



**HAL**  
open science

## Graphlet-based characterization of many ego networks

Raphaël Charbey

► **To cite this version:**

| Raphaël Charbey. Graphlet-based characterization of many ego networks. 2018. hal-01764253v1

**HAL Id: hal-01764253**

**<https://hal.science/hal-01764253v1>**

Preprint submitted on 11 Apr 2018 (v1), last revised 26 Nov 2020 (v3)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Graphlet-based characterization of many ego networks

Raphaël Charbey\* and Christophe Prieur\*

April 11, 2018

## Abstract

Network science gathers methods coming from various disciplines which sometimes hardly cross the boundaries between these disciplines. Widely used in molecular biology in the study of protein interaction networks, the enumeration, in a network, of all possible subgraphs of a limited size (usually around five nodes), often called graphlets, can only be found in a few works dealing with social networks. In the present work, we apply this approach to an original corpus of about 10 000 Facebook ego networks gathered from voluntary participants by a survey application. We define here an original measure that we call graphlet representativity, with which we produce a clustering of graphlets into five groups (paths, star-like, holes, light triangles, dense), along with an original visualization scheme for the comparison of ego networks. We then use the visualization in order to compare several clusterings of our corpus of networks, using various state-of-the art metrics. The graphlet representativity ends up producing the most discriminative clustering, so we describe the distinct structural characteristics of the five clusters of ego networks so obtained, and discuss the differences between 4-node and 5-node graphlets. We also provide many suggestions of followups of this work, both in sociology and in network science.

## 1 Introduction

For nearly twenty years, there has been a broad interest on networks across many fields of research, gathered into the terms of complex networks (Newman, 2003) or network science (Brandes et al., 2013). This trend comes after a long tradition of research in social sciences, traced back to the early 20th century with G. Simmel's social circles (Simmel, 1908), then J.-L. Moreno's sociograms (Moreno, 1934). After a frenzy of activity of anthropologists in the 1950's among which J. Barnes who coined the expression *social networks* (Barnes, 1954), an extensive quantitative framework has been formalized in the 1970's around H. White. Many accounts can be found of the formation of this field known as social network analysis (SNA) (Scott, 2017; Freeman, 2004).

---

\*Département Sciences Économiques et Sociales, Télécom Paristech, Paris, France. raphael.charbey / christophe.prieur@telecom-paristech.fr

A new boom of interest has come after the discovery and formalization of non-trivial properties that are shared by large networks coming from very diverse contexts (Faloutsos et al., 1999; Barabási and Albert, 1999; Watts and Strogatz, 1998; Newman, 2003). To study these so-called complex, small-world, networks, new methods and tools have been designed, coming from statistical physics, computer science, applied mathematics, computational biology, economics, etc. (Bornholdt and Schuster, 2006; Easley and Kleinberg, 2010), sometimes with little (if any) reference to previous works from now classical social network analysis.

Along these two paths of research on networks (namely social network analysis and the complex networks field), some methodological approaches may produce sophisticated developments which remain quite concurrent and separate according to the disciplines in which they are used and spread. An example of this phenomenon is provided by structural analyses based on the enumeration of elementary structures. Triads, *i.e.* triples of entities and the possible ties between them, have been extensively used since the 1970's as a methodological tool in social network analysis (Holland and Leinhardt (1976); see for instance Faust (2010) for an overview). Decades later, this trend has led to a very active subfield on a statistical method called *ergm*, for exponential random graph models, in which specific (small) patterns are enumerated in the studied networks, and the result compared to an expected value, given a specific type of probabilistic model of random networks (Robins et al., 2007). Meanwhile in the study of biological networks (protein-protein interaction networks for instance), a now widely used method based on the seminal work of Milo et al (Milo, 2002) is to enumerate so-called *motifs*, which are small structures overrepresented in comparison to random networks with similar properties. The two approaches are thus very similar but both the references, the exact methods and the terms used are widely distinct.

