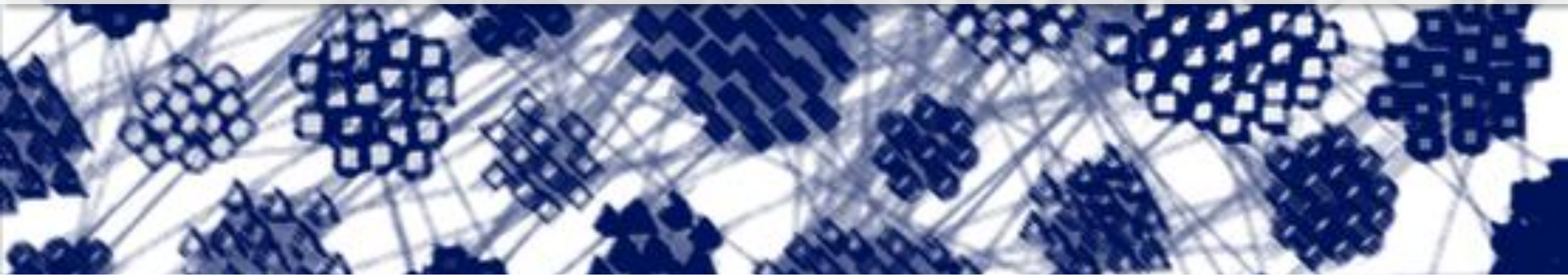
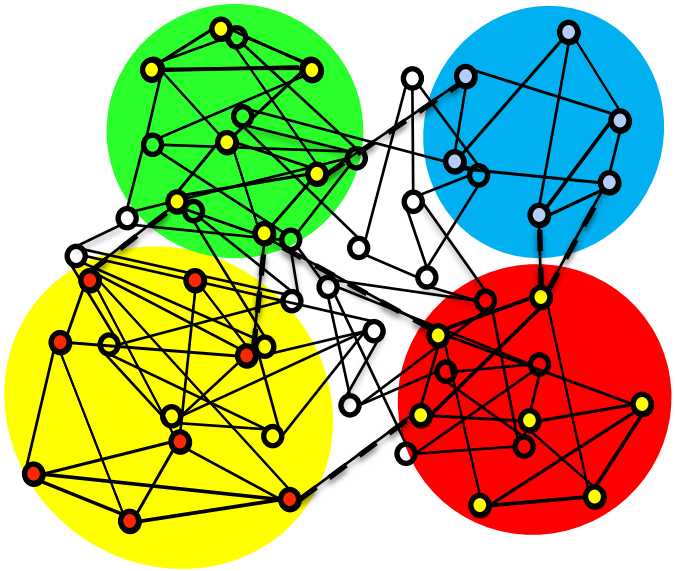
An Evaluation of Community Detection Algorithms on Large-Scale Email Traffic

Farnaz Moradi,
Tomas Olovsson, Philippas Tsigas



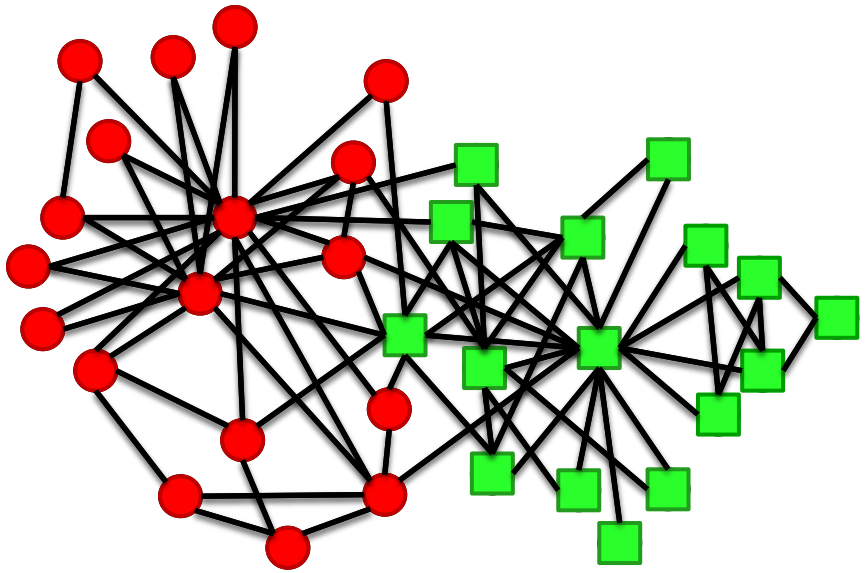
Community

- A community is a group of related nodes that
  - are densely interconnected
  - have fewer connections with the rest of the network



