A Classification for Community Discovery Methods in Complex Networks

Michele Coscia

Harvard University & KDDLab ISTI CNR

June 10th, 2013

ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

Complex Networks

Mathematical model of interaction phenomena that take place in the real world

Examples:

- WWW: nodes = webpages; edges = hyperlinks
- Metabolic: nodes = enzymes; edges = reactions
- Facebook: nodes = profiles; edges = friendship
- ...

Communities in Complex Networks

A set of entities where each entity is closer, in the network sense, to the other entities within the community than to the entities outside it

Examples

WWW:

Pages of related topics

Metabolic Networks:

Functional modules in the cells

Facebook:

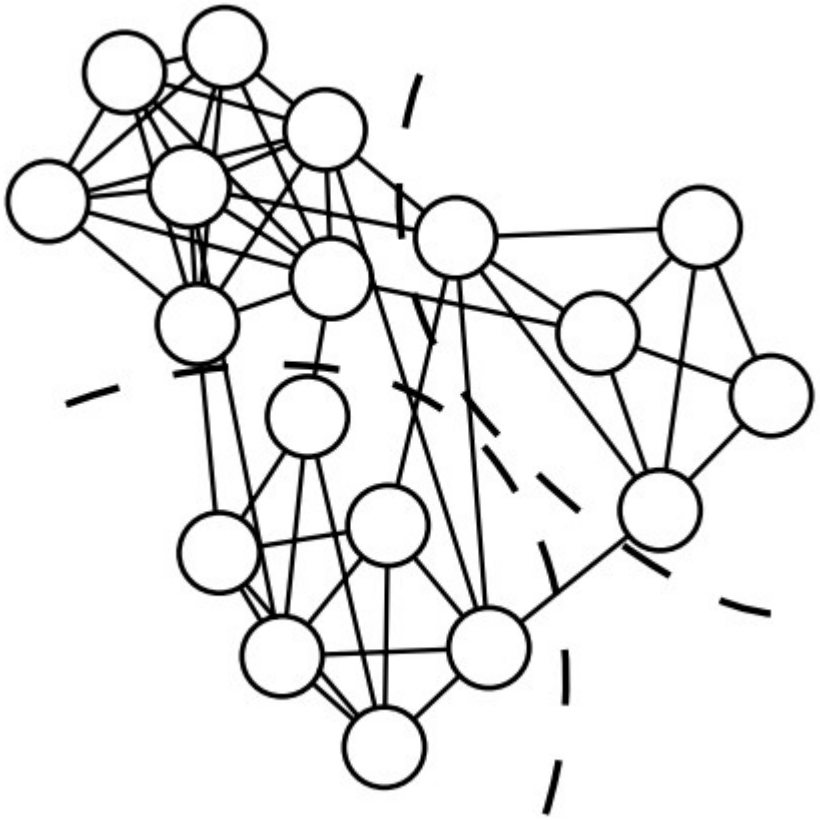
Groups of friends

Traditionally

Network proximity is based on the topology of the edges

Edges are not homogeneous in their distribution

Denser distinct groups can be isolated = Communities



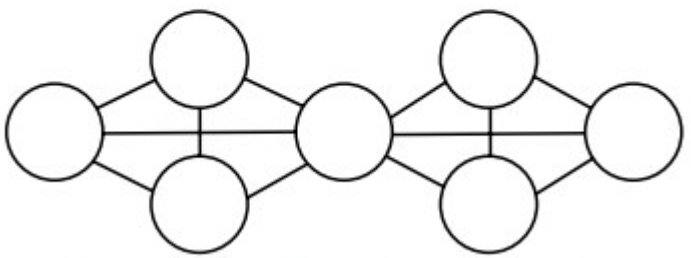
However

Network representations can be much richer than just nodes and edges

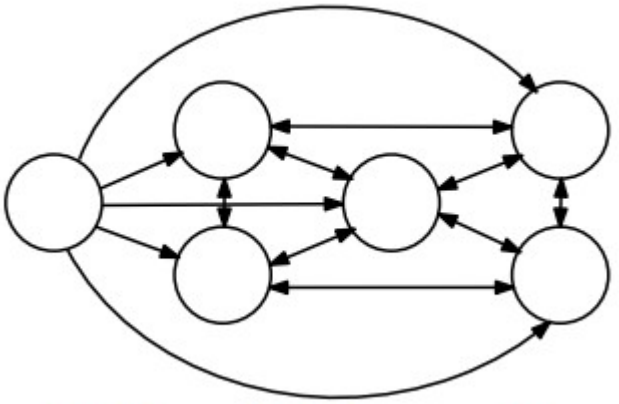
There are different ways to define a community in a network

