Enhancing Space-Aware Community Detection Using Degree Constrained Spatial Null Model

Rémy Cazabet, Pierre Borgnat, Pablo Jensen

Enhancing Space-Aware Community Detection Using Degree Constrained Spatial Null Model

Remy Cazabet[1], Pierre Borgnat[2], and Pablo Jensen[2]

[1] Sorbonne Universites, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, France (remy.cazabet@gmail.com)
[2] Univ Lyon, Ens de Lyon, Univ Claude Bernard, CNRS, Laboratoire de Physique, France

Abstract. Null models have many applications on networks, from testing the significance of observations to the conception of algorithms such as community detection. They ususally preserve some network properties, such as degree distribution. Recently, some null-models have been proposed for spatial networks, and applied to the community detection problem. In this article, we propose a new null-model adapted to spatial networks, that, unlike previous ones, preserves both the spatial structure and the degrees of nodes. We show the efficacy of this null-model in the community detection case both on synthetic and collected networks.

1 Introduction

In recent years, complex networks have become an important topic of research, and are used to model systems and interactions in many different fields, from social sciences to biology.

When elements represented as vertices have a location in space, and the distance between them plays a role, we use *spatial networks* to represent them. Examples of networks modelled by spatial networks include transportation networks, infrastructure networks, mobility networks, or even neural networks. Several models of spatial networks exist, such as random planar graph [1], or generalizations of the Watts-Strogatz model. The distinctive characteristic of spatial network models is that the probability of observing an edge between vertices depends on the distance between them. This characteristic can be represented by a *deterrence function*. For a broad overview of existing work on spatial networks, one can turn to [2].

In complex networks, *null-models* are frequently used to compare the observed properties (assortativity, diffusion, clustering, frequency of patterns, etc.) of a collected network with the ones in a randomized version of it. Another common usage is in community detection, where a quality function called Modularity compares the fraction of edges found inside communities in the observed network and in the corresponding null-model. The most commonly used null model, often called the configuration model (see sec. 2.1), rewires randomly connections between vertices while conserving the degree distribution.

Previously proposed null-models for spatial networks conserve the position of nodes, the deterrence function and the total number of edges, but not the degree distribution. In this article, we propose a null model for spatial networks that preserves as much as possible both the spatial properties and the degrees of nodes.

1.1 Related Works

In [3], the authors study several socio-spatial properties of location-based social networks, such as the average geographical distance between friends or the distribution of social link length. They compare these properties to two randomized version: the *social null-model* conserves the ties, but shuffles the position of users, while the *Geo null-model* uses a probability of friendship $P(d)$ to assign edges randomly between nodes. $P(d)$ is defined as the probability of observing a friendship between individuals situated at a given distance. This model conserves the total number of edges and the deterrence function as much as possible, but not the degree distribution of nodes, which only depends on the distance to other nodes.

In [4], the authors propose a method to find space-independent communities in spatial networks. They successfully uncover a linguistic partition in a Belgian mobile phone calls dataset, that was otherwise hidden by geographical proximities. To do so, they use a modified version of the quality function called Modularity (see sec. 2.3). We will detail this gravity-based null model in section 2.1. They use a modified version of the Louvain algorithm [5] to optimize their variant of modularity. Several articles, for instance [6, 7], applied this approach on different case studies.

In [8], the authors propose to use a mechanism similar than the one in [4], but replace the gravity-based spatial null-model by a radiation-based one. The radiation model has been recently proposed as an alternative to the gravity one, and has attracted a lot of attention since then. Their model is described in section 2.1. They do not use the exact same method than [4] to optimize their quality function, but a a variant of it. They also propose an extension to multiplex and mutislice networks.

Organization of the article: This article is organized as follows: in the first section, we describe the different null-models we will consider, the synthetic benchmarks on which to test them, and describe which method we used for community detection and validation. Section two introduces our new null-model with conservation of degrees. Section three presents the results of the evaluation process. Finally, section four presents an example of application on a collected dataset of shared bicycles.

