Random Graphs

Zastosowanie grafów losowych do projektowania algorytmów

Paweł Prałat

Updated: 2021/10/23

Department of Mathematics, Ryerson University

File: Bedlewo

Overview

1. Introduction
2. Community Detection
3. Graph Embeddings

Introduction

Why do we care?

Theory of random graphs = the intersection of graph theory and probability.

Very active area of research. Why?

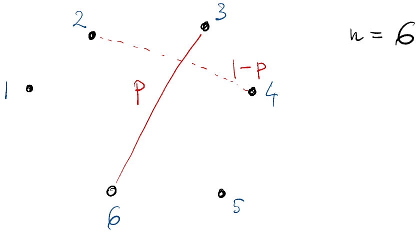
Why do we care?

(i) interesting and surprising objects, uncovering properties of typical graphs, supporting conjectures but sometimes provides counterexamples.

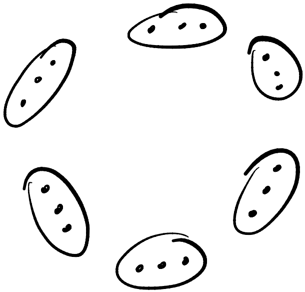
– Binomial random graphs (expected degree d), 1959
– Random d-regular graphs
– Chung-Lu model (given expected degree distribution $\mathbf{w}$)
– Random graphs with given degree distribution $\mathbf{w}$

Binomial Random Graph $\mathcal{G}(n, p)$

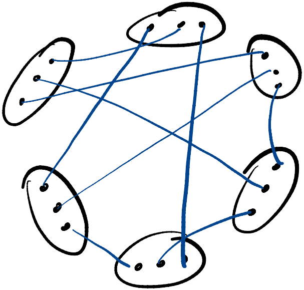
The binomial random graph $\mathcal{G}(n, p)$ can be generated by starting with the empty graph on the set of nodes $[n] = \{1, 2, \ldots, n\}$. For each pair of nodes i, j such that $1 \leq i < j \leq n$, we independently introduce an edge ij in G with probability p.



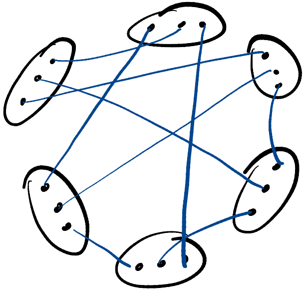
Random d-regular graphs

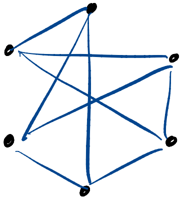


Random d-regular graphs

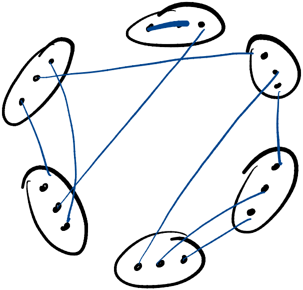


Random d-regular graphs

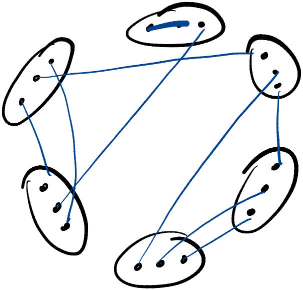


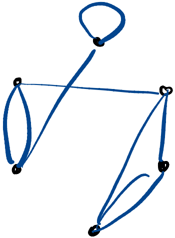


Random d-regular graphs



Random d-regular graphs





Why do we care?

(ii) provide better understanding of underlying mechanisms that create networks.

– Preferential attachment model explains power-law degree distribution ("rich get richer"), 1999

– . . .

Power-law distribution

Real-world networks typically do not have Poisson distribution: think of Instagram with Cristiano Ronaldo and Ariana Grande, having 216M+ and, respectively, 183M+ followers (May 2020).

Power-law distribution

Real-world networks typically do not have Poisson distribution: think of Instagram with Cristiano Ronaldo and Ariana Grande, having 216M+ and, respectively, 183M+ followers (May 2020).

Typically, degree distribution follows power law:

$$d_\ell \approx c \cdot \ell^{-\gamma}$$

for some parameter $\gamma > 0$ (degree exponent) and normalizing constant $c > 0$.

Power-law distribution

Real-world networks typically do not have Poisson distribution: think of Instagram with Cristiano Ronaldo and Ariana Grande, having 216M+ and, respectively, 183M+ followers (May 2020).

