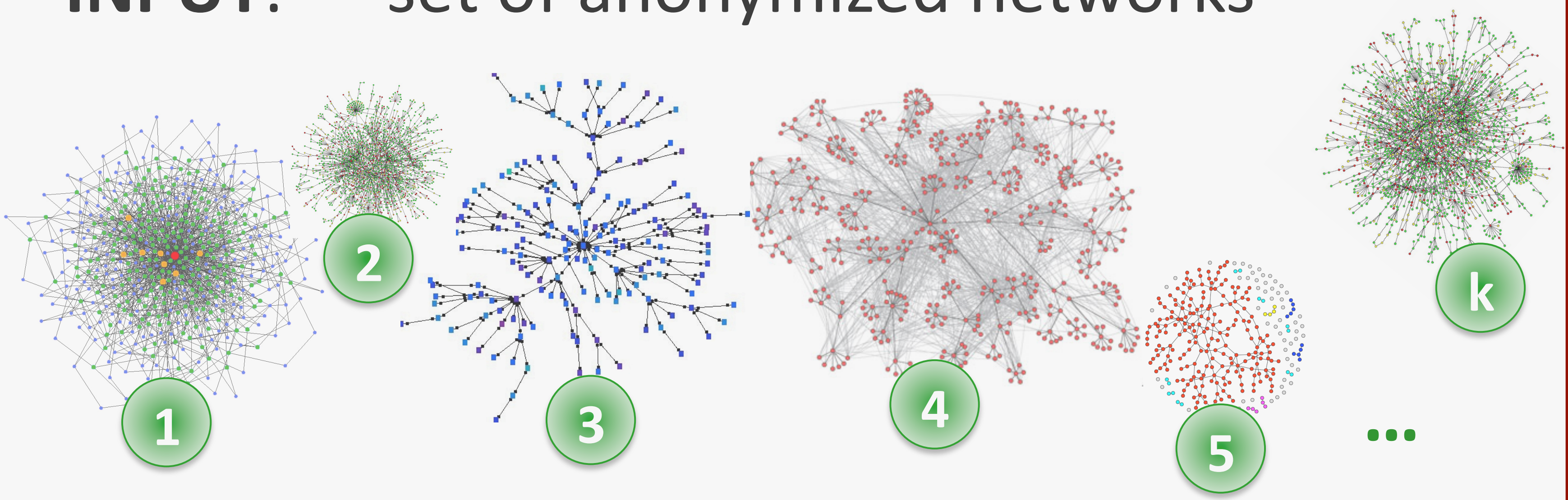
- **INPUT**: 2 anonymized networks
  - **GIVEN:** node IDs
  - **NOT GIVEN:** side-info
    class labels



- **OUTPUT**: structural similarity score

Network Similarity: Extension

- **INPUT**: set of anonymized networks



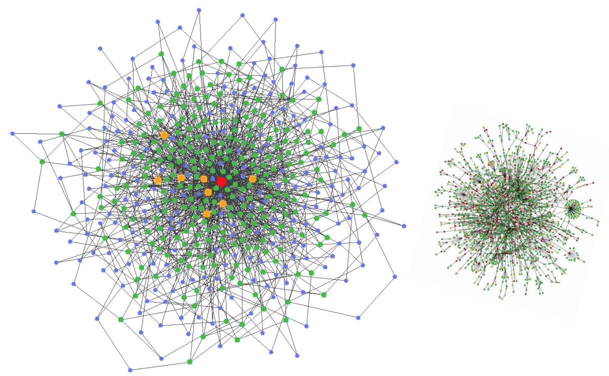
- **OUTPUT**: pairwise structural similarity scores

and…

$s_{12} = 0.8$
$s_{13} = 0.1$
...
$s_{5k} = 1$
...

Required Properties

- **P1. effectiveness**
  - size-independence
  - intuitiveness
  - interpretability
- **P2. scalability**



RoadMap

- Problem Definition
- ***NetSimile***
- Experiments
  - Applications
- Conclusions