Density is not a good quality measure to optimize anyway

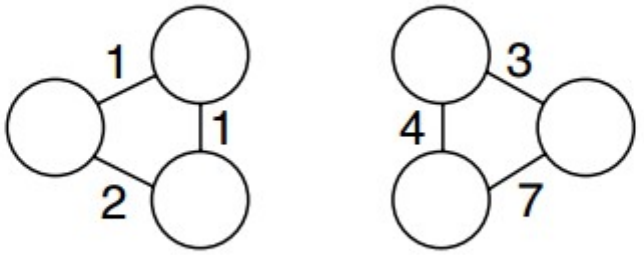
Some Features



(a) Overlapping Communities



(b) Directed Community



(c) Weighted Communities

Different Points of View

For some purposes, a community may be:

- Nodes connected by almost all possible edges (quasi-clique)
- Nodes that perform the same actions
- Nodes that will stay together if we delete some connections from the network
- ...

Density Problems

Real-world networks are sparse

Fat-tail degree distribution

Cliques maximize density, but they are a too-stringent criterion

Detecting cliques is time expensive

Our Proposal

To identify the most used different community definitions and to classify the most important works according to the community definition that they use

Community Definitions

1) Feature Distance
2) Internal Density
3) Bridge Detection
4) Diffusion
5) Closeness
6) Structure Definition
7) Link Clustering
8) Meta Clustering

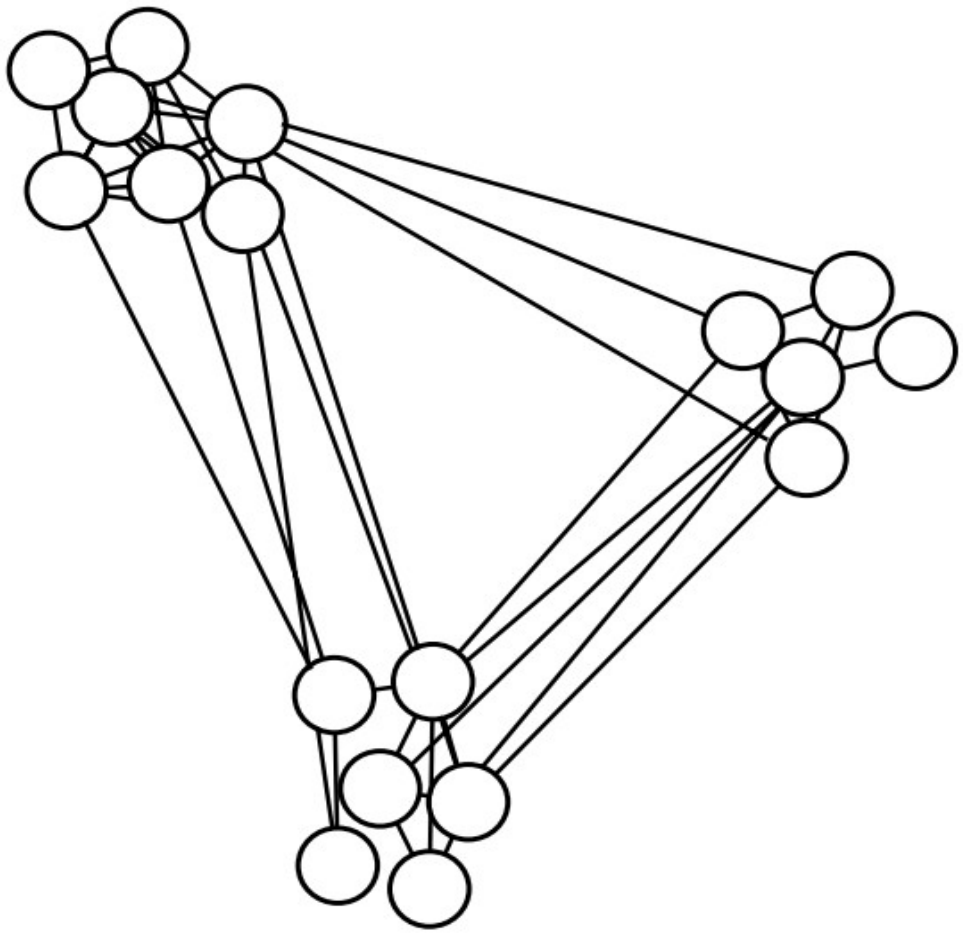
1) Feature Distance

Nodes have properties, edges, actions...

