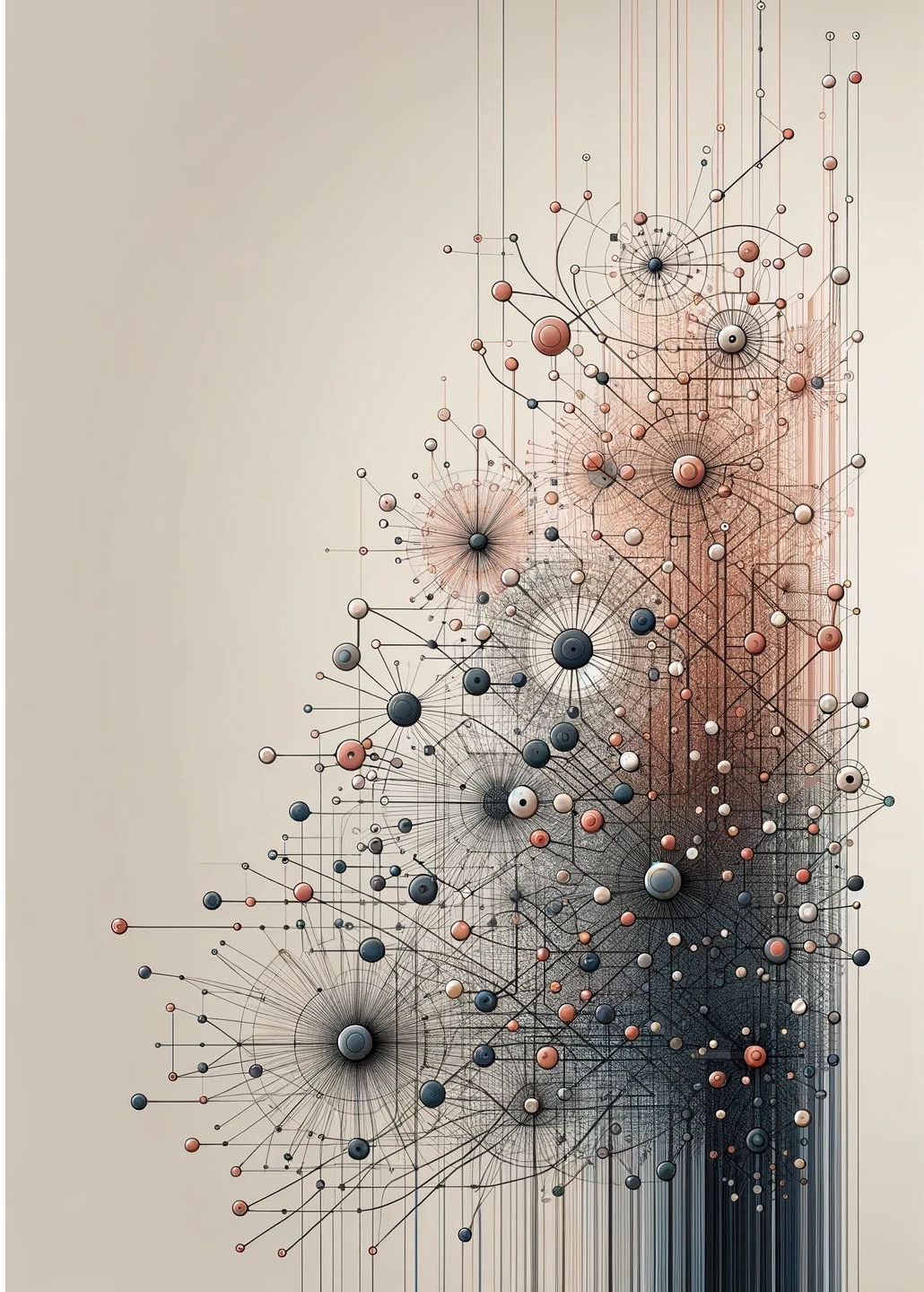
UNIVERSITÀ DI TORINO

Analisi e Visualizzazione delle Reti Complesse

NS14 - Communities - Benchmarks and Tutorial

Prof. Rossano Schifanella



UNIVERSITÀ DI TORINO

Outline

- Method evaluation
- Artificial benchmarks
- Real benchmarks
- Partition similarity
- Exercises

UNIVERSITÀ DI TORINO

Method evaluation

Method evaluation

- **Problem:** How can we tell how good a clustering algorithm is?
- **Solution:** The natural way to evaluate a method is to check whether it is able to find clusters in benchmark graphs, i.e., networks known to have a natural community structure
- **Two classes of benchmarks:**
  - **Artificial benchmarks:** computer-generated networks, built via some model
  - **Real benchmarks:** real networks in which the communities are suggested by the history of the system or by attributes of the nodes