NetSimile: overview

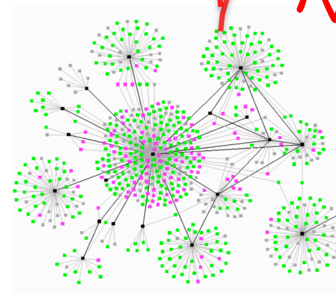
1 – Feature Extraction

2 – Feature Aggregation

3 - Comparison

deg, cc

E_{ego}, N_{ego}

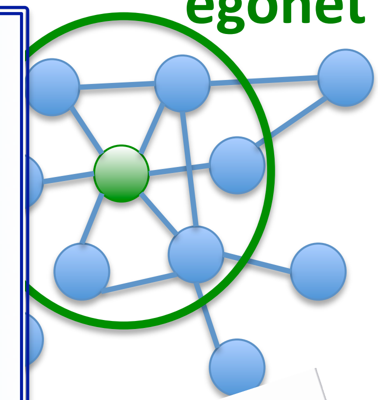


$$\frac{1}{n}\sum$$

$$\sqrt{\mathrm{E}[X^2]-\mathrm{E}[X]^2}$$

similarity metric

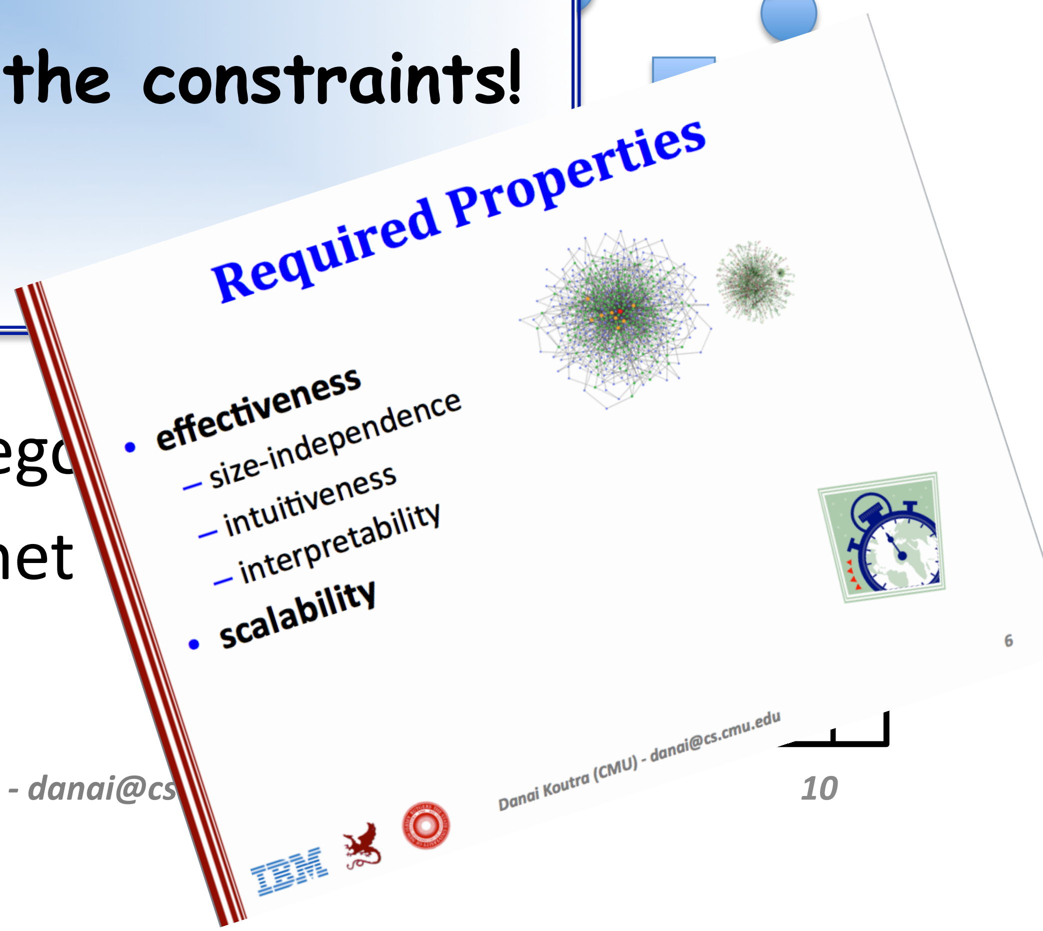
Step 1: Feature Extraction



- Local
  - ① #
  - ② cl
  - ③ a
  - ④ a
  - ⑤ edges in egonet
  - ⑥ outgoing edges from eg
  - ⑦ # of neighbors of egonet

Why these features?

They satisfy all the constraints!