We can represent all of them as a vector of features

We then use some clustering algorithm (k-Means, SOM, …) to cluster them

In the example, besides the edges we have two features that we used for the node coordinate



1) Feature Distance

B. Long, X. Wu, Z. M. Zhang, and P. S. Yu, Unsupervised learning on k-partite graphs, KDD 2006

L. Tang and H. Liu, Relational learning via latent social dimensions, KDD 2009

L. Tang, X. Wang, and H. Liu, Uncovering groups via heterogeneous interaction analysis, ICDM 2009

A. Banerjee, S. Basu, and S. Merugu, Multi-way clustering on relation graphs, SDM 2007

C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda, Learning systems of concepts with an infinite relational model, AAA 2006

L. Friedland and D. Jensen, Finding tribes: identifying close-knit individuals from employment patterns, KDD 2007

D. Chakrabarti, Autopart: parameter-free graph partitioning and outlier detection, PKDD 2004

J. Ferlez, C. Faloutsos, J. Leskovec, D. Mladenic, and M. Grobelnik, Monitoring network evolution using MDL, ICDE 2008

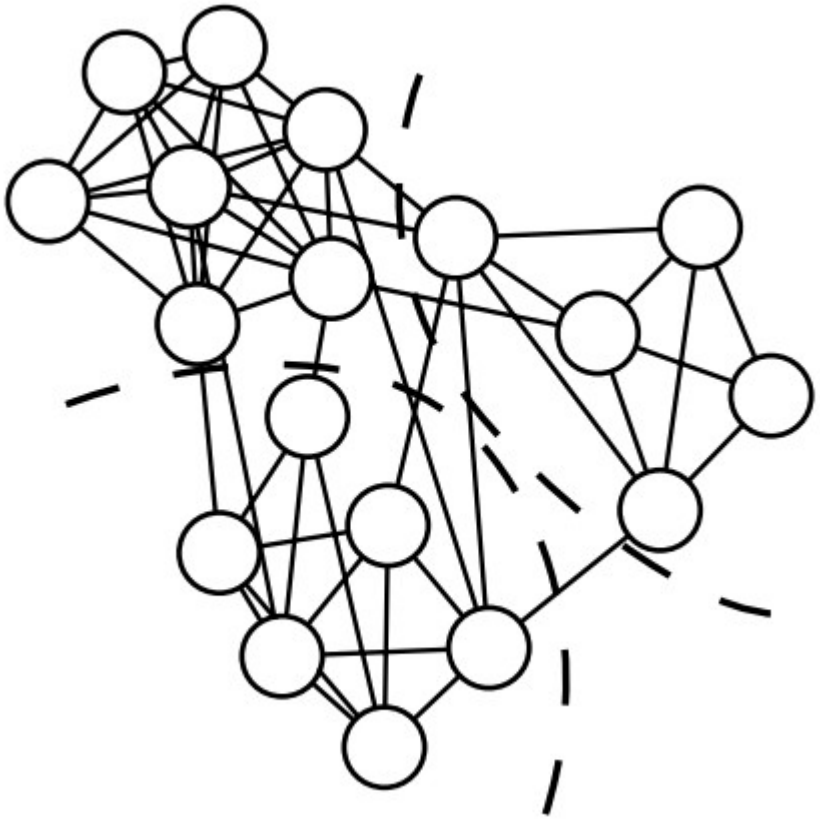
S. Papadimitriou, J. Sun, C. Faloutsos, and P. S. Yu, Hierarchical, parameter-free community discovery, PKDD 2008

2) Internal Density

Nodes connect to other similar nodes

If a group of nodes has more edges than expected (given the degree distribution) then those nodes have to be something in common