Planted partition model

- Stochastic block models (SBMs) are often used to generate artificial benchmarks

```
# network with communities with sizes in the list S
# and stochastic block matrix P

G = nx.generators.stochastic_block_model(S,P)
```

- Special version of SBMs regularly used in method evaluations: **planted partition model**
- The planted partition model has only **two link probabilities**: the probability p_{int} that nodes in the same community are connected and the probability p_{ext} that nodes in different communities are connected
- If $p_{int} > p_{ext}$ the groups are communities, as two nodes are more likely to be connected if they are within the same group than if they are in different ones

Planted partition model

- q groups of identical size $\frac{N}{q}$
- Expected internal degree of the nodes:

$$\langle k^{int} \rangle = p_{int} \left(\frac{N}{q} - 1 \right)$$

- Expected external degree of the nodes:

$$\langle k^{ext} \rangle = p_{ext} \frac{N}{q} (q - 1)$$

- Expected total degree of the nodes:

$$\langle k \rangle = \langle k^{int} \rangle + \langle k^{ext} \rangle = p_{int} \left(\frac{N}{q} - 1 \right) + p_{ext} \frac{N}{q} (q - 1)$$

In NetworkX:

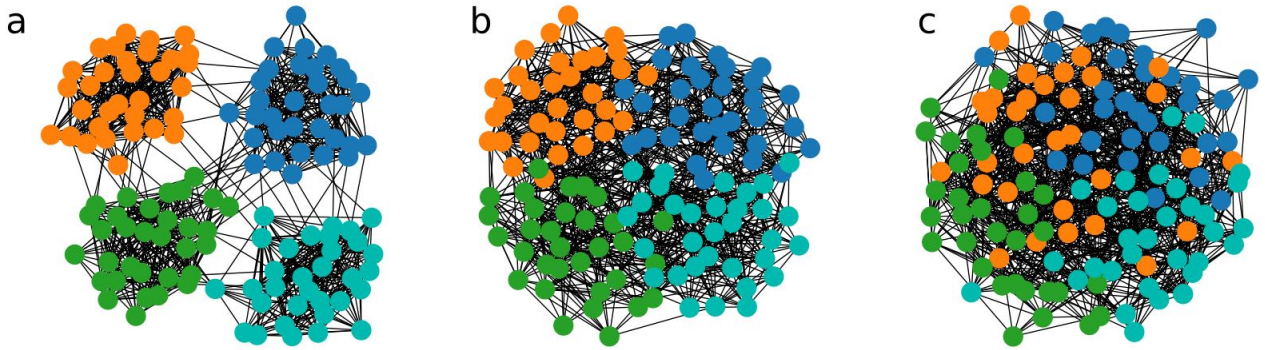
```
# network with q communities of nc nodes each
# and link probabilities p_int and p_ext

G = nx.generators.planted_partition_graph(q,nc,p_int,p_ext)
```

GN benchmark

- Specific implementation of the planted partition model, in which the network size, nodes degree, and number and size of the communities are set to particular values:
  - $N = 128$, $q = 4$, $\langle k \rangle = 16$
- Since
  - $31 p_{int} + 96 p_{ext} = 19$
- Then
  - p_{int} and p_{ext} are not independent parameters
- **GN (Girvan-Newman) benchmark** networks are constructed with a procedure similar to the one adopted for Erdős–Rényi random graphs:
  - we go over all pairs of nodes and connect each with probability p_{int} or p_{ext}, depending on whether or not the nodes are in the same community