2 Description of the evaluation settings

2.1 Description of state-of-the-art null-models

Configuration Model The configuration model, or NG model, has been introduced in [9]. It proposes to rewire randomly the graph while keeping the degrees of nodes. More formally, in the case of undirected, weighted networks, the expected weight of a node is $P_{ij}^{NG} = \frac{k_i k_j}{2w}$, with k_i the strength of node i ($k_i = \sum_j W_{ij}$) and $2w$ the sum of all edge weights in the network $2w = \sum_{ij} W_{ij}$.

This null model is not spatial, but is the most commonly used in the definition of modularity.

Gravity-Based This null-model introduced in [4] is based on works coming from the transportation domain, where gravity models have long been used to model the repartition of trips among areas such as cities, countries or neighboroods. It is originally defined in analogy with Newton's law of gravity as $P_{ij} = K\frac{n_i n_j}{d_{ij}^{\sigma}}$, with n_i a notion of the *intrinsic strength* of node i (for instance, depending on application cases, its population, number of jobs, parking lots, etc.), d_{ij} is the distance between nodes i and j, and σ is a parameter allowing to tune the influence of distance.

In recent works, a more general version of the law is often used [10],

$$P_{ij}^{Gra} = n_i n_j f(d_{ij}) \tag{1}$$

with $f(d)$ any deterrence function, and n_i same as before. Instead of being decided *a priori*, the deterrence function can be learned from the data as follows [8]:

$$f(d) = \frac{\sum_{i,j|d_{ij}=d} A_{ij}}{\sum_{i,j|d_{ij}=d} n_i n_j} \tag{2}$$

with A_{ij} the observed flow (number of trips, communications, etc.) between nodes i and j, and d_{ij}, n_i same as in eq.1.

We can note that if the distance has no effect, the deterrence function is a constant function, and the gravity-based model becomes exactly the configuration model.

Radiation-Based Just as the gravity law is an analogy of Newton's law of gravity, the radiation model takes his inspiration form laws of radiation in physics. It has first been introduced in [11], and has been successfully applied in several cases since.. It is defined as:

$$P_{ij}^{Rad} = T_i \frac{n_i n_j}{(n_i + r_{ij})(n_i + n_j + r_{ij})} \tag{3}$$

With $r_{ij} = q_{ij} - (n_i + n_j)$, q_{ij} being the sum of n_k for all k in the circle of center i and radius d_{ij} (population closer from i than j). Other notations identical to eq. 8.

A particularity of this model is that it does not need an explicit deterrence function, as the interactions between nodes depends on their *intrinsic strength* and of the presence of other nodes around them.

To be able to tune the importance of distance, however, the variant of the radiation model introduced in [8] adds a deterrence function effect learned from data, identical to the one previously introduced. The Distance Tuned radiation model becomes:

$$P_{ij}^{DTRad} = P_{ij}^{rad} f(d_{ij}) \tag{4}$$

with $f(d_{ij})$ a deterrence function defined as in the gravity-based case.

2.2 Synthetic benchmarks for space-corrected community detection

To compare the results obtained by using different null-models, the most straightforward solution is to use synthetic benchmarks mixing a community structure and a spatial structure.

The benchmark introduced in [4] in a gravity-based version and extended in [8] to a radiation process generates a network with both a planted community structure and a spatial structure. Its distinctive feature is that all edges probabilities have to respect the spatial structure. Compared with the version presented in [8], we introduce two minor modifications:

- We generalize it in order to allow any deterrence function
- We allow the gravity version to handle variable intrinsic weights

The generic test benchmark is defined as:

$$p_{ij}^{Inc} = \lambda(c_i, c_j) P_{ij}^{SNM}(f(d_{ij})) Z_1 \tag{5}$$

with c_i the community containing node i, the function $\lambda(c_i, c_j) = 1$ if nodes i and j are in the same community, and $\lambda(c_i, c_j) = \lambda_d$ otherwise, $P_{ij}^{SNM}(f(d))$ a probability given by the chosen spatial null model with deterrence function $f(d)$, and Z_1 a normalization constant ensuring that $\Sigma_{i>j} p_{ij}^{Inc} = 1$.