Typically, degree distribution follows power law:

$$d_\ell \approx c \cdot \ell^{-\gamma}$$

for some parameter $\gamma > 0$ (degree exponent) and normalizing constant $c > 0$.

First observed by Vilfredo Pareto, a 19th-century economist, who observed that a few wealthy individuals posses the majority of world wealth.

Why do we care?

(ii) provide better understanding of underlying mechanisms that create networks.

– Preferential attachment model explains power-law degree distribution ("rich get richer"), 1999

– . . .

– Protean graphs, 2006

– . . .

– Models of social learning ("homophily and aversion implies segregation")

– . . .

Why do we care?

(ii) provide better understanding of underlying mechanisms that create networks.

– Preferential attachment model explains power-law degree distribution (“rich get richer”), 1999

– . . .

– Protean graphs, 2006

– . . .

– Models of social learning (“homophily and aversion implies segregation”)

– . . .

...but these two reasons are not related to data science.

Why do we care?

(iii) create synthetic networks that closely resemble real-world networks but are flexible so that one can test various scenarios.

Why do we care?

(iii) create synthetic networks that closely resemble real-world networks but are flexible so that one can test various scenarios.

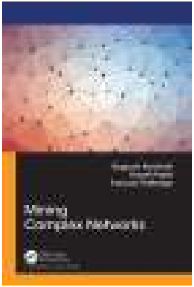
(iv) can be used to benchmark the outcomes of algorithms (for example clustering algorithms); serve as the so-called null-models.

Why do we care?

(iii) create synthetic networks that closely resemble real-world networks but are flexible so that one can test various scenarios.

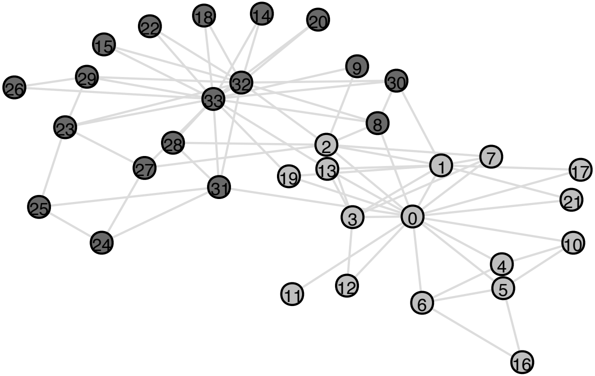
(iv) can be used to benchmark the outcomes of algorithms (for example clustering algorithms); serve as the so-called null-models.

Very active area of research.



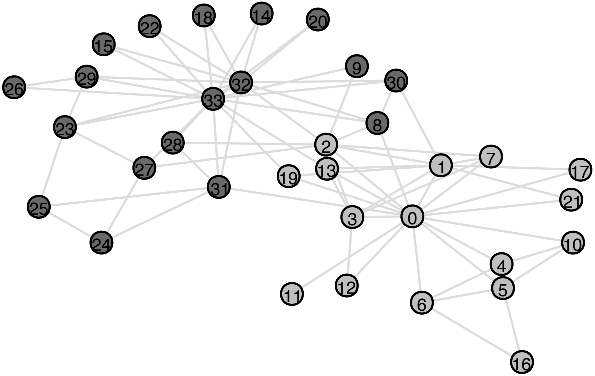
Community Detection

Community Detection — Introduction



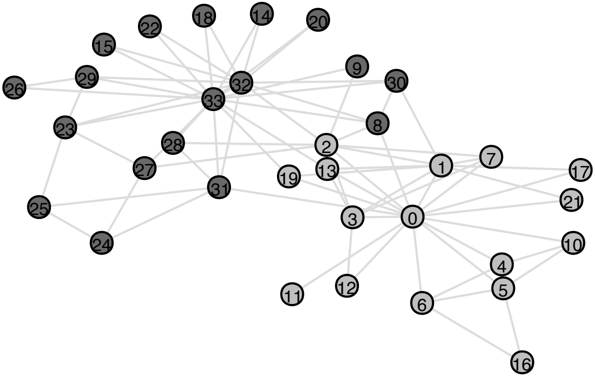
A network has community structure if its set of nodes can be split into a number of subsets such that each subset is densely internally connected.