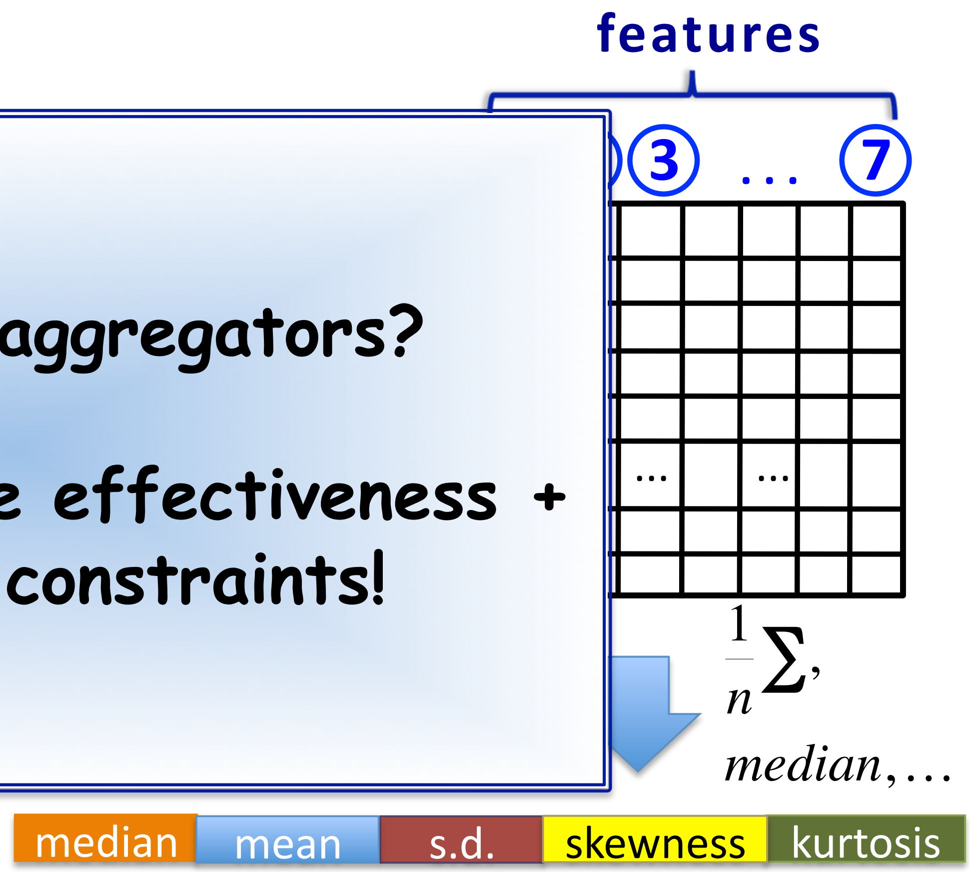


Step 2: Feature Aggregation

- 5 aggr
  - med
  - mea
  - stan
  - skev
  - kurt

Why these aggregators?

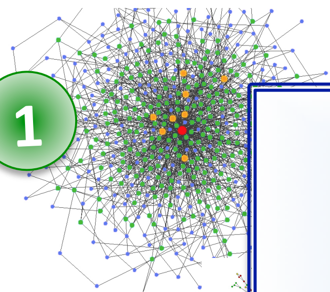
They satisfy the effectiveness + scalability constraints!



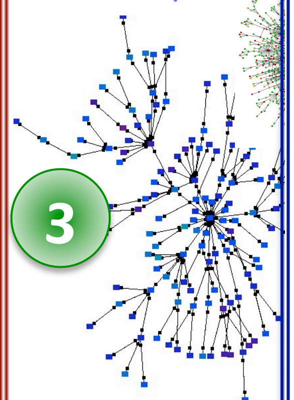
single 'signature' vector per network

Step 3: Comparison

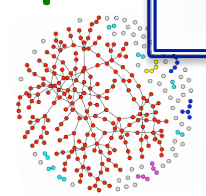
Networks



1



3



k

'Signature' Vectors
(aggr. features)

$\vec{s}_{Gk}$

Similarity Scores

12
13
1k
23
2k

Kolmogorov-Smirnov
...

$s_{k-1,k}$

Why Canberra distance?

$$canberra(\vec{P},\vec{Q}) = \sum_{i=1}^{d} \frac{|P_i - Q_i|}{|P_i| + |Q_i|}$$

① sensitive to small changes near 0
② normalizes the absolute difference of the individual comparisons.