Given the parameters:

- $N \in \mathbb{Z}, N > 0$: number of nodes
- $C \in \mathbb{Z}, C > 0$: number of communities
- $l \in \mathbb{N}, l > 0$: length of the sides of the considered, square 2-dimensional space
- $\mu \in \mathbb{N}, \mu > 0$: graph's density
- $\lambda_d \in [0, 1]$: mixing coefficient
- $f(d)$: deterrence function
- $I_{min}, I_{max} \in \mathbb{Z}, I_{max} \geq I_{min} > 0$: minimum and maximum intrinsic strengths

We generate graphs according to the following procedure:

1. Attribute a position to each of the N nodes in space, defined uniformly at random such that $n_x \in [0, l], n_y \in [0, l]$
2. Attribute an *intrinsic strength* to each node, uniformly at random such that $n_I \in [I_{min}, I_{max}]$
3. Attribute a community to each node, taken uniformly at random in the set $\{1, .., C\}$

4. Compute p_{ij}^{Inc} for all i,j, for the chosen λ_d,P_{ij}^{SNM},$f(d)$
5. Distribute uniformly at random $L = \mu N(N-1)/2$ edges, where there is an edge between i and j with probability p_{ij}^{Inc}, and multiple edges are interpreted as weights.

2.3 Community detection algorithm

The community detection procedure we use is identical to the one in [4]. The idea is to use the Louvain algorithm [5], one of the most widely used algorithms for community detection. The principle of this algorithm is to optimize a quality function called modularity, using a fast greedy approach. Modularity is a quality function of a network partition, that compares the fraction of internal edges in a given partition and in a null-model. In its original definition, modularity uses the configuration model as null-model. Here, we replace it by the null-model to test. The generic version of modularity can be expressed as [8]:

$$Q = \frac{1}{2w} \Sigma_{ij} (W_{ij} - P_{ij}^{NM}) \delta(c_i, c_j) \tag{6}$$

with $2w = \Sigma_{ij} W_{ij}$ the total edge weight, c_i the community of node i, δ the Kronecker delta, and P_{ij}^{NM} the ij-th element of the null model matrix.

2.4 Community partition evaluation

For each set of benchmark's parameters to test, a graph is generated, and communities are found for each tested null-model using modified Louvain. It then becomes possible to compare the detected partition, result of the algorithm, with the planted partition. As in previous works [4,8], we use the Normalized Mutual Information (NMI) [12], an information-theoretic similarity mesure that gives a score between 0 and 1, 1 meaning that the two partitions are identical, and 0 meaning that there is no more correlation between them than expected by chance.

3 Definition of a Degree Constrained Gravity-Based model

In the previously introduced spatial null models, there is no simple relation between the *intrinsic strength* of a node and its actual strength (sum of weights of adjacent edges) in a network generated according to this null model. This means that if the only available data is an observed network, and we use observed degrees of nodes as a proxy for their intrinsic importance, then any of the previously proposed spatial null model fitted on this observed network will not conserve the degrees of nodes. The null model we propose is searching for a degree constrained solution, i.e a spatial null-model preserving the degrees of nodes.

To do so, we take inspiration from the doubly constrained gravity model [13], and adapt it to the case of spatial networks with estimated deterrence function. The intuition is that we are searching for values of intrinsic strength that would best explain the observed degrees. We present the method in its more general form, adapted to oriented weighted networks. Therefore, we compute separately for each node an *Incoming estimated Intrinsic trength* (n^{Ieis}) and an *Outgoing estimated Intrinsic Strength* (n^{Oeis}). For non-oriented networks, $n^{Ieis} = n^{Oeis}$.

The method consists in iteratively estimating the new values for n^{Ieis} and n^{Oeis} that satisfies the observed indegrees (deg^{in}) and outdegrees (deg^{out}) constraints.

We can define them recusively as :

$$n^{Ieis} = \frac{deg^{out}(i)}{\sum_i n^{Oeis} f(d_{ij}))}, n^{Oeis} = \frac{deg^{in}(i)}{\sum_i n^{Ieis} f(d_{ij})} \tag{7}$$

And the corresponding Degree Constrained gravity model is:

$$P_{ij}^{DCgrav} = n^{Oeis} n^{Ieis} f(d_{ij}) \tag{8}$$

Starting with initial values $n^{Oeis} = deg^{out}$ and $n^{Ieis} = deg^{in}$, we first compute all values for n^{Oeis}, then all values for n^{Ieis}, and so on and so forth until the degrees obtained in the gravity model defined in Eq. 8 are close enough to the target network. Although this process is known to converge [13], in this article we will use a fix number of iterations, $i = 5$, to avoid discussions on stopping criterium and convergence time.