Usually we identify them by maximizing a density function like the Modularity



2) Internal Density

A. Clauset, M. E. J. Newman, and C. Moore, Finding community structure in very large networks, Phys Rev E 2004

Y.-R. Lin, J. Sun, P. Castro, R. Konuru, H. Sundaram, and A. Kelliher, Metafac: community discovery via relational hypergraph factorization, KDD 2009

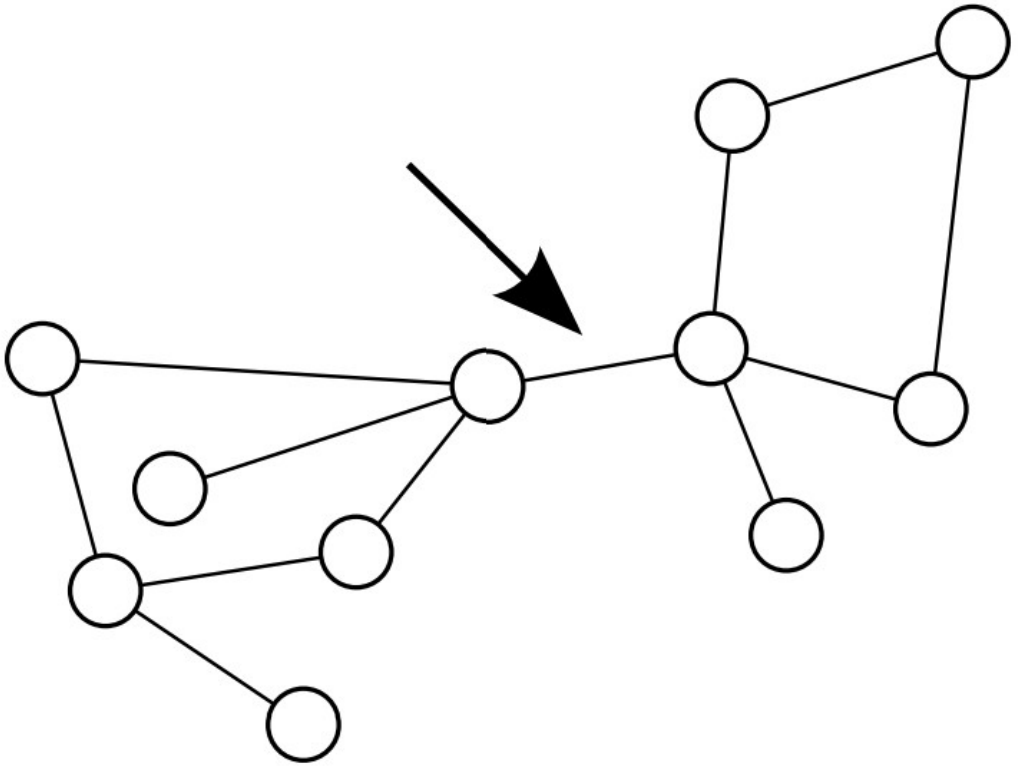
J. M. Hofman, and C. H. Wiggins, A bayesian approach to network modularity, Phys Rev Lett