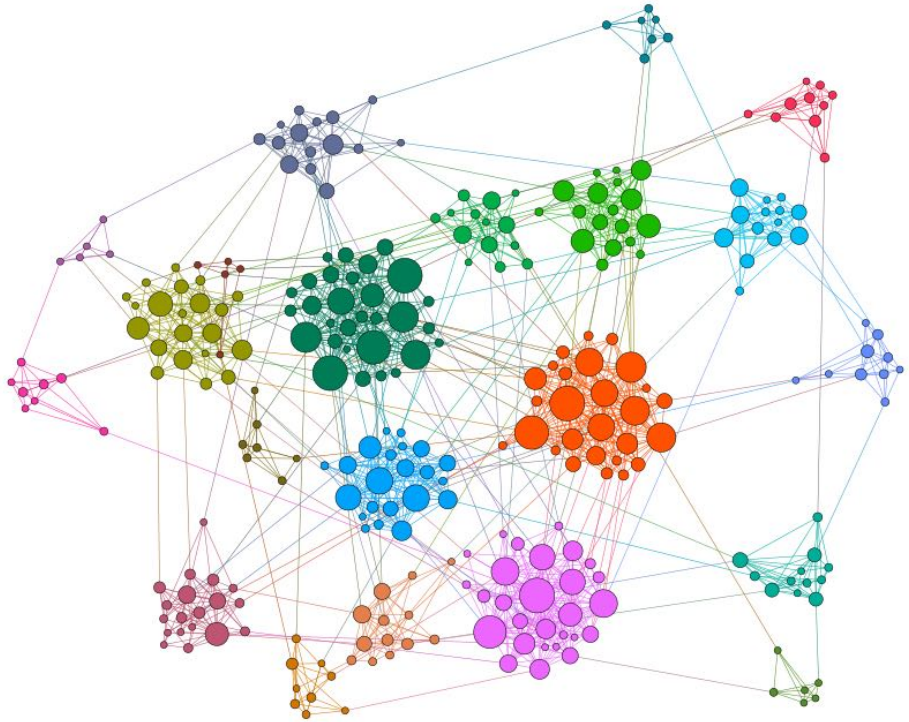
GN benchmark

- The higher the expected external degree and the lower the expected internal degree, the more difficult it is to detect the communities
- **Expectation:** communities should be detectable as long as $p_{int} > p_{ext}$, i.e., $\langle k \rangle$ smaller than (about) 12
- **Surprise:** the detectability threshold is quite a bit lower, around 9, due to random fluctuations in the placement of the links (detectability limit)
  - Reading material: Girvan, M., and Newman, M. E. J. 2002. Community structure in social and biological networks. PNAS, 99(12), 7821–7826.



LFR benchmark

- The GN benchmark is not realistic!
- **Problem 1:** nodes have approximately the same degree
- **Problem 2:** communities have approximately the same size
- In real networks the distributions of node degree and community size are quite heterogeneous
- The **LFR (Lancichinetti–Fortunato–Radicchi) benchmark** produces networks having heavy-tailed distributions of degree and community size
  - Reading material: Benchmark graphs for testing community detection algorithms. Physical Review E, 78(4), 046110.



In NetworkX:

```
G = nx.generators.LFR_benchmark_graph(n, tau1, tau2, mu)
```

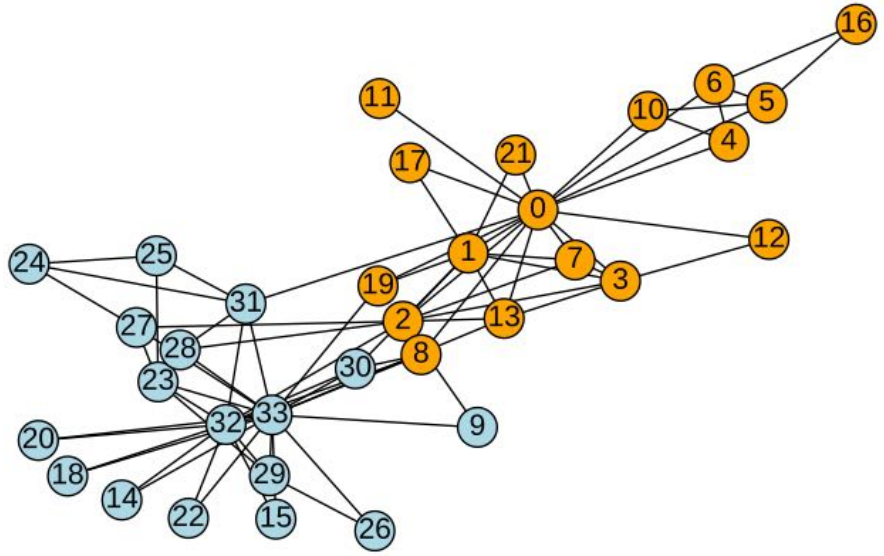
The importance of negative tests

- **Negative test:** is the algorithm finding communities also in networks without communities?
- **Example:** if a method finds non-trivial partitions in random networks, it is unreliable
- **Trivial partitions:** one cluster including everything or N clusters of one node each
- Any non-trivial partition would signal the method's inability to distinguish actual communities from subnetworks with high concentrations of links generated by random fluctuations

Real benchmarks

- Classic example: Zachary's karate club network
- Network of social relationships between members of a karate club in the US
- Following a disagreement between the instructor (node 0) and the president (node 33), the club split in two different groups
- The task is to identify those groups (identified by the node colors) using a clustering algorithm

```
G = nx.karate_club_graph()
```