3.1 Recomputation of the deterrence function

Because the computed deterrence function depends on the *intrinsic strength* of nodes, estimating it using observed degrees as a proxy leads to a biased approximation. By recomputing the deterrence function after each iteration of the algorithm, we can in part correct this bias.

3.2 Summary of the proposed method

The complete process for constructing a spatial null model can be summarized as follows:

1. Initialise n^{Ieis} and n^{Oeis} with nodes out and in degrees
2. Compute the deterrence function according to Eq. 2
3. Update all n^{Oeis} according to Eq. 7
4. Update all n^{Ieis} according to Eq. 7
5. If stopping criterium is not reached, return to step 2)
6. Compute all P_{ij}^{DCgrav} according to the gravity model defined in Eq. 8

4 Validation of null models on synthetic benchmarks

To test which null model is better suited to discover communities in which case, we generate networks according to the proposed benchmark with different parameters, search for communities using the same algorithm with different null-models, and compare discovered communities with planted partitions known by construction. In this section, we first describe the sets of parameters considered, secondly, we describe the community detection process and its evaluation, before studying the results.

4.1 Benchmark parameters

To limit the number of cases to study, we decided to fix some parameters. The influence of these parameters has already been studied in [8], and minor changes do not affect much the results. Of course, major changes can have strong effect, for instance is the graph becomes extremely sparse, finding communities becomes harder for all methods.

We choose values close to the ones studied in [8]. Fixed parameters and their values: $N = 100$, $l = 10$, $\mu = 100$, $I_{min} = 10, I_{max} = 100$, $C = 2$

For the deterrence function, whose impact was not studied in [8], we consider several values: For gravity based benchmarks, we take $f(d)$ among $\{f(x) = 1/x, f(x) = 1/x^{0.5}, f(x) = x^2\}$. For the Radiaton case, we consider $f(d) \in \{f(x) = 1, f(x) = 1/x\}$. $f(x) = 1$ corresponds to the original definition of the Radiation model, with no explicit definition of deterrence function.

As in [8], we allow the mixing parameter λ_d to vary from 0 to 1, i.e from perfectly unambiguous community structure to a network with only a spatial structure.

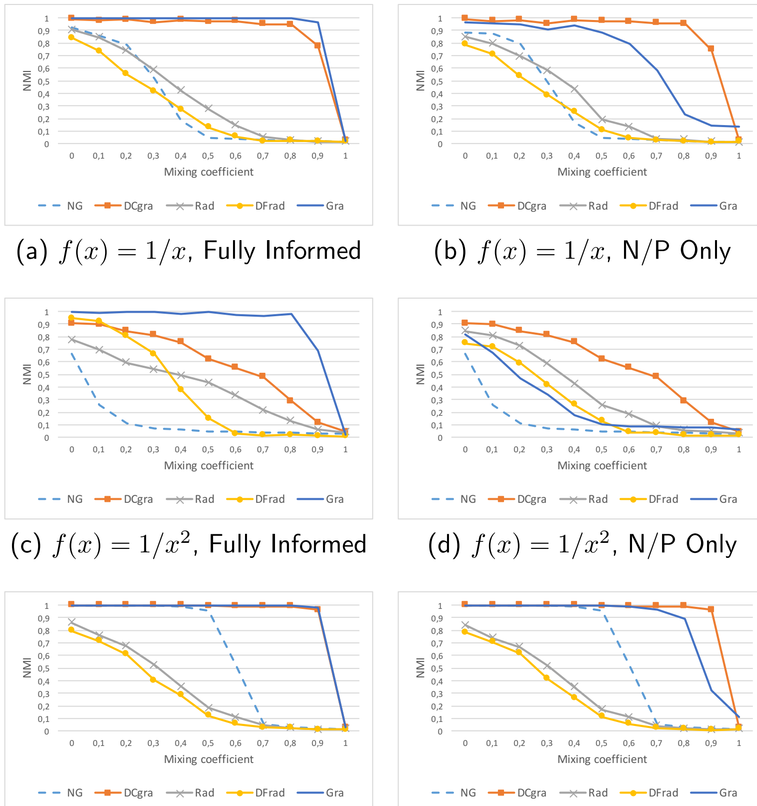
4.2 Evaluation process

For each set of parameters, we generate 50 instances of networks. For each instance, we run the modified Louvain algorithm with each of the following null models:

- Configuration model [14], noted as NG
- Gravity-based [4], noted as Gra
- Radiation-based (original) [11], noted as Rad
- Radiation-based with deterrence function [8], noted as $DFrad$
- Degree constrained Gravity-based, introduced in the present paper, noted as $DCgra$

For each of these algorithms, we consider two cases, Fully Informed and Network/Position Only.