Community Structure

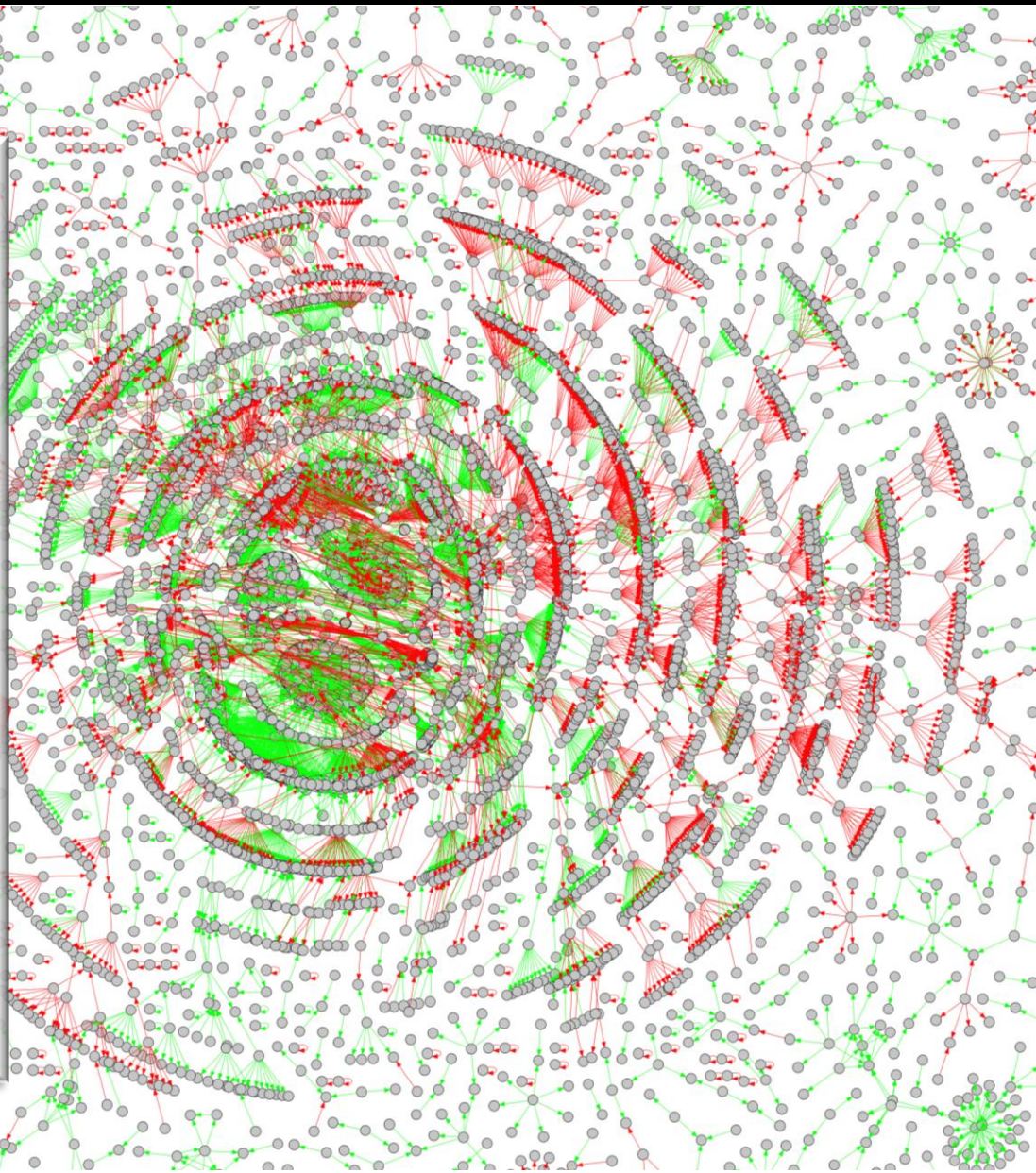
- Many real networks have community structure
  - Social networks
  - Web graph
  - P2P networks
  - Biological networks
  - Email networks



Zachary's Karate Club

- Community detection aims at unfolding the logical communities by only using the structral properties of the networks.

