Network classification-based structural analysis of real networks and their model-generated counterparts

Marcell Nagy, and Roland Molontay

Department of Stochastics, Budapest University of Technology and Economics, Hungary

Human and Social Data Science Lab, Hungary

Nagy, M., & Molontay, R. (2022). Network classification-based structural analysis of real networks and their model-generated counterparts. *Network Science*, 1-24.

MŰEGYETEM 1782

Human and Social Data Science Lab

- Our mission: translate fundamental research in data and network science into lasting impact in social sciences and humanities
- We conduct both methodology oriented basic research in **data and network science** and applied research with a human-centered and societal focus
- Our research revolves around four main areas:

- **Network science**
- Educational data science
- Health data analytics
- Intelligent monitoring & anomaly detection

- Webpage: `hsdslab.math.bme.hu`

Research objectives

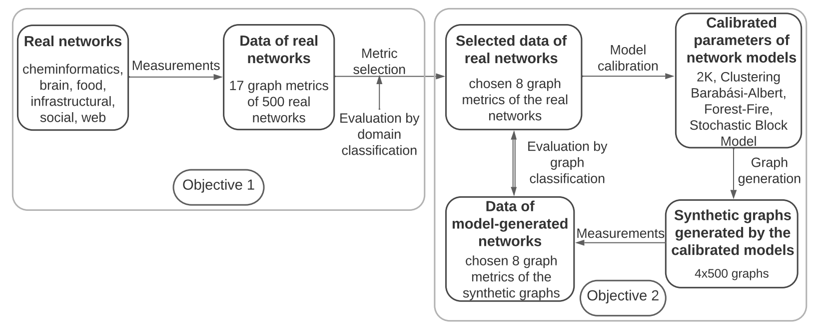
Goal
To **study** and compare the structural properties of **real networks** and synthetic graphs generated by **measurement-calibrated** network **models**

Research objectives

1. Study the **correlation** structure **of** graph **metrics** of real networks
2. Identify a **non-redundant** subset of graph **metrics** that describe real networks well enough
3. Using a **data-driven** approach, **calibrate** the **parameters of** network **models** to mimic real networks as closely as possible

A key contribution
Our proposed framework unifies more branches of data-driven network analysis: **correlation analysis of metrics, network similarity, model calibration**, and **graph classification**.

Workflow



Real networks
cheminformatics, brain, food, infrastructural, social, web
Measurements
Data of real networks
17 graph metrics of 500 real networks
Metric selection
Evaluation by domain classification
Objective 1
Selected data of real networks
chosen 8 graph metrics of the real networks
Model calibration
Calibrated parameters of network models
2K, Clustering Barabási-Albert, Forest-Fire, Stochastic Block Model
Graph generation
Evaluation by graph classification
Synthetic graphs generated by the calibrated models
4x500 graphs
Measurements
Data of model-generated networks
chosen 8 graph metrics of the synthetic graphs
Objective 2

Analysis of real networks

Analysis of real networks

- The study is based on 17 measurements of 500 real networks
- The networks are from 6 domains:
  - **Brain**: human and animal connectomes
  - **Cheminformatics**: protein-protein (enzyme) interaction networks
  - **Food**: what-eats-what, consumer-resource networks
  - **Infrastructure**: Transportation and distribution networks
  - **Social**: online social, co-authorship, and collaboration networks.
  - **Web**: samples of the World Wide Web

Methodology

1. **Collect** real-world networks (Network Repository, ICON, KONECT)
2. **Calculate** graph metrics of the real networks, and **normalize** the size-dependent ones (e.g. maximum degree)
3. **Study** and visualize the **relationship** of **metrics**
4. **Select** a **non-redundant subset** of metrics