In the motifs trend in biological networks, many works have been published during the last fifteen years in computational biology, in computer science and in statistics on the question of enumerating what has been called *graphlets*, small networks of size up to 5 or 6. Algorithms have been devised to do it efficiently (Wernicke, 2006), including the enumeration of all possible positions of nodes in the graphlets (Stoica and Prieur, 2009; Hocevar and Demsar, 2014; Hočevar and Demšar, 2017), sometimes called *orbits*, and measures have been defined to use the results of the enumeration to compare networks of various origins (Pržulj, 2007; Yaveroğlu et al., 2015a). As suggested above, the motif or graphlet framework, mostly used on biological networks, has had a very low impact on the field of social network analysis, where the *ergm* method is prominent. On the specific case of ego networks, a key topic in social network analysis where the focus is made on individuals and their relational environment, only a few works have been done using the graphlet approach (Stoica and Prieur, 2009; Cunningham et al., 2013).

Dating back to the classical work of E. Bott on the network structure of households (Bott, 1957), the study of personal networks, or ego-centered networks, or ego networks, has long been an important trend to address questions ranging from homophily to agency, life change or social capital (see for instance Wellman (2007) for a survey). The development of online social network sites during the mid-2000's has come partly as a consequence of both the small world narrative and this ego-centered perspective in network analysis, reinforcing the relevance of this approach while bringing new ways to adopt it. Studying ego networks extracted from Facebook has then been a very stimulating perspective to study tie strength

(Backstrom and Kleinberg, 2014), cohesion (Friggeri et al., 2011), social capital (Brooks et al., 2014) or online ties and the uses of Facebook itself (Spiliotopoulos and Oakley, 2013; Park et al., 2012)... until Facebook decided to remove access to the users' ego networks for third-party applications as of May 2015 (Hogan, 2018; Nasim et al., 2016).

In the present work, we apply the graphlet approach to a corpus of about 10 000 Facebook ego networks. They have been gathered by a survey application using the platform's programming interface (API) which collected participants' data. Rather than using graphlets to categorize networks of different types (such as biological, social, or transportation networks for instance), we describe an original research framework enabling to compare similar networks to exhibit fine-grained structural properties. Without relying on random networks, we have designed a so-called *graphlet representativity* measure which provides insights for the analysis of specific individuals or groups of networks in the corpus, and also seems to outperform existing measures to split our corpus of networks into structurally significant clusters. For each network of the corpus, our indicator compares the proportion of its graphlets with the proportions in the whole corpus. The same measure is used also to cluster the graphlets themselves, providing reading guides for the analysis of the corpus. The outcome is twofold, in network science and in sociology. Indeed it carries on the graphlet approach to exhibit a rich variety of structural properties in corpuses of similar networks, which may occur in many contexts (biology, sociology, etc.). In the sociological study of personal networks in particular, our work brings new methodological tools and clues for future research on structural properties of networks of relationships.

The paper is structured as follows. In Section 2, we recall previous works on graphlet counting and graphlet-based distances between networks, and present the analytical method that we have developed and the data we will use. We define the graphlet representativity and show the distribution of graphlets among our corpus that constitute the reference values for the representativity measure. In Section 3 we look at the similarities between the different graphlets of size 5 within our dataset, providing a clustering of graphlets based on these. We then consider, in Section 4, clustering the Facebook ego networks and exhibit five clusters of ego networks which provide significant structural characteristics. Finally, in Section 5, we discuss the interest of counting the graphlets only up to the size 4. In the conclusion, we provide many suggestions of followups of this work, both in sociology and in network science.

## 2 Framework

### 2.1 Graphlets

A network is usually modeled as a *graph*, which is a pair  $(V, E)$  where  $V$  is a set of *vertices* (or *nodes*) and  $E$  a set of *edges* (or *links*, or *ties*), pairs of vertices defining a binary relation between them. In the present paper, we only consider undirected graphs, *i.e.* where the relation  $E$  is symmetric. A graph is said to be *connected* if for any pair of its nodes, there exists a sequence of edges that form a path from one node to the other. Following Pržulj et al. (2004), we call *graphlets of size  $k$*  the collection of connected graphs with  $k$  vertices.