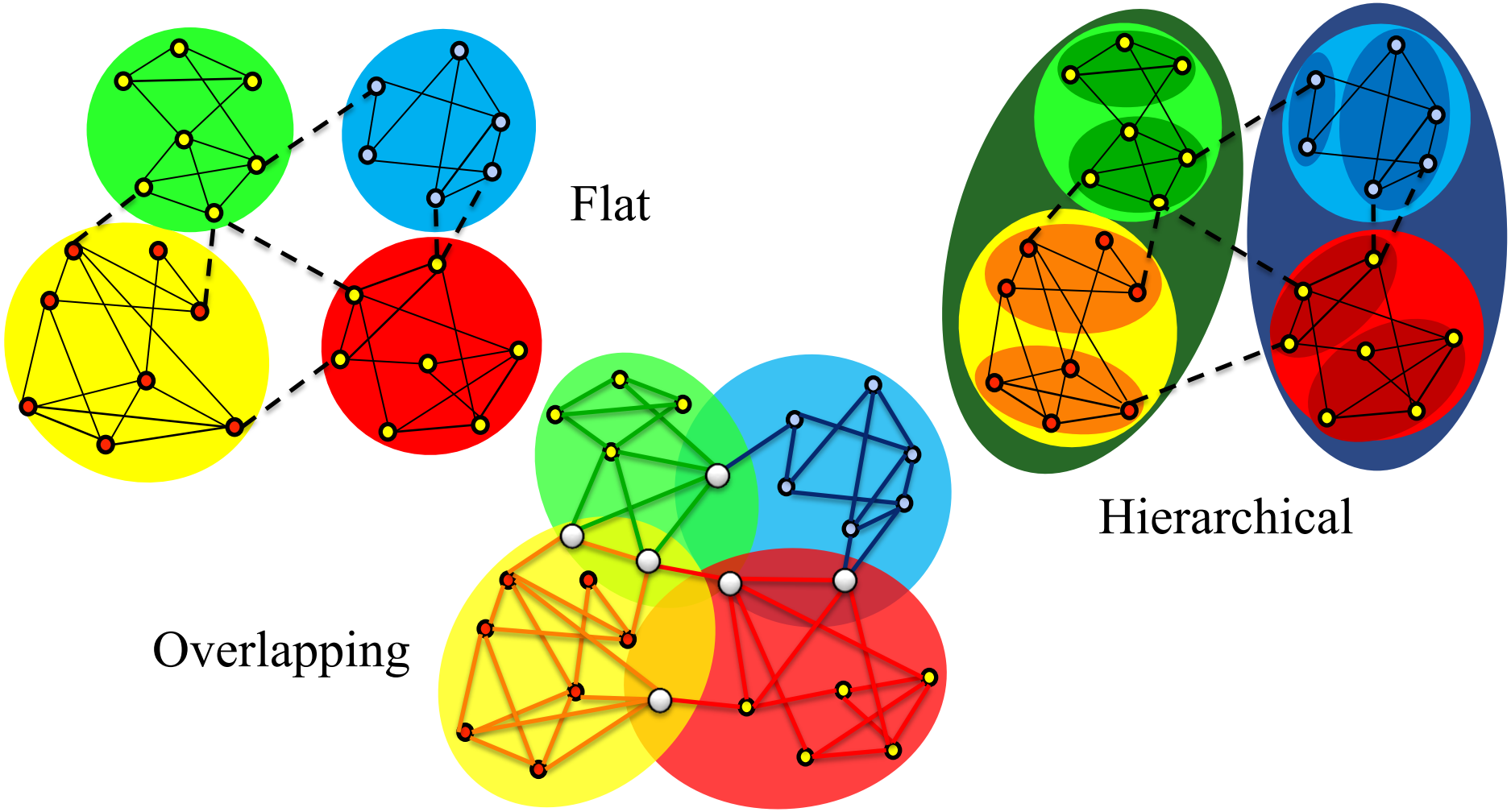
- Separating legitimate (ham) and unsolicited (spam) email in a large-scale email network generated from real email traffic.
- Assessing the quality of community detection algorithms in creating structural and logical communities.