3) Bridge Detection

Communities are semi-isolated modules of the network

Few edges (bridges) connect between them

If we remove them, we end up with our communities



3) Bridge Detection

M. Girvan and M. E. J. Newman, Community structure in social and biological networks, PNAS 2002

S. Gregory, A fast algorithm to find overlapping communities in networks, PKDD 2008

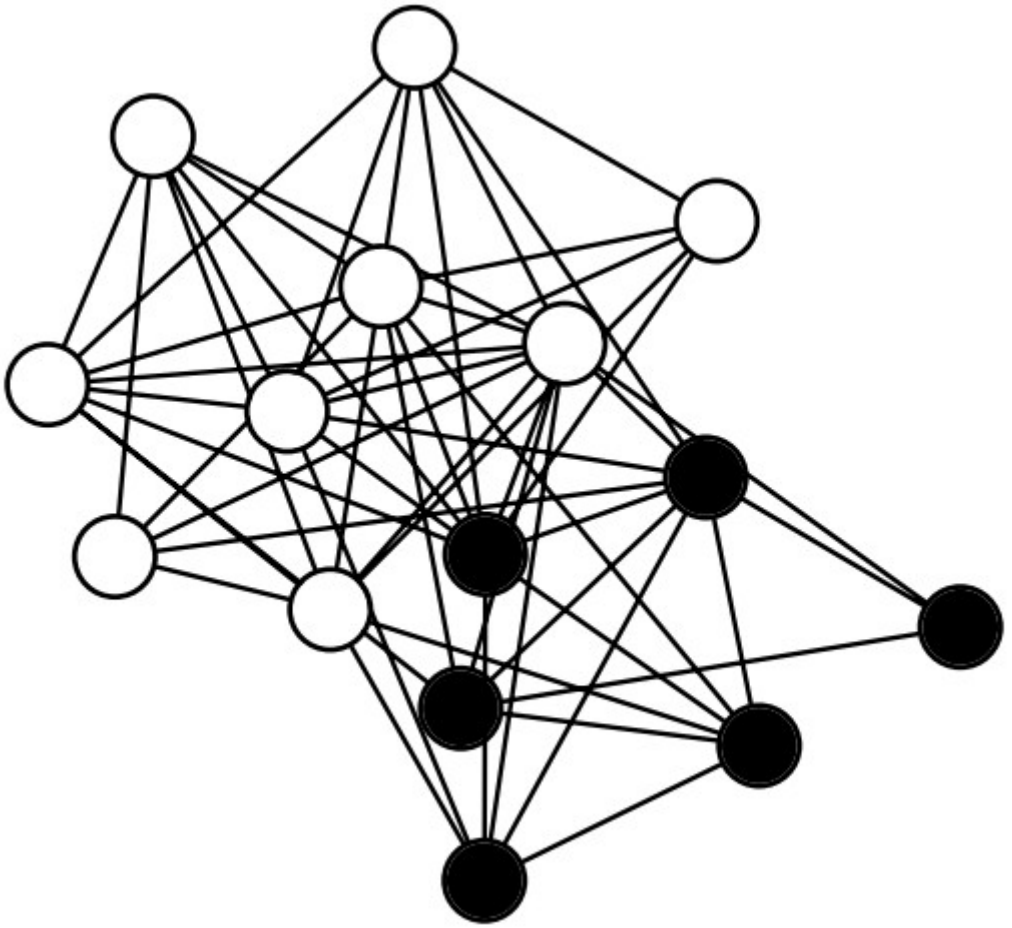
J. Bagrow and E. Bollt, A local method for detecting communities, Phys Rev E 2005

4) Diffusion

Nodes perform actions or are subject to contagion

After the event, nodes that end up in the same state with other nodes locally have something to do

We perform simulated diffusing events (percolation, social influence, etc...) and we extract the nodes which end up in the same state



4) Diffusion

A. Goyal, F. Bonchi, and L. V. Lakshmanan, Discovering leaders from community actions, CIKM 2008

U. N. Raghavan, R. Albert, and S. Kumara, Near linear time algorithm to detect community structures in largescale networks, Phys Rev E 2007

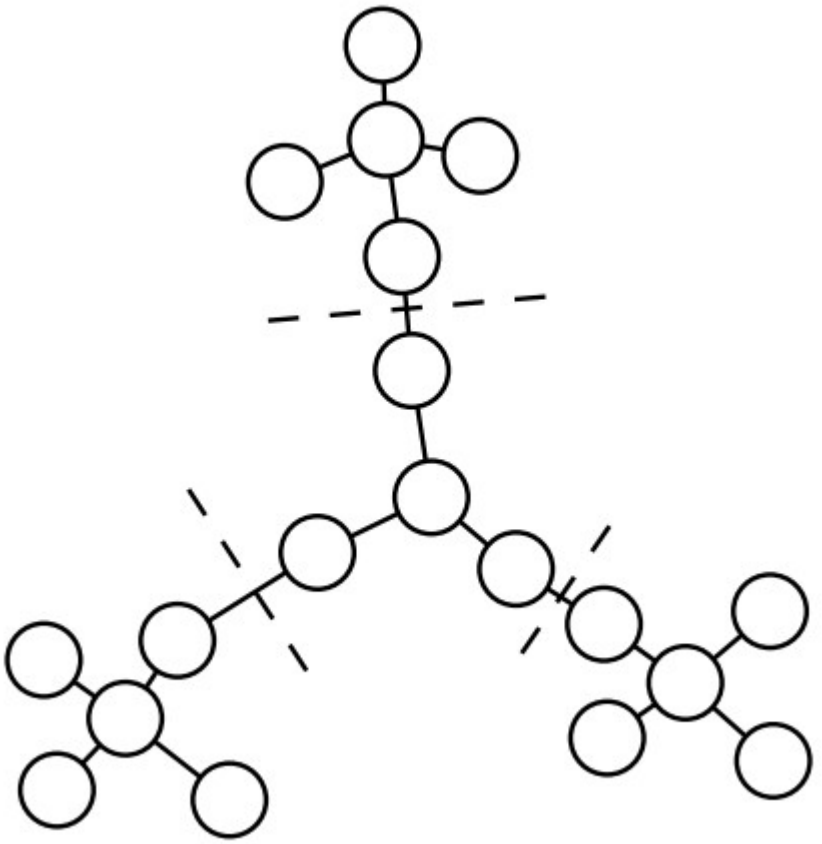
C. Tantipathananandh, T. Berger-Wolf, and D. Kempe, A framework for community