Required Properties: NetSimile

☑ **P1. effectiveness**

- ☑ size-independence
- ☑ intuitiveness
- ☑ interpretability
- ☑ avoids the node-correspondence problem

☑ **P2. scalability**

LEMMA

The runtime complexity for generating NetSimile's 'signature' vectors is linear on the number of edges in the input networks:

nodes

$$O(\sum_{j=1}^{k} f \cdot n_j + f \cdot n_j \cdot \log(n_j))$$

RoadMap

- Problem Definition
- *NetSimile*
- **Experiments**
  - Applications
- Conclusions

Experiments: Data

- 30 real-world networks

Cornell University Library
arXiv.org

dblp

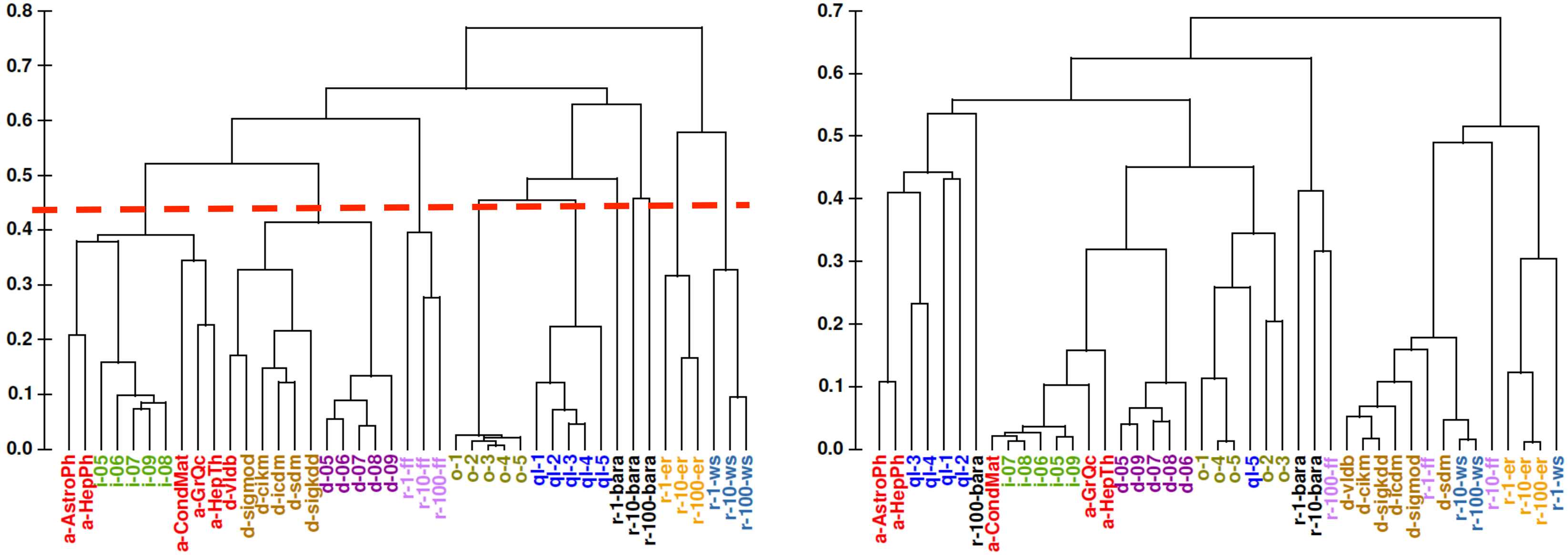
IMDb

querylog

Oregon AS

- multiple synthetic networks
  - Barabási-Albert
  - Forest Fire
  - Erdös-Rényi
  - Watts-Strogatz