Community Detection — Introduction



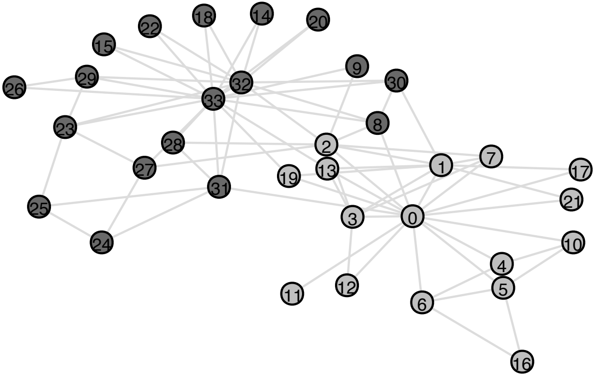
social networks: communities based on common location of their users, their interests, occupation, gender, age, etc.

Community Detection — Introduction



web graph: web pages that belong to the same community are on a similar topic.

Community Detection — Introduction



protein-protein interaction networks: proteins that belong to the same community are often associated with a particular biological function within the organizm.

Finding the right partition that represents the community structure is a challenging but important problem for a number of reasons.

Community Detection — Introduction

Finding the right partition that represents the community structure is a challenging but important problem for a number of reasons.

Communities allow us to...
– see a "big picture" (large scale map with individual communities represented as meta-nodes),

Community Detection — Introduction

Finding the right partition that represents the community structure is a challenging but important problem for a number of reasons.

Communities allow us to...
– see a “big picture” (large scale map with individual communities represented as meta-nodes),
– better understand the function of the system represented by the network,

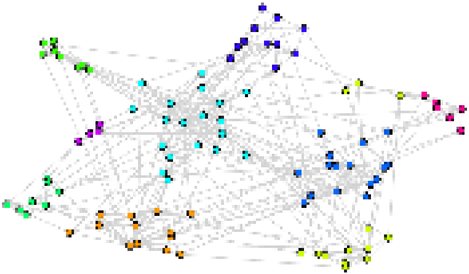
Community Detection — Introduction

Finding the right partition that represents the community structure is a challenging but important problem for a number of reasons.

Communities allow us to...
– see a "big picture" (large scale map with individual communities represented as meta-nodes),
– better understand the function of the system represented by the network,
– classify the nodes based on the position they have in their own clusters and how they are connected to other clusters: roles and importance,
– . . .

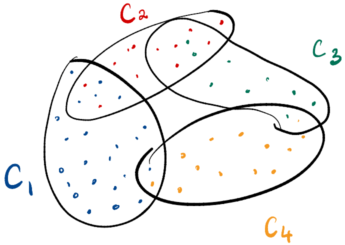
Generating Synthetic Networks

Purpose: testing and tuning unsupervised algorithms (typically the ground truth is not available!).



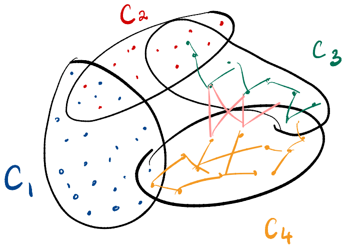
– SBM (Stochastic Block Model),
– LFR,
– ABCD + ABCDe (parallel counterpart),
– New trend: generating synthetic higher-order structures.

Graph Modularity — Definition



Let $\mathscr{A} = \{A_1, A_2, \ldots, A_\ell\}$ be a given partition of V.

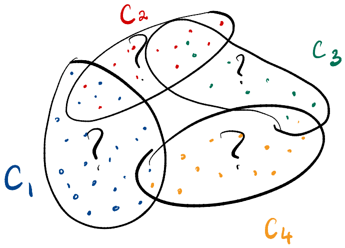
Graph Modularity — Definition



Let $\mathscr{A} = \{A_1, A_2, \ldots, A_\ell\}$ be a given partition of V.

This partition captured $\sum_{A_i \in \mathscr{A}} e_G(A_i)/|E|$ fraction of edges (edge contribution). Should we be happy with this?

Graph Modularity — Definition



Let $\mathscr{A} = \{A_1, A_2, \ldots, A_\ell\}$ be a given partition of V.

This partition captured $\sum_{A_i \in \mathscr{A}} e_G(A_i)/|E|$ fraction of edges (edge contribution). Should we be happy with this? No!

Compare it to the expected fraction of edges captured by this partition in the Chung-Lu model with the (expected) degree distribution $\mathbf{d} = \big(\deg(v_1), \deg(v_2), \ldots, \deg(v_n)\big)$.