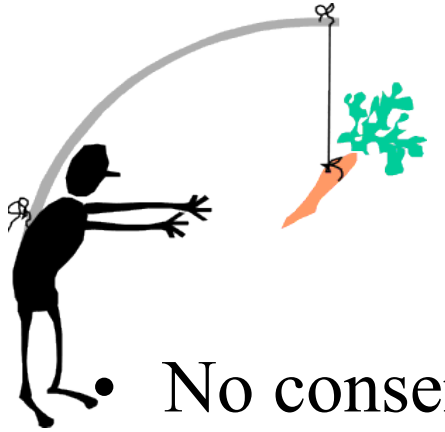


Outline

- Community detection algorithms
- Quality functions
  - Structural quality
  - Logical quality
- Experimental evaluation
  - Real email traffic

Community Detection





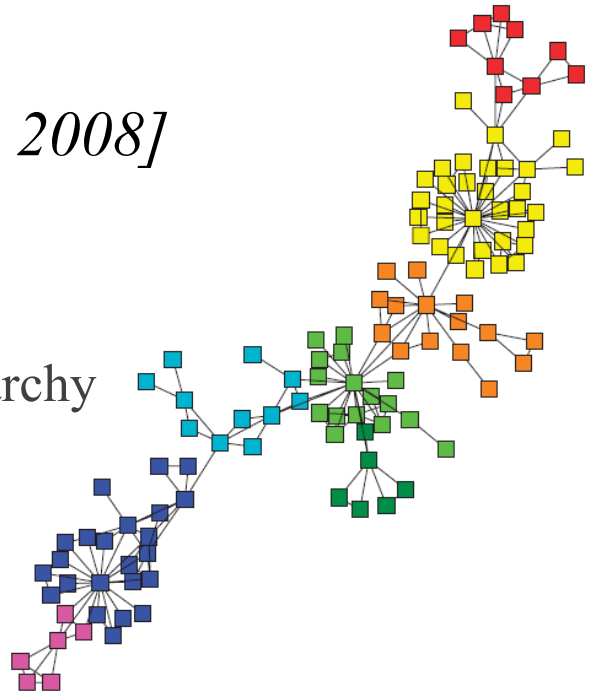
Motivation

Experimental Evaluation

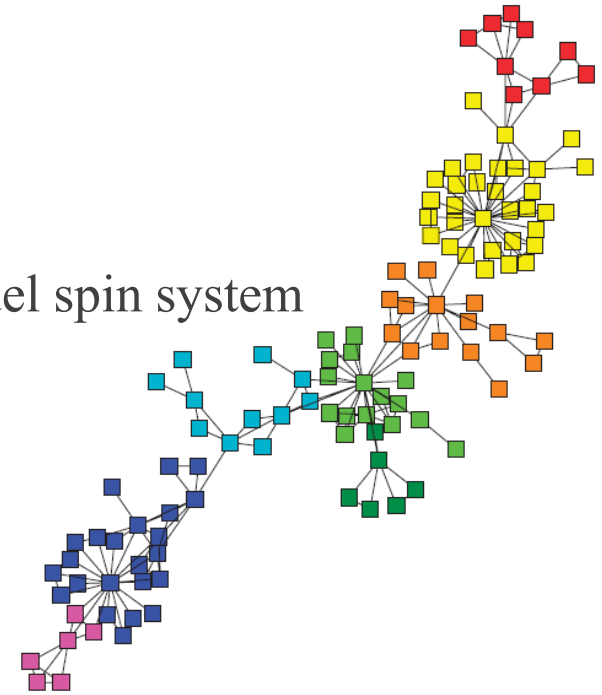
- No consensus on which algorithm is more suitable for which type of network.
- Experimental evaluation on synthetic graphs is not completely realistic *[Delling et al. 2006]*:
  - Implicit dependencies between:
    - community detection algorithms
    - synthetic graph generators
    - quality functions used to assess the performance of the algorithms
- Empirical studies on real-world networks are crucial.

Community Detection Algorithms

- Blondel (Louvian method), *[Blondel et al. 2008]*
  - Fast Modularity Optimization
  - Hierarchical clustering
  - Blondel L1: the first level of clustering hierarchy
- Infomap, *[Rosvall & Bergstrom 2008]*
  - Maps of Random Walks
  - Flow-based and information theoretic
- InfoH (InfoHiermap), *[Rosvall & Bergstrom 2011]*
  - Multilevel Compression of Random Walks
  - Hierarchical version of Infomap