5) Closeness

Communities are groups of nodes that can reach the fellow community members in few hops

We perform random walks on the network

Nodes that constantly end up in the same random walks are in the same community



5) Closeness

P. Pons, and M. Latapy, Computing communities in large networks using random walks, J Graph Algor Appl 2006

F. Wei, W. Qian, C. Wang, and A. Zhou, Detecting overlapping community structures in networks, World Wide Web 2009

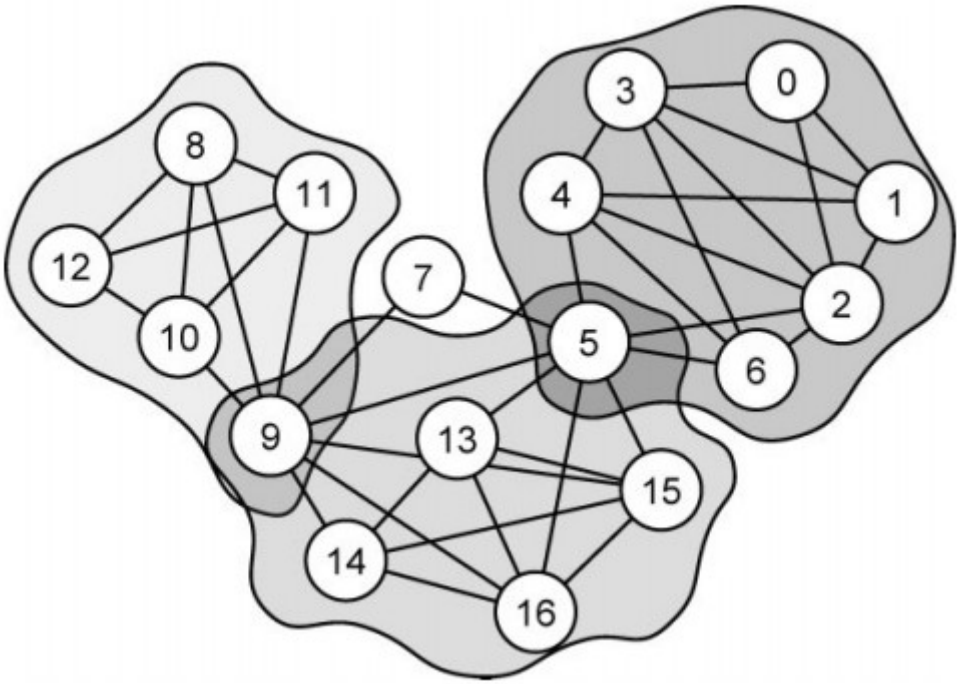
M. Rosvall, and C. T. Bergstrom, Maps of random walks on complex networks reveal community structure, PNAS 2008

6) Structure Definition

Sometimes we know exactly the exact structure of the communities we are looking for

We may look for cliques, or quasi-cliques with only a given amount of edges missing