Experiments: intuitiveness + interpretability of NetSimile



(a) NETSIMILE

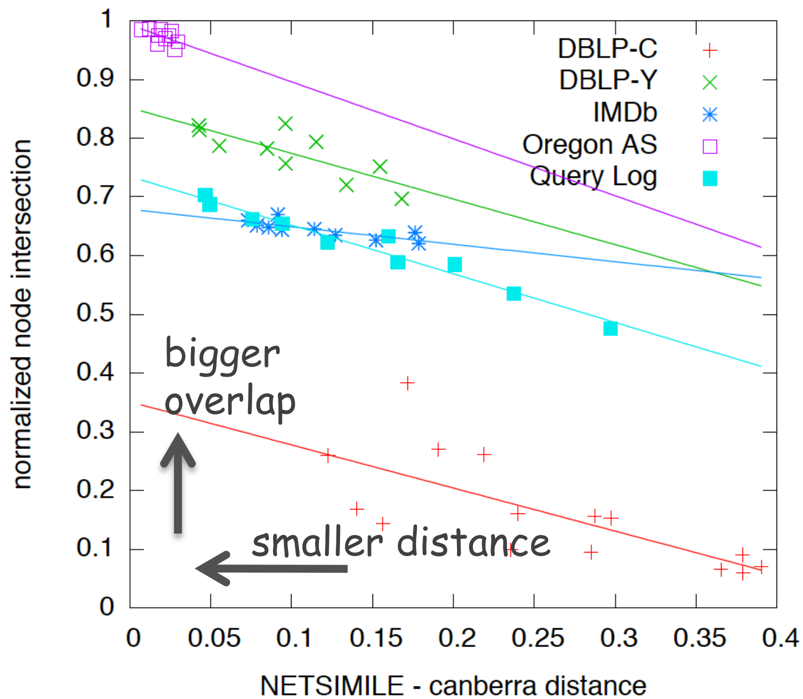
homogeneity in colors

(b) EIG

Observation:

NetSimile gives better and more intuitive graph clusters than the EIG method (eval-based competitor method).

Experiments: NetSimile and node-overlap

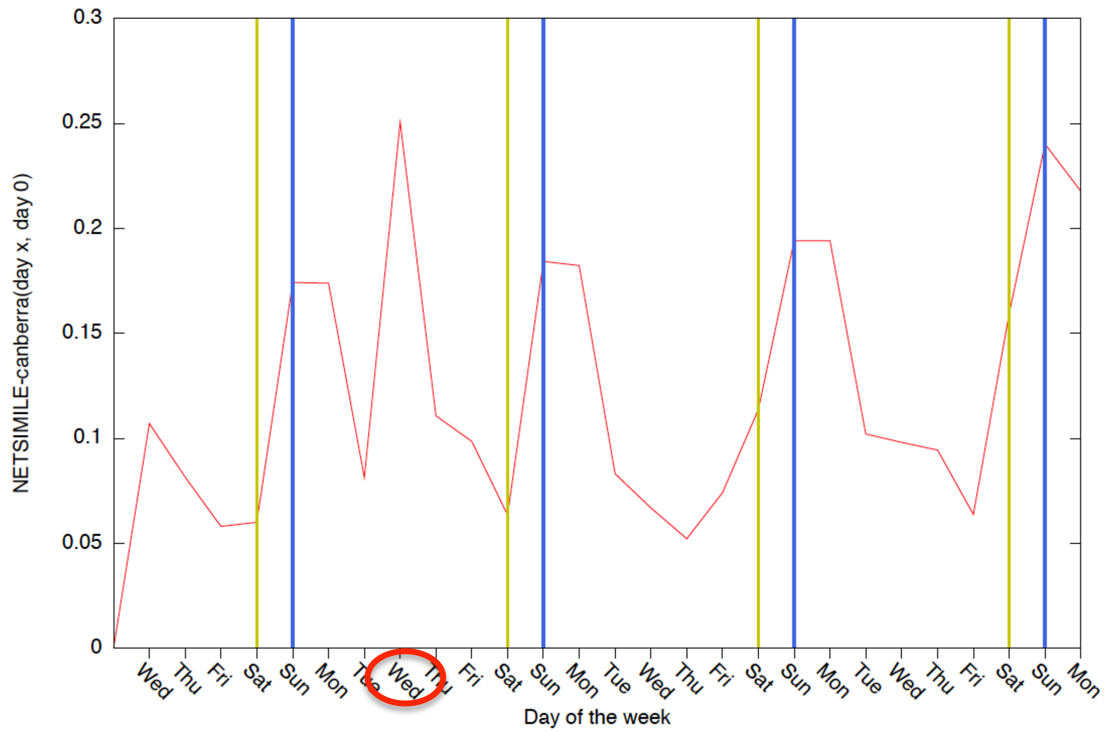


Hypothesis:
bigger node overlap =>
greater similarity

Implicit Assumption:
networks are from the same domain

Observation:
The lower the NetSimile score (greater similarity), the higher the normalized node intersection of the input networks.

