Supplementary Information:
Measuring the effect of node aggregation on community detection

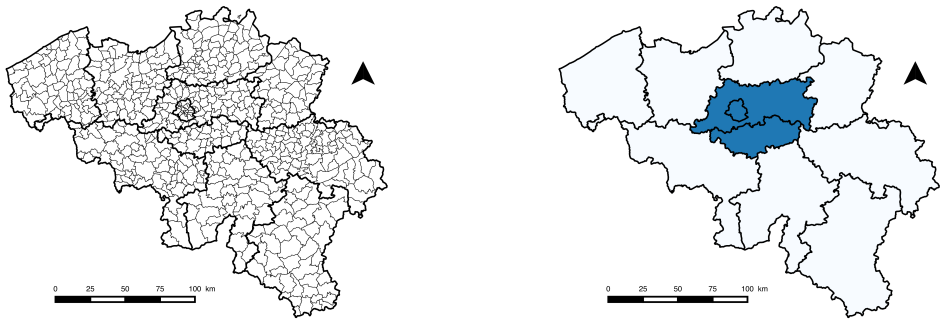
Yérali Gandica[1,2,*], Adeline Decuyper[1], Christophe Cloquet[3], Isabelle Thomas[1] and

Jean-Charles Delvenne[1,2]

1 Center for Operations Research and Econometrics (CORE), Université Catholique de Louvain, Louvain-la-Neuve, Belgium.
2 Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM), Université Catholique de Louvain, Louvain-la-Neuve, Belgium
3 Poppy, Rue Van Bortonne, 7, 1090 Jette, Belgium.

SA.1. Description of the territory

The Belgian territory extends over 30,528 square kilometers, and counts 11 million inhabitants. Belgium is divided into 10 provinces + 1 capital Region. Historically, Belgium was divided into 2,675 municipalities (hereafter referred to as "former municipalities"), that were merged in the year 1979 into the 589 current municipalities (Figure 1a). The capital city, Brussels, is surrounded by the former province of Brabant, now split into two provinces.



(a) Belgium (b) Former province of Brabant

Figure 1: Areas covered by the two datasets: Twitter data are collected over the whole Belgian territory, mobile phone data counts phone calls between towers within the former province of Brabant (blue area).

SA.2. Study of the aggregability index on synthetic data

We analyse the behaviour of the aggregability index, defined in Eq. 2 of the main text, on synthetic data.

We start with a graph of 4 disconnected cliques of 160 nodes each (640 nodes in total), creating a perfect partition into 4 communities. Each clique (160 nodes) is divided into 4 subsets of 40 nodes, arbitrarily. These 4 subsets are the aggregation classes. Thus the aggregability index is $\eta = 1$, as each of the 16 aggregation class is embedded into a community.

We then gradually rewire the links so that the 4 cliques start to mix with each other. In this way, the community structure gradually ceases to be the union of aggregation classes, which are always fixed. (The codes we used are available at `github.com/yerali/aggregability_index_on_synthetic_data`. The black curve of Fig. 2 shows the aggregability index in terms of the rewiring probability, where the decay of the index is clear for rewiring probability on the range of $(0.5 - 0.55)$. It is within that range where the community structure becomes chiefly determined by the random fluctuations of densities created by the rewiring process and loses any alignment with the four initial cliques. Thus, the aggregation classes are no longer subsets of the communities.

Let us now analyse the robustness of the index against heterogeneity on the density of interactions (number of links). For that purpose, we repeat the experiment on an initial network that is union of 4 cliques of *different* sizes (instead of 160 nodes per clique, in the homogeneous case). For comparative purposes, the networks will always have 640 nodes. Each clique is then subdivided into arbitrary aggregation classes of 40 nodes.

In the same Fig. 2 we show the curve of aggregability index as a function of the rewiring for different sizes of the initial cliques. The network with original cliques of 40, 40, 40, and 520, shows the earliest decay of the aggregability to the rewiring. This suggests that the heterogeneous communities are less robust to rewiring than the homogeneous communities, thus leading to a quick disalignment with the original cliques and aggregation classes.

SA.3. Methodology

Details of how both datasets were collected

The phone dataset was collected from 13/04/2015 to 24/05/2015 over the towers located the region of Brabant in Belgium. The dataset contains the number of phone calls whose origin and