In the **Fully Informed** version, we consider that we know not only the observed network, but also the intrinsic strength of nodes and the deterrence function used to generate the network. This is the same setting than tests conduced in [4, 8].



(a) $f(x) = 1/x$, Fully Informed (b) $f(x) = 1/x$, N/P Only

(c) $f(x) = 1/x^2$, Fully Informed (d) $f(x) = 1/x^2$, N/P Only

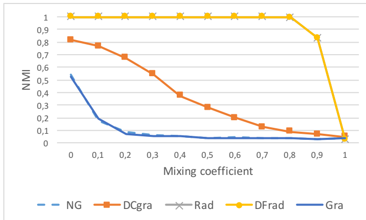
(e) $f(x) = 1/x^{0.5}$, Fully Informed (f) $f(x) = 1/x^{0.5}$, N/P Only

Fig. 1. Results for the synthetic benchmark, using a **generative Gravity model**. In fully informed cases, the gravity null-model is the most efficient, while the proposed DCgra model gives best results when only the network and position of nodes is known.

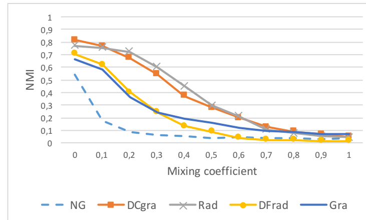
In the **Network/Position Only** version, we consider that we only know the observed network, and the position of nodes. The deterrence function is first computed from these data, when needed, and the degree of nodes is used as proxy for the intrinsic importance of nodes, as it is often done in applications to collected datasets, for instance in [7, 8]. This setting is more realistic, for applications to real world datasets.

4.3 Results

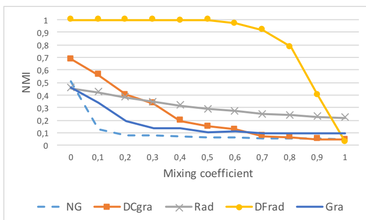
In Fig. 1, left column, we present the results for the synthetic benchmark with a generative gravity model, and the fully informed case. As expected, the *Gra* null-model is the most efficient. We can observe that the problem becomes harder with the increase in the exponent of the deterrence function. In fact, the more



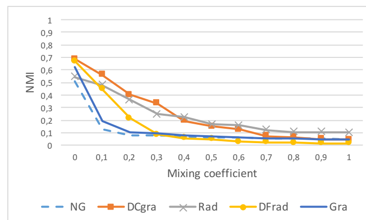
(a) $f(x) = 1$, Fully informed



(b) $f(x) = 1$, N/P Only



(c) $f(x) = 1/x^2$, Fully informed



(d) $f(x) = 1/x^2$, N/P Only

Fig. 2. Results on the synthetic benchmark, using a **Radiation generative model**, both for **Fully Informed** and **Network/Position Only** settings. While the DFrad model gives by far the best results in fully informed cases, its efficacy dwindle when less information is available, and the DCgra model and the original Radiation Null-models give the best results.

this exponent is low, the more the network resemble a non-spatial network. The proposed DCgravity model, that does not benefit from full information, comes nevertheless second in most settings.

In fig. 1, right column, tests are conduced with same settings but in Network/Position Only version, i.e similar to a collected dataset. In this configuration, results for the original gravity model dwindle, in particular with a high exponent for the deterrence function, in which cases the radiation models give better results. The DCgravity algorithm gives best results in most settings.