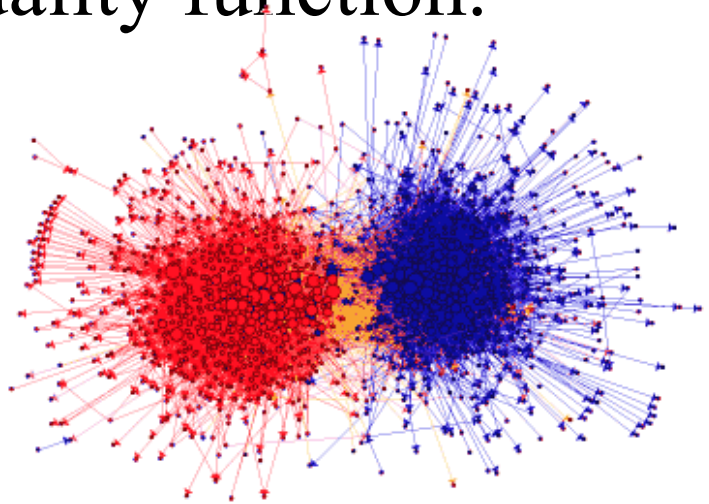
Community Detection Algorithms



- RN, *[Ronhovde & Nussinov 2009]*
  - Potts Model Community Detection
  - Minimization of Hamiltonian of an Potts model spin system
- MCL, *[Dongen 2000]*
  - Markov Clustering
  - Random walks stay longer in dense clusters
- LC, *[Ahn et al. 2010]*
  - Link Community Detection
  - A community is redefined as a set of closely interrelated edges
  - Overlapping and hierarchical clustering

Quality Functions

- Used to assess the quality of the algorithms when the true community structure of the network is not known.
- There is no single perfect quality function. *[Almedia et al. 2011]*
  - Structural quality
  - Logical quality



Structural Quality

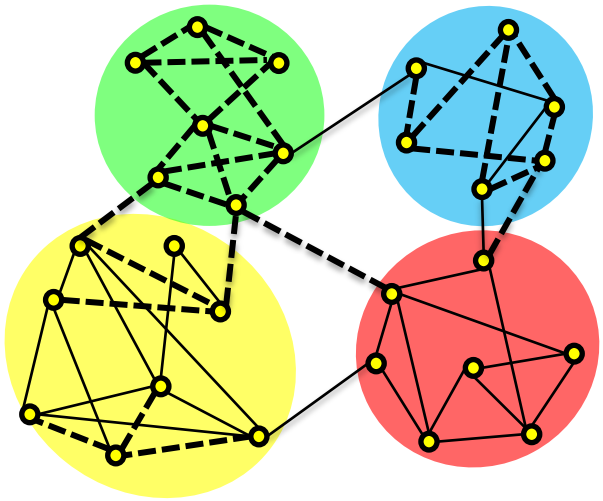
Coverage	$Cov(C) = \frac{m(C)}{m}$
Modularity	$Q(C) = \frac{m(C)}{m} - \frac{1}{4m^2}\sum_{c \in C}(\sum_{v \in c} \deg(v))^2$
Conductance	$\varphi(c) = \frac{\bar{m}(c)}{\min(\sum_{v \in c} \deg(v), \sum_{v \in V \setminus c} \deg(v))}$
Inter-cluster conductance	$\delta(C) = 1 - max_i \varphi(c_i), \; i \in \{1, \ldots, k\}$
Average conductance	$\frac{1}{\lvert c \rvert}\sum_{c \in C} \varphi(c)$

- Community coverage
- Overlap coverage

} Overlapping Clusterings

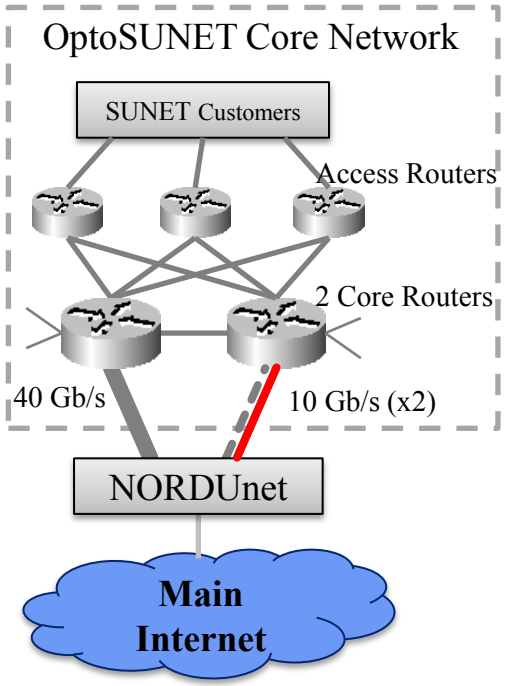
Logical Quality

- We define the logical quality based on the type of the edges inside the communities.
  - Homogeneous communities have perfect logical quality
  - The percentage of homogeneous communities in a network can be used to assess the logical quality of the network.