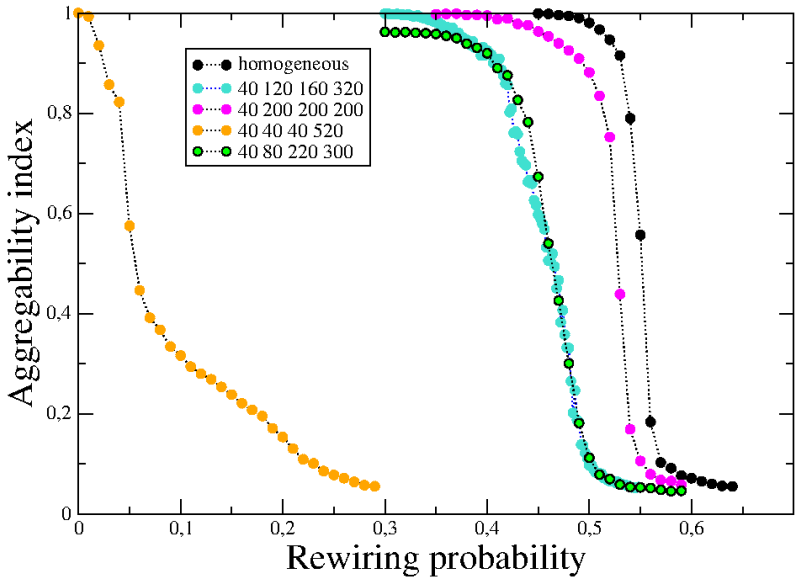


Figure 2: **Aggregability index on synthetic data**. Aggregability index vs. the rewiring probability on the graphs before calculating community detection, for five different distribution of nodes into cliques for the original network. Before rewiring, the graphs have 4 disconnected cliques (which we take as 4 unambiguous communities). For the case of a homogeneous density, each clique has 160 nodes and is divided into 4 aggregation classes of 40 nodes each. For the case of heterogeneous densities, the number of nodes in each clique (initial community) varies in order to test different posibilities. The number of nodes in each aggregation class is kept to 40 and the total number of nodes in preserved to 640. As the rewiring process is pursued, the initial connectivity is altered, allowing to create connection between different cliques. The communities of the network and the aggregability index are recomputed for each step of the rewiring process. Communities are computed as maximising modularity ($\rho = 1$), as estimated with the Louvain method. For heterogeneous sizes of the cliques, we see that the aggregability index curve drops earlier in the rewiring process than in the homogeneous case.

destination are towers within the territory of Brabant, and where both the caller and the callee are clients of the phone company providing the data.

We use Twitter feeds that consist in 140-character messages publicly broadcasted with a latitude and longitude. We have selected the tweets geotagged with coordinates within the boundaries of Belgium, used as reply-to messages to specific users. This allows to define a user-to-user social network, where nodes are users and undirected edges are weighted by the number of exchanged tweets. A GPS coordinate is attached to every user, which is the barycenter of the GPS coordinates of the tweets issued by the user, making it a place-to-place, geographic network as well.

To collect it, we used the Twitter STREAMING API through the Twython module [1]. The collection resulted in an average of 0.59 tweets per second. Twitter is known to enforce rate

limitations on its APIs, to keep the acquired tweets rate below 1 % of the total tweets streams. This may result in bias in the collected data (see for instance [2, 3]), especially during the peak hours. The researchers have no control over this. However, contrarily to [2, 3], we requested a much smaller sample of the Twitter stream and we collected about 0.59 tweets per second, representing only 0.6 % of the total 9,100 tweets per second (58M/day) [4]. We collected 8,730,973 geo-tagged tweets from 143,314 users between 29/10/2013 and 18/2/2014, for which the WGS-84 geographical coordinates fell into the rectangle of Figure 1a.

We removed tweets from some known automatic accounts, such as Touring Mobilis, helping locating traffic events, that automatically locates tweets at the place that the tweet is about, although it does not represent an actual user tweeting from these places. Among the remaining tweets, 291,552 from 18,327 users were reply-to messages sent to other users from the collected databases. These 291,552 tweets form our database of edges.

As a word of warning, we show in Fig. 3 the density of users for each of the 579 municipalities of Belgium, showing large discrepancies across the territory.

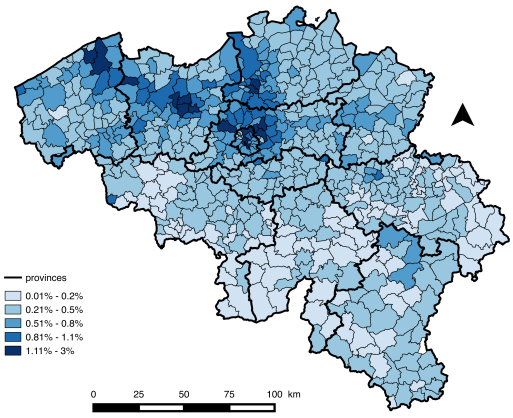


Figure 3: Number of Twitter users per inhabitant in each municipality.

SA.4. Tables

The networks derived from both datasets following the procedure explained in the main text are described in terms of number o f nodes and edges in Tables 1 and 2.

network	aggregation level	# nodes	# edges
N_0	users	15,453	41,326
N_m	muni.	1,569	22,711
N_{fm}	former muni.	534	12,729
N_{125}	125m	14,593	41,212
N_{250}	250m	13,136	41,051
N_{500}	500m	10,342	40,371
N_{1k}	1km	6,795	37,493
N_{2k}	2km	3,335	30,610
N_{4k}	4km	1,281	20,161
N_{8k}	8km	420	8,911
N_{16k}	16km	134	2,294
N_{32k}	32km	42	433

Table 1: Characteristics of the Twitter networks analyzed.

network	aggregation level	# nodes	# edges
M_0	towers	1,168	367,388
M_m	muni.	111	5,722
M_{125}	125m	1,119	356,815
M_{250}	250m	1,012	320,769
M_{500}	500m	877	266,163
M_{1k}	1km	656	157,073
M_{2k}	2km	433	68,544
M_{4k}	4km	201	16,241
M_{8k}	8km	62	1,777
M_{16k}	16km	20	204
M_{32k}	32km	8	35

Table 2: Characteristics of the mobile phone networks analyzed.

References

[1] Twython Module. Https://github.com/ryanmcgrath/twython

[2] Gonzalez-Bailon, S., Wang, N., Rivero, A., Borge-Holthoefer, J., Moreno, Y.: Assessing the bias in samples of large online networks. Social Networks **38**, 16–27 (2014)

[3] Morstatter, F., Pfeffer, J., Liu, H.: When is it biased?: assessing the representativeness of twitter's streaming api. Proceedings of the 23rd International Conference on World Wide Web, 555–556 (2014)

[4] Twitter Statistics. Http://www.statisticbrain.com/twitter-statistics