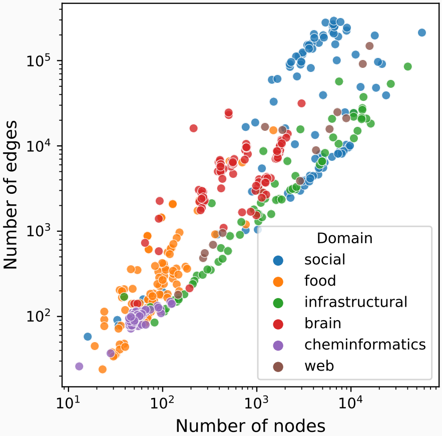
Data of real networks

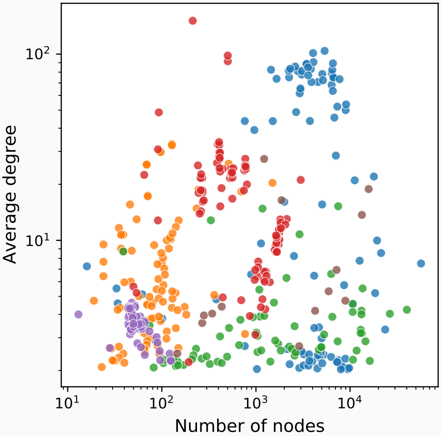
17 graph metrics are chosen that **describe every aspect** of networks:

1. **Degree distribution related**
   e.g., interval degree probabilities (4 metrics), average degree, maximum degree (normalized by the size)
2. **Shortest paths related**
   Diameter and average path length (normalized by the log of the size)
3. **Centrality related**
   e.g. maximum eigenvector and betweenness centrality
4. **Clustering related**
   e.g. Global clustering coefficient, average local clustering coefficient

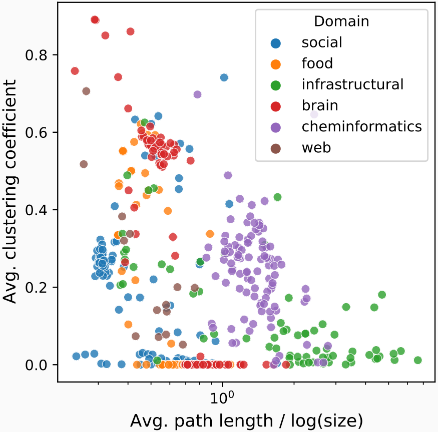
The data of real networks are publicly available on GitHub

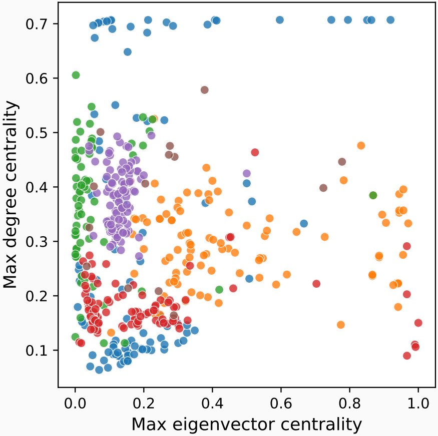
Real networks





Real networks





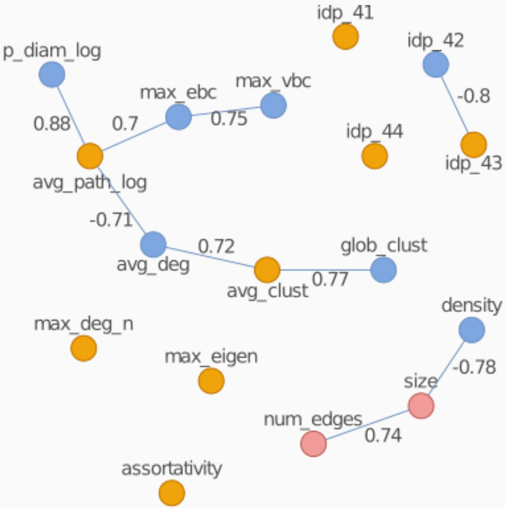
Correlation network and metric selection

Network construction

Two metrics are connected if their **domain-averaged absolute Spearman's** rank **correlation** is greater than 0.65.

Metric selection

1. Exclude the nodes strongly correlated with the number of nodes and edges
2. Select a **maximal independent set** of the remaining vertices

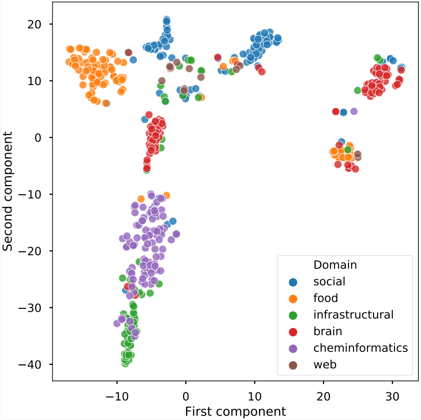


Evaluation of selected metrics

Evaluation of the selection

We evaluated the selected graph metrics by performing **domain classification** using all attributes and the selected ones.

