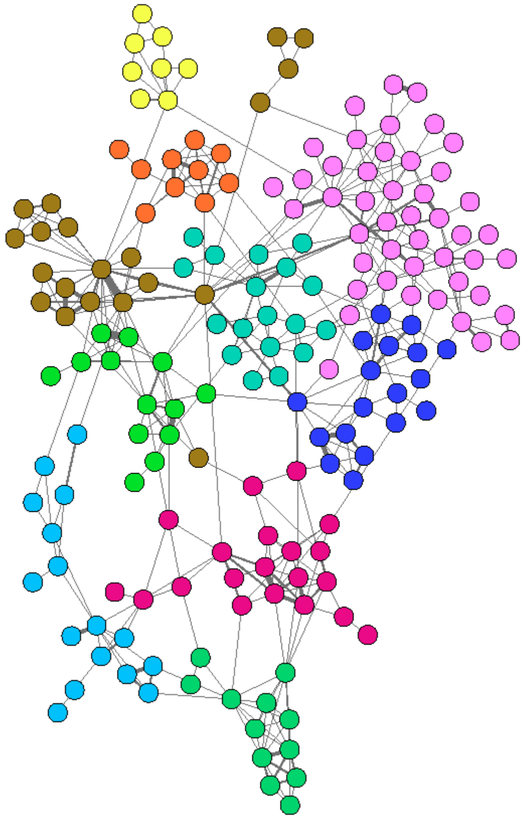
Community detection in networks

Dafne van Kuppevelt

2018-05-31 Statistics SIG



Find groups of nodes that are relatively close to each other in the network

Why?
To understand meso-scale structure, cohesion/fragmentation, relationships between individual nodes

Image: Frank Takes

Different methods for community detection

- Modularity
- Generative models
- Other models

Intuition: More edges **within-community** edges than expected.

Optimize number of **within-community edges** compared to **expected number of within-community edges.**

Intuition: More edges **within-community** edges than expected.

Optimize number of **within-community edges** compared to **expected number of within-community edges.**

$$Q = \frac{1}{2m}\sum_{ij}(A_{ij} - \frac{k_i k_j}{2m})\delta(C_i, C_j)$$
$$= \frac{1}{2m}\sum_{c}(m_c - \frac{K_c^2}{4m})$$

A_{ij} = Adjacency between node i and j
C_i = Class assignment of node i
k_i = degree of node i
m_c = nr of edges in community *c*
K_c = total degree of nodes in community c

Upside:

- Fast approximate optimization (Louvain Method)
- No fixed nr of communities

Downside:

- Resolution limit: prefers communities of certain sizes
- Exponential many near-optimal results
- Optimization gives no significance

Intuition: there is a **latent assignment in classes**, and links between nodes are formed based on a distribution c**onditioned on the class** assignment.

C_i = class assignment of node I
$P_{ij} = \Omega_{C_i,C_j}$ = probability of link forming between node i and j

Degree-corrected SBM:
$P_{ij} \sim d_i d_j \Omega_{C_i,C_j}$

To find communities with the SBM model, infer parameters by **maximizing likelihood**.

Model selection is necessary to determine the number of communities.

Upside:

- Explicit assumptions in model
- Can give statistical significance
- Can generate benchmark models

Downside:

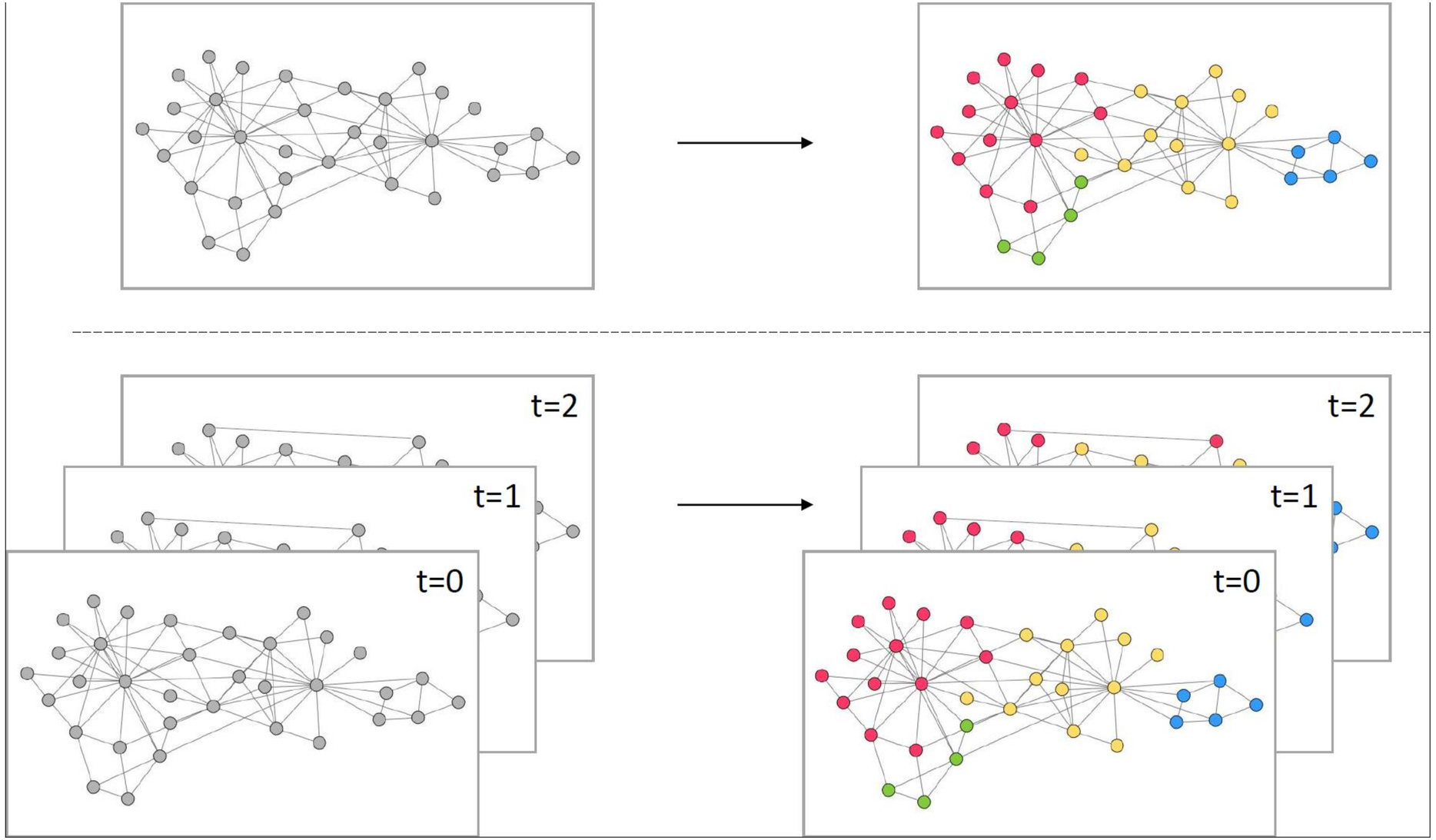
- Assumptions in model are not reflecting reality