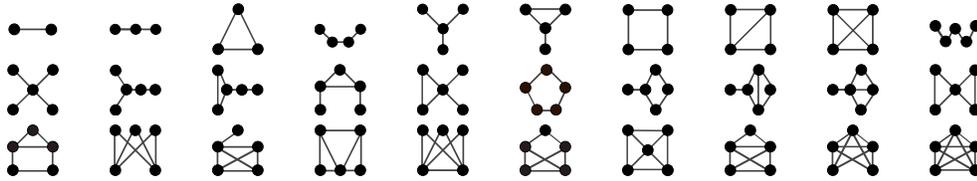


Figure 1: The 30 graphlets of size five or less.

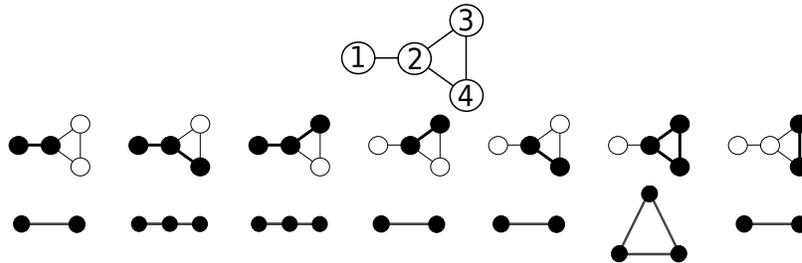


Figure 2: Example of graphlet counting up to size 3 for a small graph (on top). The middle line shows the induced subgraphs as they are visited. These subgraphs are identified to their corresponding graphlets (bottom line). All connected induced subgraphs of size less than 4 have been visited during the process. Notice that, in particular, the 2-path  $\bullet\text{---}\bullet$  is only visited twice even though it is a subgraph of the right-most portion of the graph (vertices 2, 3, 4), but the subgraph *induced* by  $\{2, 3, 4\}$  is the triangle graphlet  $\triangle$ .

The list of all graphlets of size up to 5 is presented in Figure 1. In order to avoid too heavy notations, we consider in the whole paper that the set of graphlets is given as a numbered sequence, so we denote a graphlet by its number.

Given a graph  $G = (V, E)$  and a subset  $V'$  of  $V$ , a graph  $G' = (V', E')$  with  $V' \subseteq V$  and  $E' \subseteq E$  is called a *subgraph* of  $G$ . It is said to be *induced by*  $V'$  if  $E'$  is obtained by removing from  $E$  all the edges containing at least one vertex outside  $V'$ . The enumeration of graphlets within a graph (a network) consists in visiting (and counting) all of its connected induced subgraphs. An example for 2- and 3-node graphlets is illustrated in Figure 2. In practice, the maximum number of nodes within the combinations is usually of 3 for directed networks (Holland and Leinhardt, 1976; Milo, 2002) and 4 or 5 (Ali et al., 2014; Pržulj et al., 2004; Yaveroglu et al., 2015a), due to combinatorial limitations. Indeed, both the number of different graphlets and the time of computation increase exponentially with the value of  $k$ .

Given a graph  $G$  and an integer  $i \in \{2, 30\}$ , the *relative frequency* of the  $i$ -th graphlet is the ratio  $N_i(G)/T(G)$ , where  $N_i(G)$  is the number of occurrences of the graphlet within  $G$  and  $T(G) = \sum_{i=2}^{30} N_i(G)$  its total amount of graphlets. N. Pržulj and her colleagues have defined this relative frequency to design a graphlet-based distance metric for networks that

they call the *relative graphlet frequency distance* (Pržulj et al., 2004). In his seminal paper, Milo uses the very similar notion of *concentration*, on directed graphs (Milo, 2002). To tone down the difference of importance between graphlets which occur a lot in real-world networks (like the fork  or the star-like  for instance) and unusual graphlets (like many with a hole , a logarithmic function is applied to reduce the amplitude of the proportions. Formally, the distance between two graphs  $G_1$  and  $G_2$  is defined as:

$$D(G_1, G_2) = \sum_{i=2}^{30} |F_i(G_1) - F_i(G_2)|, \quad \text{where } F_i(G) = -\log(N_i(G)/T(G)).$$

They used it to compare protein-protein interaction networks with graphs generated according to four different random graph models.