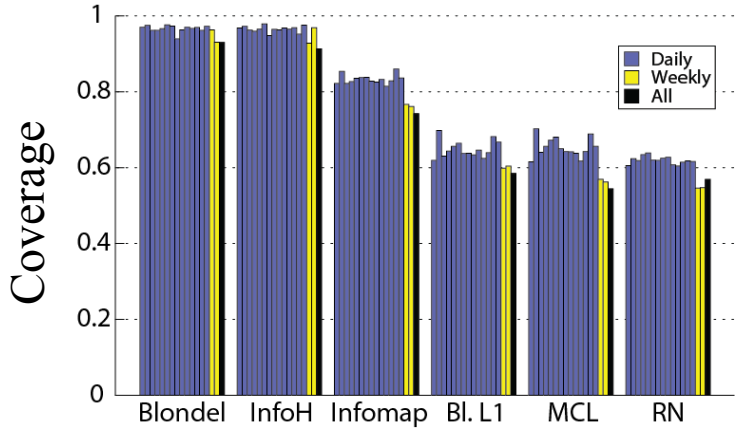
Experimental Evaluation

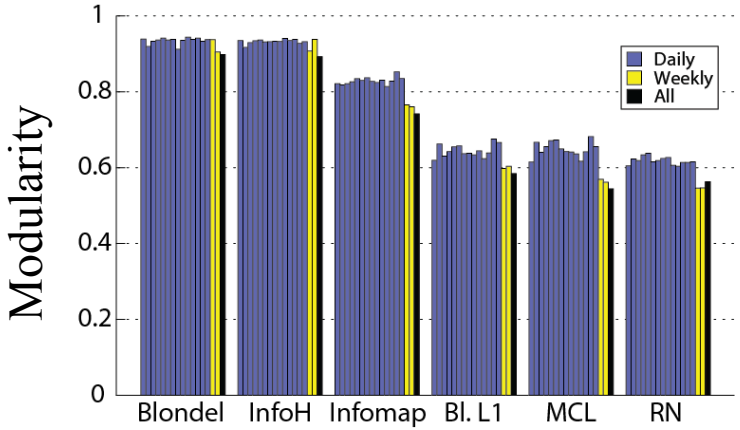
- Email traffic was collected on a 10 Gbps backbone link during 14 days
- Emails were classified as:
  - Legitimate (Ham)
  - Unsolicited (Spam)
- Implicit social network were created:
  - Nodes: Email addresses
  - Edges: Transmitted Emails
- Daily and weekly email networks were studied:
  - 14 daily networks
  - 2 weekly networks
  - 1 complete network
    - 1.6 million nodes and 2.8 million edges