Classifier	All attributes	Selected attributes
kNN	78.5%	87.0%
Decision Tree	85.2%	85.5%
Random Forest	89.3%	91.2%



Analysis of network models

Analysis of network models

Goal

Realistically synthesize networks with given network models

Models

1. 2K model (captures the joint-degree distribution of a graph)
2. Clustering variant of Barabási–Albert model (CBA)
3. Forest-fire model (FF)
4. Stochastic block model (SBM)

Methodology

1. **Calibrate** the parameters of each model for each real network to mimic them as closely as possible
2. **Generate** graphs with the calibrated parameters
3. **Compare** the graph metrics of the synthetic and real networks
4. **Classify** real-word and model-generated graphs.

Model calibration

Notations

- G_T: target real network
- $M(\theta)$: network model with parameter vector θ
- $G_{M(\theta)}$: a graph instance generated by model $M(\theta)$
- d: distance function defined on graphs: Canberra distance of vectors of the chosen metrics

Parameter calibration

For a given target graph G_T and network model $M(\theta)$, find θ^* that minimizes the distance between G_T and $G_{M(\theta)}$, i.e.

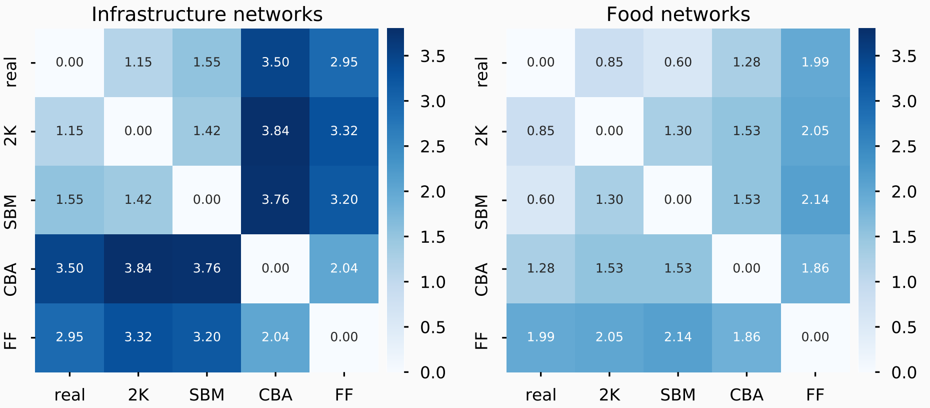
$$\theta^* = \arg\min_{\theta} d\big(G_{M(\theta)},\, G_T\big).$$

The minimization is carried out with **grid search**: for a given $M(\theta)$, where $\theta = (\theta_1, ..., \theta_p)$, we perform a grid search on the p-dimensional state space of the parameters.

Results

Mean Canberra distances

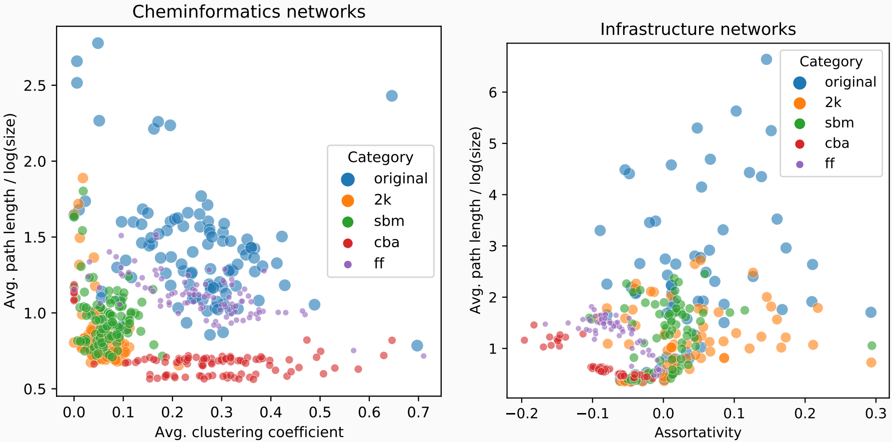
The mean Canberra distances between the original and the synthetic graphs.



Findings

Food networks are **easiest** and **infrastructure networks** are the **most difficult** to mimic for all models.