Graph Modularity — Definition

The expected fraction of edges is equal to

$$\frac{1}{|E|}\left(\sum_{v_j v_k \in \binom{A_i}{2}} \frac{\deg(v_j)\deg(v_k)}{2|E|} + \sum_{v_j \in A_i} \frac{\deg^2(v_j)}{4|E|}\right)$$

$$= \frac{1}{4|E|^2} \sum_{v_j \in A_i} \sum_{v_k \in A_i} \deg(v_j)\deg(v_k)$$

$$= \frac{1}{4|E|^2}\left(\sum_{v_j \in A_i} \deg(v_j)\right)^2 = \frac{(\mathrm{vol}(A_i))^2}{(\mathrm{vol}(V))^2}.$$

Graph Modularity — Definition

Modularity for graphs is based on the comparison between:

a) the actual density of edges inside a community (the edge contribution), and

b) the expected density if nodes of the graph were wired randomly, regardless of community structure (the degree tax).

Graph Modularity — Definition

Modularity for graphs is based on the comparison between:

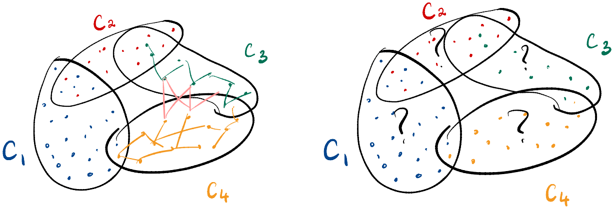
a) the actual density of edges inside a community (the edge contribution), and

b) the expected density if nodes of the graph were wired randomly, regardless of community structure (the degree tax).

Such reference random graph is known in this context as the null-model.

Modularity function:

$$q_G(\mathscr{A}) = \sum_{A_i \in \mathscr{A}} \frac{e_G(A_i)}{|E|} - \sum_{A_i \in \mathscr{A}} \frac{(\mathrm{vol}(A_i))^2}{(\mathrm{vol}(V))^2}$$



Graph Modularity — Definition

Modularity function:

$$q_G(\mathscr{A}) = \sum_{A_i \in \mathscr{A}} \frac{e_G(A_i)}{|E|} - \sum_{A_i \in \mathscr{A}} \frac{(\mathrm{vol}(A_i))^2}{(\mathrm{vol}(V))^2}$$

Some properties:

- $q_G(\mathscr{A}) \leq 1$,
- If $\mathscr{A} = \{V\}$, then $q_G(\mathscr{A}) = 0$,
- If $\mathscr{A} = \{\{v_1\}, \ldots, \{v_n\}\}$, then $q_G(\mathscr{A}) = -\frac{\sum \deg^2(v)}{4|E|^2} < 0$,
- $q_G(\mathscr{A}) \geq -1/2$.

Graph Modularity — Definition

Modularity function:

$$q_G(\mathscr{A}) = \sum_{A_i \in \mathscr{A}} \frac{e_G(A_i)}{|E|} - \sum_{A_i \in \mathscr{A}} \frac{(\mathrm{vol}(A_i))^2}{(\mathrm{vol}(V))^2}$$

Some properties:

– $q_G(\mathscr{A}) \leq 1$,

– If $\mathscr{A} = \{V\}$, then $q_G(\mathscr{A}) = 0$,

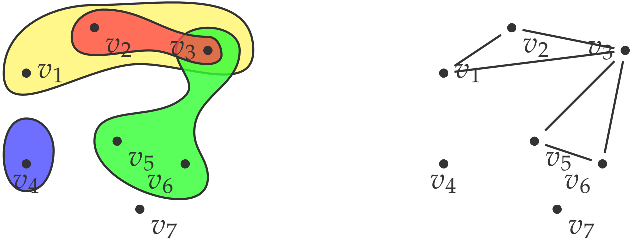
– If $\mathscr{A} = \{\{v_1\}, \ldots, \{v_n\}\}$, then $q_G(\mathscr{A}) = -\frac{\sum \deg^2(v)}{4|E|^2} < 0$,

– $q_G(\mathscr{A}) \geq -1/2$.

$$q^*(G) = \max_{\mathscr{A}} q_G(\mathscr{A})$$

(Well defined but impossible to find in practice.
Used to guide a heuristic algorithms that try to maximize it.)

Hypergraphs



– hypergraphs (left) better represent many complex networks, including social networks,

– unfortunately, there are very few tools and so they are usually reduced to their 2-sections (right),

– but the situation changes: hypergraph modularity function.