With graphlets comes the notion of *orbits* (once again using the term proposed by Pržulj et al., Pržulj (2007)). The *orbits* of a graphlet are the equivalence classes defined on its vertices by the automorphic equivalence. Two vertices of a graph are *automorphic* if they cannot be distinguished from each other. For instance, in the star , the four nodes at the end of the branches are in the same orbit while the central node is the only one in its orbit. Another example: all the cliques     have only one orbit each. Pržulj et al. base their work on the 73 different orbits included in the 30 graphlets up to size 5.

In a more recent development, the same team has defined two other distances based on orbits. The so-called *Graphlet Degree Distribution Agreement* (Pržulj, 2007) compares networks depending on the distribution of what they call the graphlet degrees of their nodes, which is a generalization of the degree. Without getting into too technical details, let us just describe the *graphlet degree* of a vertex as the vector of the numbers of its appearances in each of the 73 orbits (note that the traditional degree, *i.e.* the number of edges incident to a given vertex, is exactly the number of its appearances in the only orbit of the simple edge graphlet ). In Yaveroğlu et al. (2015a), from all the graphlet degrees of a graph, they consider, for each pair of orbits, a correlation value, which defines the so-called *Graphlet Correlation Matrix* of this graph. The *Graphlet Correlation Distance* of two graphs is then the Euclidean distance between the upper triangles of their respective matrices. They also define a variant where they remove a few orbits they show as being combinatorially redundant. They showed the efficiency of this method by distinguishing clusters of networks according to their origins (social networks, protein interaction networks, trade networks, various types of random networks, etc.).

Concurrently, other researchers have defined a function, called *NetDis*, to compare two graphs based on the count of the graphlets in ego networks extracted from each of the two graphs. Ego networks in this case are induced by 2-neighborhoods: for a given vertex  $v$ , its 2-neighborhood is the set of vertices which can be reached from  $v$  by a path of length at most 2. The graphlet count is then compared to an expectation which is computed from a reference input graph. The deviation from this reference is the basis for the comparison between the two graphs Ali et al. (2014). Yaveroğlu et al. argue that the Graphlet Correlation Distance gives better results (Yaveroğlu et al., 2015b).

Like Ali et al. (Ali et al., 2014), we choose not to have a concomitant use of graphlets of

different sizes. Indeed, there are much more networks of higher size in big networks which makes uncertain how to compare counts of graphlets not having the same combinatorial constraints.

From an algorithmic point of view, graphlet counting is very expensive in terms of time of computation and thus different approaches have been proposed. Some algorithms are designed to exactly compute the number of appearances of each graphlet (Wernicke, 2006; Stoica and Prieur, 2009; Hocevar and Demsar, 2014; Pinar et al., 2017) while others propose sampling strategies to have an approximate value for each graphlet (Wernicke and Rasche, 2006; Zhao et al., 2012).

## 2.2 Graphlet representativity

Most of these methods aim at finding similarities between networks coming from empirical data and randomly generated networks. They may even be used to discriminate networks coming from various fields (Yaveroğlu et al., 2015a). Unlike these, the graphlet-based metric that we propose here, that we call the *graphlet representativity*, is intended to be used with (many) networks coming from one homogeneous corpus. For instance a set of ego networks from a sociological survey, a set of protein-protein interaction networks, or of urban street networks of various cities. The input is then a set of graphs  $\mathcal{C} = \{G_1, \dots, G_{|\mathcal{C}|}\}$ .

We call *global (graphlet) frequency* of the  $j$ -th graphlet the ratio of its total number of appearances among all the graphs of  $\mathcal{C}$ , over the sum of appearances of all the graphlets of the same size in all the graphs. The idea of the graphlet representativity is then to normalize the relative graphlet frequency according to this global frequency, with an additional center cut operation that we describe now.