Application: Discontinuity Detection in Yahoo! IM



(a) NETSIMILE between each day and day 0 in Yahoo! IM

1. Microsoft offers to buy Yahoo!.
2. New features for flickr were announced.

nodes: IM users
edges: communication events

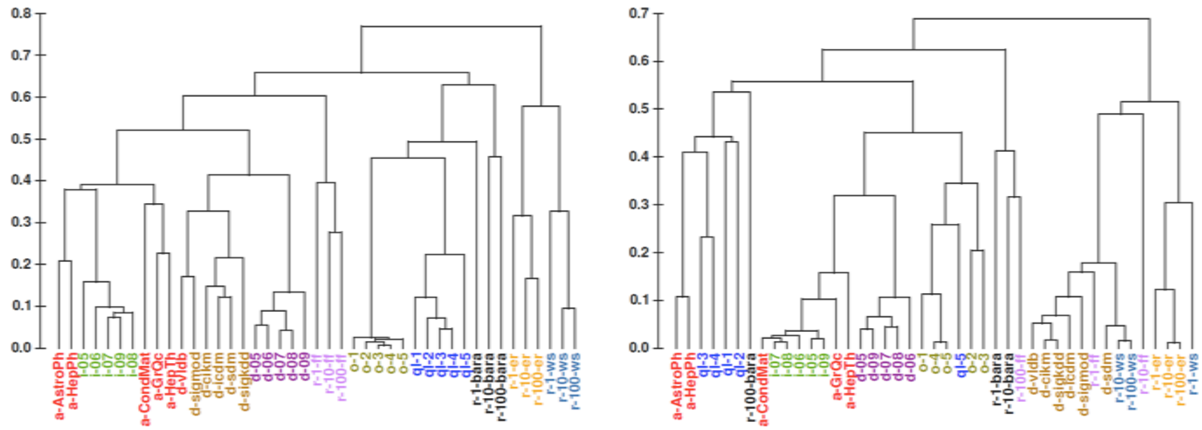
RoadMap

- Problem Definition
- *NetSimile*
- Experiments
  - Applications
- **Conclusions**

Conclusions

- **Novel approach**:
  - ‘signature’ vector for each graph (summarization)
- **NetSimile**:
  - effective
    - size-independent, intuitive, interpretable
  - scalable
- **Applicability** to a variety of problems

Thank you!

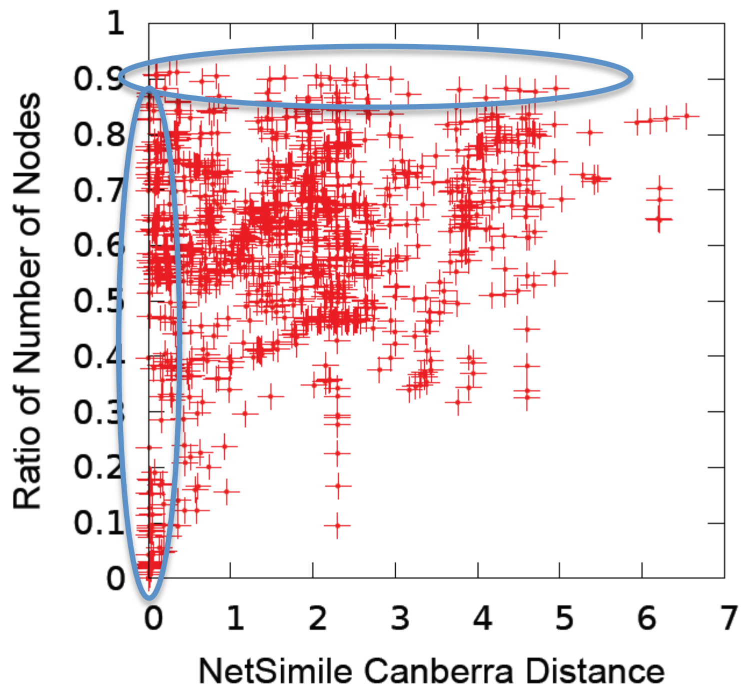


(a) NetSimile (b) EIG

http://www.cs.cmu.edu/~dkoutra

danai@cs.cmu.edu

Experiments (2): Are we measuring size?



Observation:
NetSimile is **not** measuring size – there is no correlation between extracted features and network size.

Advantages of NetSimile

☑ size-invariant

☑ scalable

LEMMA

The runtime complexity for generating NetSimile's 'signature' vectors is linear on the number of edges in the input networks:

nodes

$$O\left(\sum_{j=1}^{k} f \cdot n_j + f \cdot n_j \cdot \log(n_j)\right)$$

☑ avoids the node-correspondence problem