Properties that models cannot capture

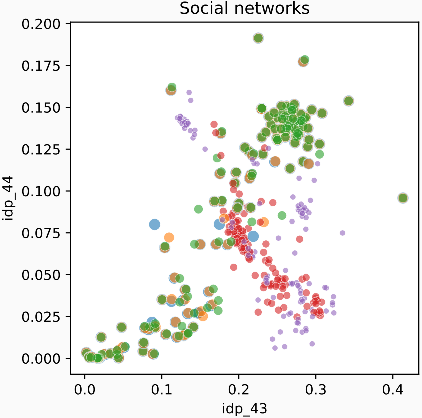


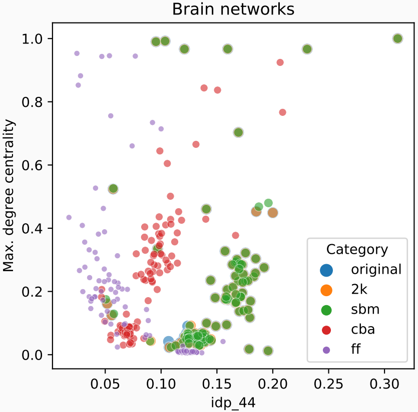
The properties models cannot capture

The distribution and relationship of **avg. path length**, **clustering coefficient**, **max. eigenvector centrality**, and **assortativity**[1] observed in real networks.

[1] Except for the 2K model

Properties that models can capture





Findings
Overall, **2K** model and **SBM** can **accurately** mimic real-world networks.

Summary

Conclusion

Our approach, that unifies correlation analysis, network similarity, graph classification, and model calibration, led us to the following observations:

Findings

- **Models cannot capture** the relationship of **avg. path length**, **max. eigenvector centrality**, and **clustering coefficient** observed in real networks
- **Brain and food** networks are **easier to mimic** (classification accuracy is below 80% baseline) than the others.
- Overall, **2K** and **SBM** can **accurately** mimic real networks

Extension of this work
To compare the descriptive ability of graph metrics to graph embedding methods

Thank you for your attention!

Canberra distance

Definition
Let p and q be two n-dimensional vectors. The Canberra distance d_{Can} between the vectors p and q is defined as follows:

$$d_{\mathrm{Can}}(p, q) = \sum_{i}^{n} \frac{|p_i - q_i|}{|p_i| + |q_i|}$$

Interval degree probabilities

For a given graph G, let μ and σ be the average and the standard deviation of the degrees in G. Then we define the following four intervals $R(i)$, $i = 1, \dots, 4$:

$$R(i) = \begin{cases} [\,\min_v \deg(v),\ \mu - \sigma\,), & \text{if } i = 1 \\ [\,\mu - \sigma,\ \mu\,), & \text{if } i = 2 \\ [\,\mu,\ \mu + \sigma\,), & \text{if } i = 3 \\ [\,\mu + \sigma,\ \max_v \deg(v)\,], & \text{if } i = 4 \end{cases}$$

The interval degree probabilities $IDP(i)$, $i = 1, \dots, 4$ are defined as follows:

$$IDP(i) = \mathbb{P}\big[\deg(v) \in R(i)\big], \quad v \in V(G)$$

References i

Sadegh Aliakbary, Jafar Habibi, and Ali Movaghar.
Quantification and comparison of degree distributions in complex networks.
In *7'th International Symposium on Telecommunications (IST'2014)*, pages 464–469. IEEE, 2014.

Minas Gjoka, Bálint Tillman, and Athina Markopoulou.
Construction of simple graphs with a target joint degree matrix and beyond.
In *Computer Communications (INFOCOM), 2015 IEEE Conference on*, pages 1553–1561. Citeseer, 2015.

Petter Holme and Beom Jun Kim.
Growing scale-free networks with tunable clustering.
Physical Review E, 65(2):026107, 2002.

Jure Leskovec, Jon Kleinberg, and Christos Faloutsos.
Graphs over time: densification laws, shrinking diameters and possible explanations.
In *Proc. of the 11th ACM SIGKDD Int. Conf. on Knowledge Discovery in Data Mining*, pages 177–187. ACM, 2005.

Tiago P Peixoto.
Nonparametric bayesian inference of the microcanonical stochastic block model.
Physical Review E, 95(1):012317, 2017.