In Fig. 2, a radiation generative model is used. With the function $f(x) = 1$, both Rad and DFrad give similar result, because this function is implicitly assumed by the Rad null model. Although they reach high NMI scores in Full Information settings, again the results shrink in the N/P only case, in particular for DFrad. With a modified deterrence function, DFrad is the only one to give good results on Fully informed settings, but again, it does not maintain this efficiency for N/P Only. Interestingly, results for DCgra and Rad are comparable in the N/P Only cases.

In conclusion, according to these benchmarks, Gra and DFrad are the best methods to use when one already knows the nature of the underlying spatial process (gravity or radiation), the real intrinsic strengh of nodes and the deterrence function. If only the observed network is available, if the underlying process is gravity based, then the DCgra null model gives the best results in most cases, and it also gives results comparable to the Rad null-model when the underlying process is radiation-based.

4.4 Application to a Bike Sharing network

We use a BSS dataset presented in [15] and that we provide in open access [3]. It is composed of all bicycle trips done in 2011 in the city of Lyon, France, using the bicycle-sharing system called Velo'v. Each trip has a specific origin from a static station in the city, and a destination station which can be any other. It can be studied as a network as in [16], where each station corresponds to a node. The weight on edge (i, j) corresponds to the number of trips done from i to j. The network consists of more than 6 Million trips and 343 nodes (stations). We use the great circle distance between stations to learn the deterrence function, although the difference with euclidian distance is negligeable for such short distances.

In fig. 3, we can see communities discovered by the different proposed null-model. Radiation-based null-models (results similar for both) apparently fail on this dataset, probably because very few trips are actullaly observed between close stations (under a km), contrary to the assumption of radiation model, for which closest nodes necessarily have the strongest bonds. The NG null model find relevant communities, corresponding to the spatial organisation of the city. With the gravity and DCgravity, some non spatially constrained communities are discovered, in particular ones corresponding to patterns typical of leisure or convenient trips, through the bicycle friendly banks of the river and between city parks. Although there is no argument to say that one partition is better than the other in absence of any reference, DCgravity communities match even more the rivers' banks than the simple gravity one, as highlighted in Fig. 4.

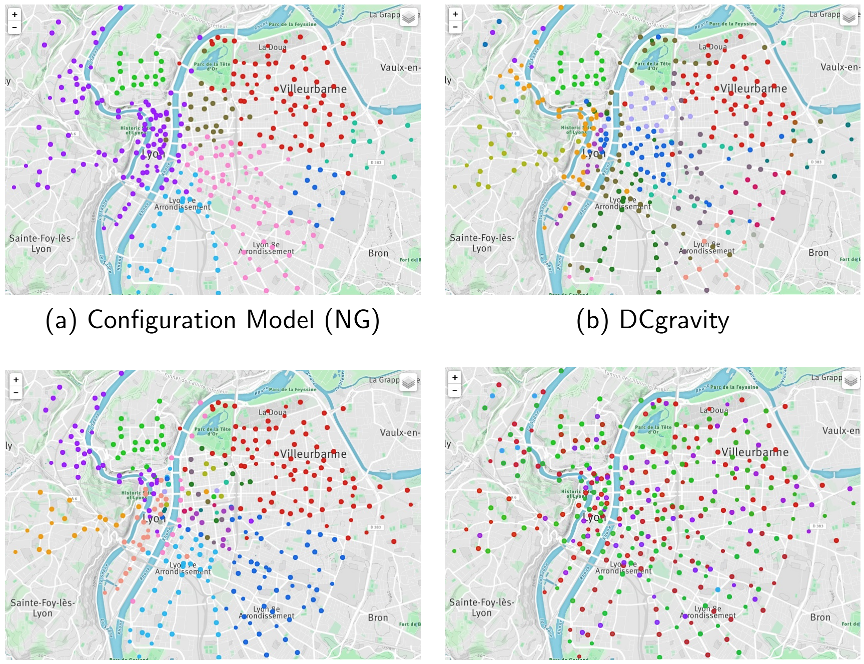
5 Conclusion

In this article, we proposed a null model for spatial networks that conserves both the spatial structure of a network and the degrees of nodes. We have shown on synthetic benchmarks that it was more efficient that non degree-constrained versions to discover community structures hidden in a spatial network. Results on a real dataset confirmed that the approach was relevant, and succeeded to discover meaningful communities while methods based on radiation approach failed.