Let  $r_{i,j}$  and  $R_j$  denote respectively the relative frequency of the graphlet  $j$  within the graph  $G_i$ , and its global frequency. The ratio  $\frac{r_{i,j}}{R_j}$  indicates whether the graphlet is over- or under-represented in  $G_i$ . However, it is either between 0 and 1 in the latter case, or greater than 1 in the former. We thus normalize that value so that it fits between 0 and 2. The *representativity*  $\rho_{i,j}$  of the graphlet  $j$  within the graph  $i$  is defined as follows:

$$\rho_{i,j} = \begin{cases} \frac{r_{i,j}}{R_j} & \text{if this value is under 1} \\ 2 - \frac{R_j}{r_{i,j}} & \text{else.} \end{cases}$$

Figure 3 shows the values computed on a sample ego network from our corpus (see below for the description of the dataset).

We will also use an extended definition of the representativity applied to sets of graphs included in the main corpus. Let  $\mathcal{C}'$  be such a subset of  $\mathcal{C}$ , then we define  $r_{\mathcal{C}',j}$  as the *relative frequency in  $\mathcal{C}'$* , which is the ratio between the number of occurrences of the  $j^{\text{th}}$  graphlet, and the total number of graphlets in  $\mathcal{C}'$ . This value is actually the relative frequency (as defined in the previous section) of a graph made of the union of the graphs in  $\mathcal{C}'$ . Likewise,  $\rho_{\mathcal{C}',j}$  computed from  $r_{\mathcal{C}',j}$  and  $R_j$  like previously, is the representativity of the  $j^{\text{th}}$  graphlet in  $\mathcal{C}'$ .

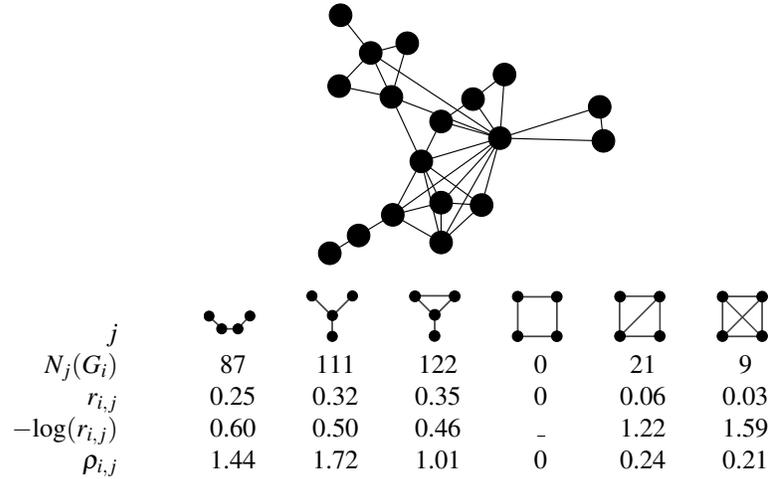


Figure 3: Values of relative graphlet frequency (raw and with the log function) and representativity for the size-4 graphlets of an example graph  $G_i$ . The global frequency, which is not given here, is computed on the corpus described in the next section.

The notion of representativity we just defined is close to the approach taken by NetDis (Ali et al., 2014) since it compares the relative frequency of graphlets in several graphs with an expectation value. The main difference is that we do not invoke, as they do, a graph that would be considered as a reference, but we use the corpus of graphs itself to produce the reference score. Moreover, the similarities between graphlet representativities of a graph might be used to produce a representative vector in a similar way as the one used by Yaveroğlu *et al.* with the orbits (Yaveroğlu et al., 2015a).

We can think of some limitations of representativity. First, a natural consequence of its very purpose is that network representativities cannot be compared to each other if they have not been extracted from the same dataset. Now, focusing on a sub-corpus, every graphlet representativity converges to 1 (the null value) as the portion of the whole corpus in this sub-corpus grows, which means that it should not be too big in order to carry significant information.