We create a procedure (similar to graph mining) to extract the given structures



6) Structure Definition

G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 2005

S. Lehmann, M. Schwartz, and L. K. Hansen, Bi-clique communities, Phys Rev 2008

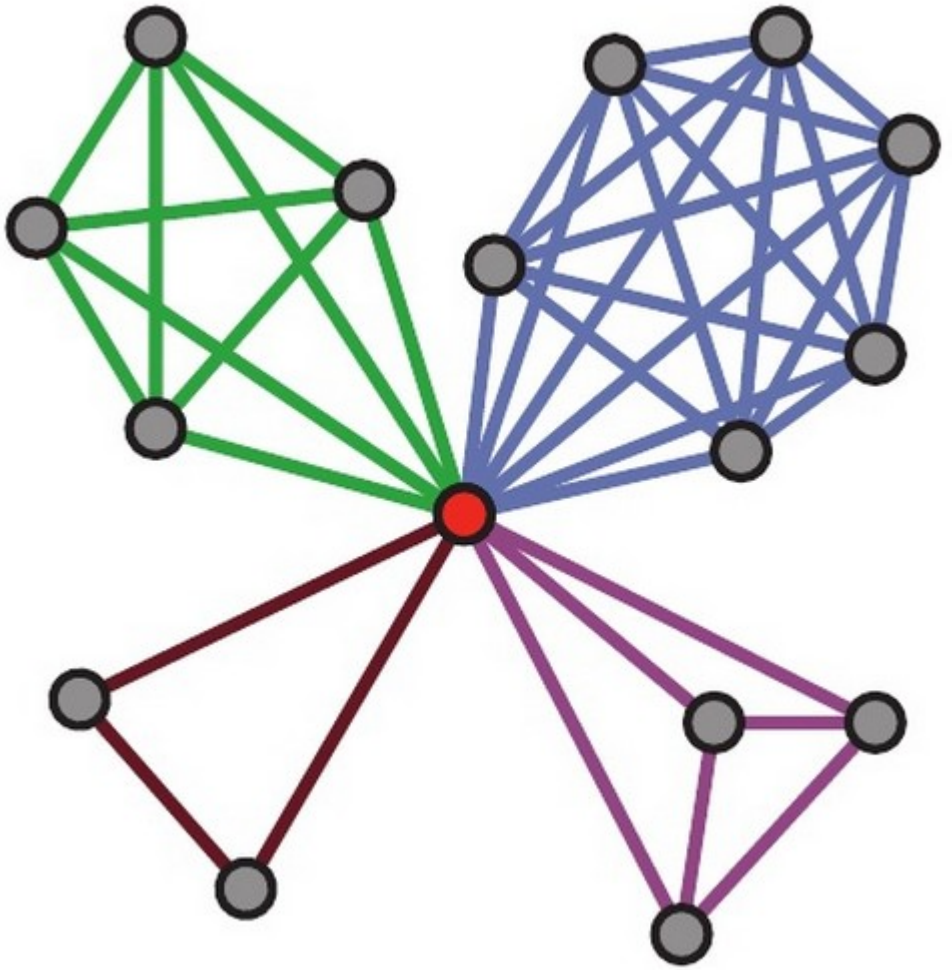
C. Komusiewicz, F. Huffner, H. Moser, and R. Niedermeier, Isolation concepts for efficiently enumerating dense subgraphs, Theor Comput

7) Link Clustering

We are not interested in clustering nodes, but links

It is the connection that belongs to a given “relational environment”

Nodes just belong to all the communities of their links



7) Link Clustering

T. S. Evans and R. Lambiotte, Line graphs, link partitions and overlapping communities, Phys Rev E 2009

Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, Link communities reveal multi-scale complexity in networks, Nature 2010

B. Ball, B. Karrer, and M. E. J. Newman, An efficient and principled method for detecting communities in networks, ArXiv e-prints, 2011

8) Meta Clustering

Some community discovery methods do not have a proper definition of what a community is

Sometimes, they just add some features (e.g. overlap) on top of the communities extracted with another method

Sometimes, they ask the analyst to provide a community definition and then they operate with it

8) Meta Clustering

M. E. J. Newman and E. A. Leicht, Mixture models and exploratory analysis in networks, PNAS 2007

T. Eliassi-Rad, K. Henderson, S. Papadimitriou, and C. Faloutsos, A hybrid community discovery framework for complex networks, SDM 2010

D. Cai, Z. Shao, X. He, X. Yan, and J. Han, Community mining from multi-relational

Conclusion

Community discovery in networks is a complex problem

We presented some of the main definitions of communities

We provided a framework for analysts who want to use an approach given her own community definition

Thank you

Questions?

ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"