In future works, we want to explore different usages of such a DC null-model. First, it could be used not only on spatial data, but also on any kind of data in which a distance between node properties could be computed, such as a difference in age, altitute, income, or even temporal data.

---

[3] `https://figshare.com/articles/Lyon_s_BSS_2011/4257128`, the authors thank JCDecaux (Cyclocity) for having provided access to the Velo'v dataset



(a) Configuration Model (NG) (b) DCgravity

(c) Gravity (d) Radiation

Fig. 3. Communities found on the Lyon BSS dataset, using different null models.

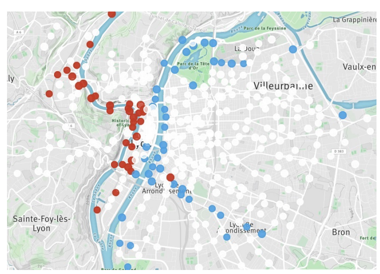


Fig. 4. Details of the two communities discovered using DCgravity null-model that correspond to enjoyable/convenient trips in the city, that were hidden by the influence of space proximity.

We could also investigate other usages besides community detection: null models are used as references for properties such as clustering coefficient, motif frequencies, or, more straightforwardly, to discover the most significant edges and nodes in a network.

References

1. A. Denise, M. Vasconcellos, and D. J. Welsh, “The random planar graph,” *Congressus numerantium*, pp. 61–80, 1996.
2. M. Barthélemy, “Spatial networks,” *Physics Reports*, vol. 499, no. 1, pp. 1–101, 2011.
3. S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo, “Socio-spatial properties of online location-based social networks.” *ICWSM*, vol. 11, pp. 329–336, 2011.
4. P. Expert, T. Evans, V. Blondel, and R. Lambiotte, “Uncovering space-independent communities in spatial networks,” *Proceedings of the National Academy of Sciences*, vol. 108, no. 19, pp. 7663–7668, 2011.
5. V. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.
6. Y. Liu, Z. Sui, C. Kang, and Y. Gao, “Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data,” *PloS one*, vol. 9, no. 1, p. e86026, 2014.
7. M. Z. Austwick, O. OBrien, E. Strano, and M. Viana, “The structure of spatial networks and communities in bicycle sharing systems,” *PloS one*, vol. 8, no. 9, p. e74685, 2013.
8. M. Sarzynska, E. A. Leicht, G. Chowell, and M. A. Porter, “Null models for community detection in spatially embedded, temporal networks,” *Journal of Complex Networks*, p. cnv027, 2015.
9. M. E. Newman, “Modularity and community structure in networks,” *Proceedings of the national academy of sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.
10. M. Lenormand, A. Bassolas, and J. J. Ramasco, “Systematic comparison of trip distribution laws and models,” *Journal of Transport Geography*, vol. 51, pp. 158–169, 2016.
11. F. Simini, M. C. González, A. Maritan, and A.-L. Barabási, “A universal model for mobility and migration patterns,” *Nature*, vol. 484, no. 7392, pp. 96–100, 2012.
12. A. Strehl and J. Ghosh, “Cluster ensembles—a knowledge reuse framework for combining multiple partitions,” *Journal of machine learning research*, vol. 3, no. Dec, pp. 583–617, 2002.
13. I. Williams, “A comparison of some calibration techniques for doubly constrained models with an exponential cost function,” *Transportation Research*, vol. 10, no. 2, pp. 91–104, 1976.
14. M. E. Newman, S. H. Strogatz, and D. J. Watts, “Random graphs with arbitrary degree distributions and their applications,” *Physical review E*, vol. 64, no. 2, p. 026118, 2001.
15. M. Vogel, R. Hamon, G. Lozenguez, L. Merchez, P. Abry, J. Barnier, P. Borgnat, P. Flandrin, I. Mallon, and C. Robardet, “From bicycle sharing system movements to users: a typology of vélo’v cyclists in lyon based on large-scale behavioural dataset,” *Journal of Transport Geography*, vol. 41, pp. 280–291, 2014.
16. P. Borgnat, C. Robardet, P. Abry, P. Flandrin, J.-B. Rouquier, and N. Tremblay, “A dynamical network view of lyon’s velo’v shared bicycle system,” in *Dynamics On and Of Complex Networks, Volume 2.* Springer, 2013, pp. 267–284.