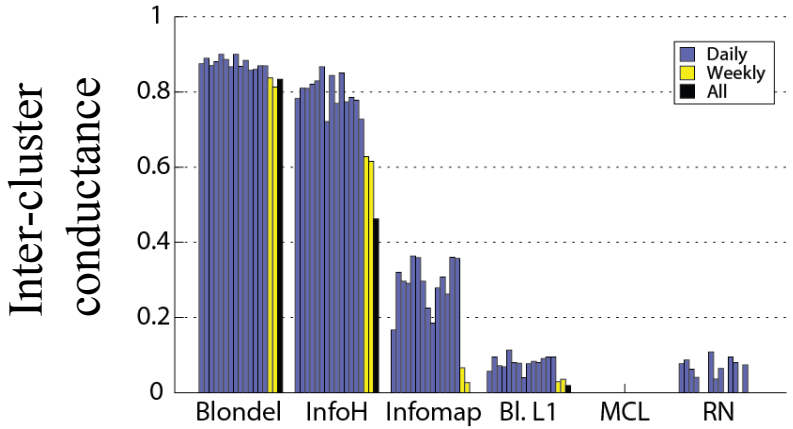


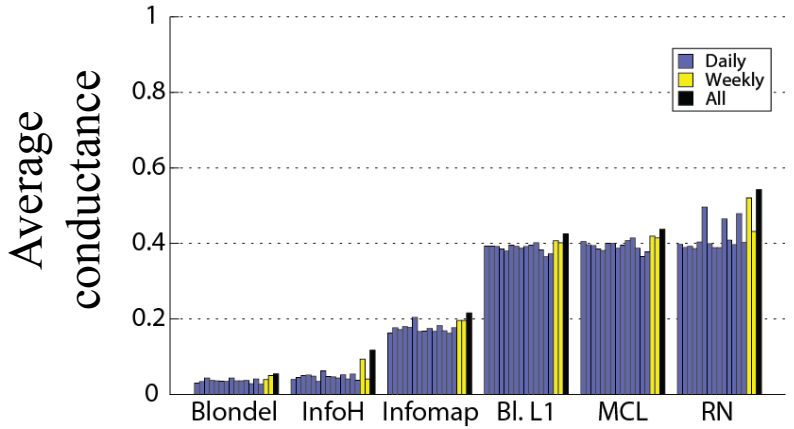
Experimental Results

Structural Quality





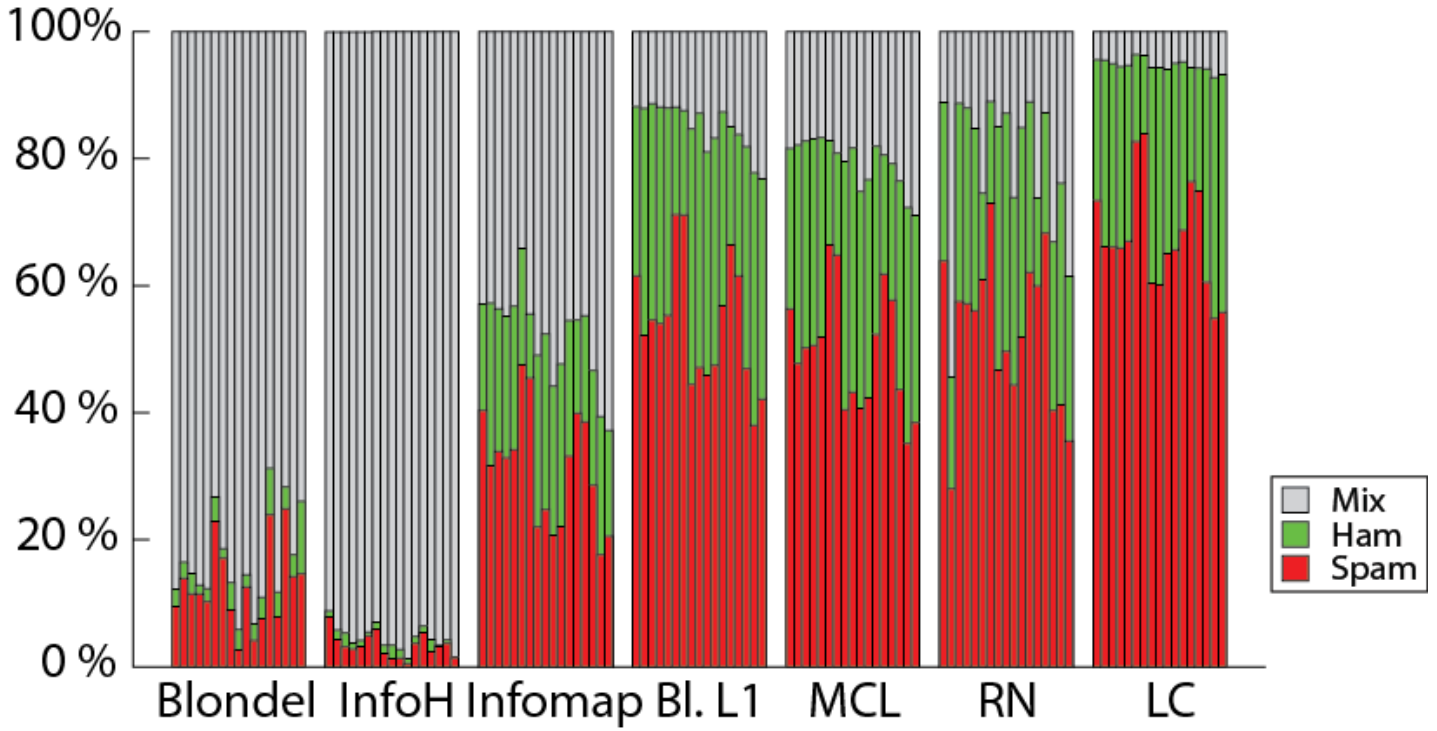




- *Community* and *overlap coverage* are used for assessing quality of LC

Experimental Results

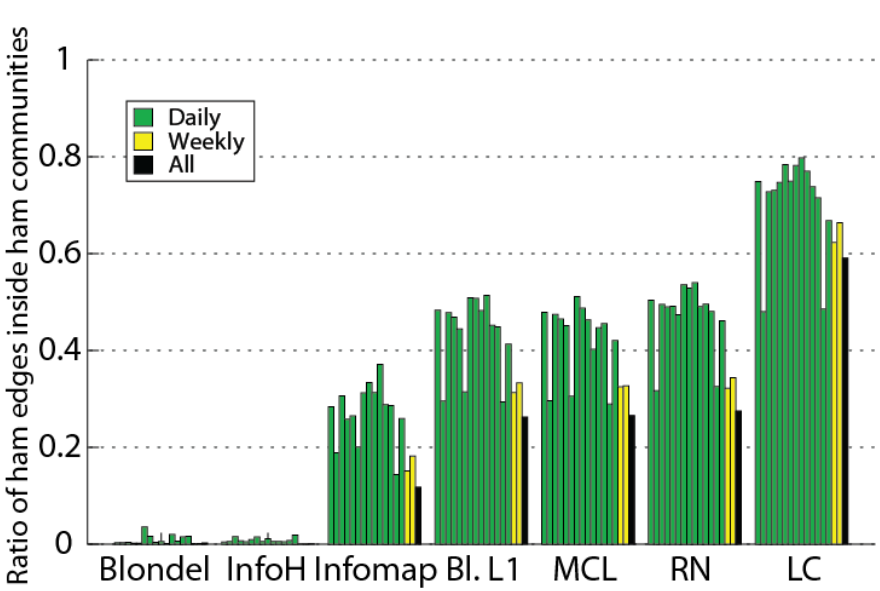
Logical Quality

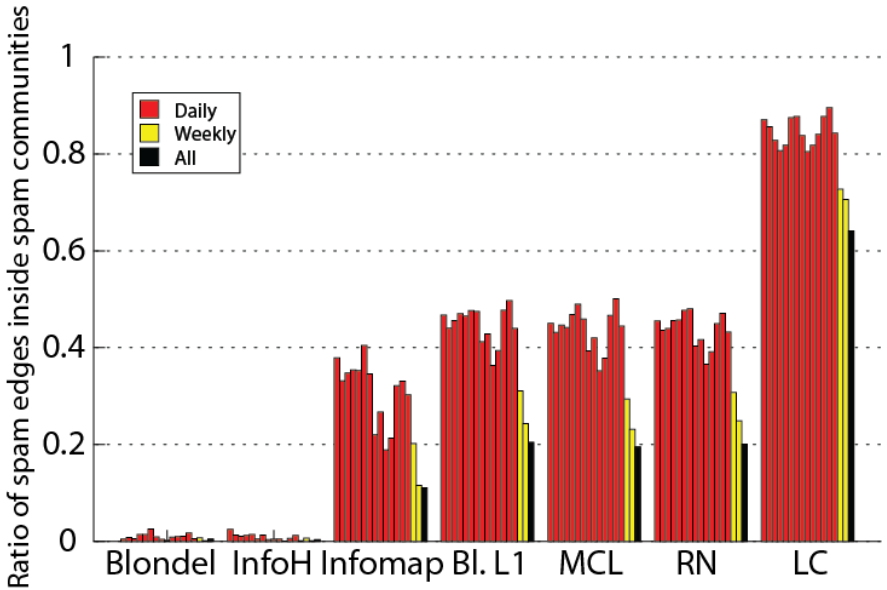


Comparison of the percentage of spam, ham, and mix communities

Experimental Results

Logical Quality





The amount of spam and ham emails that have been separated by community detection algorithms

Summary

- The algorithms that create coarse-grained communities achieve the best structural quality, but the worst logical quality.
  - Blondel and InfoH
- The algorithms that create communities with similar granularity, achieve similar structural and logical quality.
  - Blondel L1, MCL, and RN
- The algorithm that creates communities based on the edges of the network achieves the best logical quality.
  - LC

Conclusions

- Yielding high structural quality by community detection algorithms is not enough to unfold the true logical communities of the email networks.
- Link community detection is the most suitable approach for separating spam and ham emails into distinct communities.
- It is necessary to deploy more realistic measures for clustering real-world networks.

Thank You!