Real benchmarks

- Many other networks, whose nodes can be classified based on their attributes, are available for testing community detection algorithms
- **Examples:** in many social networks there are groups that users can decide to join, in citation networks papers can be grouped according to their publication venues, Internet routers can be categorized by country, etc.
- **Question:** do clusters of nodes with similar attributes match the communities found by clustering algorithms based only on the network structure?
- **Answer:** if nodes with similar attributes are strongly linked to each other, the attributes can be revealed through community detection. If instead the attributes do not play a role in the build-up of the network, they remain invisible to clustering methods

Partition similarity

- **Question:** how can we tell how close the partition found by the algorithm is to the planted partition of a benchmark network?
- **Answer:** partition similarity measures!
- Different classes of measures, all of them with pros and cons
- Here we discuss two of them:
  - **Fraction of correctly detected nodes**
  - **Normalized mutual information**

Fraction of correctly detected nodes

- The simplest measure of success is just to count the **fraction of nodes that are classified into the correct groups**.
- **Caveat:** group labels can be shuffled
- **Solution:** one commonly calculates the fraction of correctly classified nodes as the maximum over all permutations of the group labels
- It might also be the case that the **number of groups found by the algorithm is not equal** to the number of groups in the ground truth.
    - The same permutation approach works in this case, except that one must **always permute the larger of the two sets of labels**.

Normalized mutual information

- **Intuition:** If we know one partition, how much information do we need to infer the other one?
- NMI measures how much knowledge about one partition reduces uncertainty about the other partition
- If the partitions are identical, knowing one perfectly predicts the other (NMI = 1)
- If they are completely independent, knowing one tells us nothing about the other (NMI = 0)
- **Formula:**

$$NMI(X, Y) = \frac{2 \times MI(X, Y)}{H(X) + H(Y)}$$

Where:

- $MI(X, Y)$ is the mutual information between partitions X and Y
- $H(X)$ and $H(Y)$ are the Shannon entropies of each partition

UNIVERSITÀ DI TORINO

Understanding NMI

- **Shannon entropy** $H(X)$ measures the uncertainty in partition X:

$$H(X) = -\sum_{x} P(x) \log(P(x))$$

Where $P(x) = \frac{N_x}{N}$ is the probability that a random node belongs to community x

- **Mutual information** measures how much knowing one partition reduces uncertainty about the other:

$$MI(X, Y) = \sum_{x \in X} \sum_{y \in Y} P(x, y) \log \frac{P(x, y)}{P(x)P(y)}$$

Where $P(x, y)$ is the joint probability (fraction of nodes in both community x and y)

- **Key properties:**
  - $NMI = 1$ if partitions are identical
  - $NMI \approx 0$ for independent (random) partitions
  - NMI increases as partitions become more similar

Adjusted Rand Index (ARI)

- **Alternative similarity measure** based on counting pairs of nodes
- **Intuition:** Count how many pairs of nodes are classified the same way in both partitions
- For each pair of nodes, we have four possibilities:
  - **a:** Nodes in same community in both partitions (agreement)
  - **b:** Nodes in same community in X, different communities in Y (disagreement)
  - **c:** Nodes in different communities in X, same community in Y (disagreement)
  - **d:** Nodes in different communities in both partitions (agreement)
- **Rand Index:** Fraction of node pairs where partitions agree
  - $RI = \frac{a+d}{a+b+c+d}$

Adjusted Rand Index (ARI)

- **Problem with Rand Index:** Even random partitions can have high RI by chance
- **Solution:** The Adjusted Rand Index corrects for chance

$$ARI = \frac{RI - \text{Expected RI}}{\text{Max RI} - \text{Expected RI}}$$

- **Formula in terms of contingency table:**

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}$$

Where:

- n_{ij} is the number of nodes in both community i of X and community j of Y
- a_i is the sum of the i-th row of the contingency table
- b_j is the sum of the j-th column

Comparing NMI and ARI

- Both measures range from values around 0 (random partitions) to 1 (identical partitions)
- **ARI advantages:**
  - Based on counting pairs, which is intuitive
  - Well-established in clustering literature
  - Less sensitive to the number of small communities
- **NMI advantages:**
  - Based on information theory
  - Better at comparing partitions with different numbers of communities
  - More widely used in community detection literature
- Best practice: **use both metrics** for a more robust evaluation

Reading material

References

[ns1] **Chapter 6 (Communities)**

Danon, L., Díaz-Guilera, A., Duch, J., & Arenas, A. (2005). Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment, 2005(09), P09008.

Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. Journal of Machine Learning Research, 11, 2837-2854.

UNIVERSITÀ DI TORINO

Q&A