Flow based models: Communities derived from behavior of random walks or flows on the network

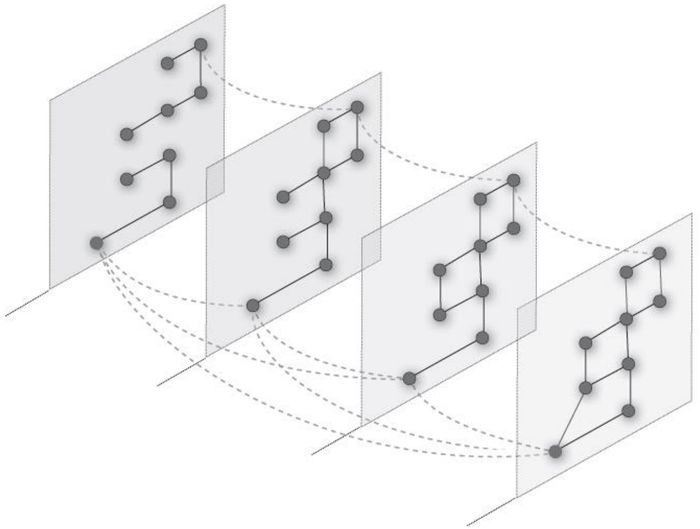
Structural methods: k-cliques. Only applicable in certain domain

Similarity based methods: (hierarchical) clustering on structural similarity matrix

Time-evolving networks



$$Q_{\text{multislice}} = \frac{1}{2\mu}\sum_{ijsr}\left[\left(A_{ijs} - \gamma_s \frac{k_{is}k_{js}}{2m_s}\right)\delta_{sr} + \delta_{ij}C_{jsr}\right]\delta(g_{is}, g_{jr})$$



Mucha et al. (2010): add **coupling** term to modularity function for intra-frame connections

- Many approaches based on Hidden Markov Model + Stochastic Block Model
- Approaches based on processes for adding and removing links
- Generative models for growing models (e.g. citation networks)

- Directed (acyclic graphs), such as citation networks
- Weighted graphs
- Nodes are not constant over time
- Bimodal networks

References

[1] Rosvall, M., Delvenne, J.-C., Schaub, M. T. and Lambiotte, R. (2017) ‘Different approaches to community detection’, in. Available at: https://arxiv.org/pdf/1712.06468.pdf (Accessed: 12 April 2018).

[2] Fortunato, S. and Hric, D. (2016) ‘Community detection in networks: A user guide’, *Physics Reports*, 659, pp. 1–44. doi: 10.1016/j.physrep.2016.09.002.

[3] Good, B. H., de Montjoye, Y.-A. and Clauset, A. (2010) ‘Performance of modularity maximization in practical contexts’, *Physical Review E*, 81(4), p. 046106. doi: 10.1103/PhysRevE.81.046106.

[4] Mucha, P. J., Richardson, T., Macon, K., Porter, M. A. and Onnela, J.-P. (2010) ‘Community Structure in Time-Dependent, Multiscale, and Multiplex Networks’, *Science*, 328(5980), pp. 876–878. doi: 10.1126/science.1184819.

[5] Coscia, M., Giannotti, F. and Pedreschi, D. (2011) ‘A classification for community discovery methods in complex networks’, *Statistical Analysis and Data Mining*, 4(5), pp. 512–546. doi: 10.1002/sam.10133.