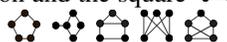
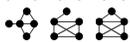
## 2.3 Data

We have applied the graphlet representativity to a corpus of ego networks extracted from 2013 to 2015 by an application designed for a survey on the uses of Facebook<sup>1</sup>. Besides a few questions, this application was asking for respondents' agreement to collect their

<sup>1</sup>Dataset available on request (and shall be put on a publicly available website before publication of the paper). The data collection process was designed in close relationship with the national public organization in charge of the privacy policies.

(anonymized) data on Facebook (including the ties between friends, information which was available via the API until April 2015). As an incentive for their participation, the application provided, among other utilities, an interactive visualization tool showing their network of friends. Before accessing this tool, they had also to answer a few questions about their relationship to five of the friends among the most active on their Facebook page (or 'wall'). They could (and some did) answer these relationship questions about more friends, by clicking on them on the map of their network. More than 16,000 people participated in the survey, among which more than 800 from a controlled representative sample recruited by a poll institute (as the application spread word-of-mouth on the web mostly through a technophile academic environment, young urban males are overrepresented among the respondents).

For each respondent (that we call *ego*), the ego network so retrieved is composed of ego's Facebook friends (that we call *alters*), along with the 'Facebook-friendship' ties between them, except for the alters who have opted out from their default visibility to third-party applications (in our data these appear to have no ties at all). For computational reasons, in the present work, we have restricted the corpus to networks according to their size. We have kept two corpuses: one with  $N_5 = 3,694$  networks with less than 150 nodes for counting size-5 graphlets, one with  $N_4 = 10,252$  with less than 350 nodes for size-4 graphlets. Only networks with at least 15 links have been considered. Figure 4 shows the global graphlet frequency computed on these two corpuses.

The first thing one can see is that the values vary a lot. To ease the description, we will use descriptive nicknames for some graphlets or groups of graphlets. The most represented graphlets of size 4 are those we call the hanger  and the path  while the clique  is less common and the square  very unusual. So are the 5-node graphlets with one or several holes  and conversely, the fork , the star-like  and some others, quite dense , are frequent.

It is interesting to note that the most common graphlets are not necessarily those with the most obvious structure, like the path , the star , the cycle  or the clique , which would have been easier to interpret and compute. This confirms the interest of an exhaustive counting of the graphlets.

### 3 Clustering the graphlets

Once computed the graphlet representativity of the ego networks in our corpus, before using it to analyze the networks, we first study the relationships between the graphlets themselves (focusing on the size 5 in the present section). Identifying correlations should help reducing the complexity of an analysis of the networks.

Moreover, some graphlets can be seen as witnesses of clear social interactions between the alters (ego's friends) while for others it may be much more complicated. This suggests to create clusters of graphlets which would be more easily interpretable than isolated graphlets. For example, the clique  is the marker of a group of classmates, coworkers, family

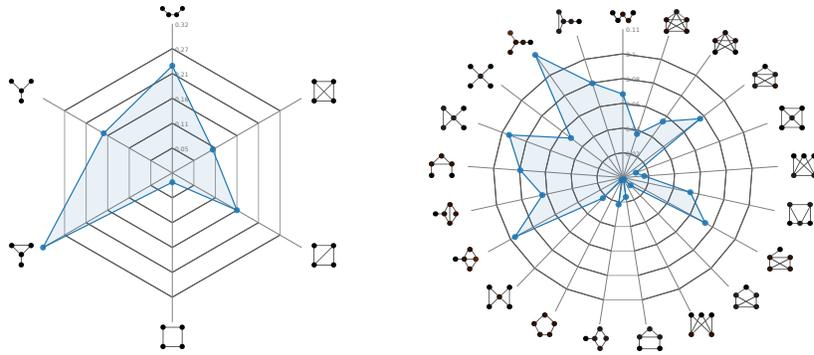


Figure 4: Relative frequencies of graphlets of 4 nodes at left, and 5 nodes at right of our corpus of personal networks. These constitute the reference values for the computation of the graphlet representativity

members, usually related to bonding ties. The central orbit of the star  indicates a special connection between ego and this alter, which is in relation with four other alters that do not know each other (try to think about who is in this position in your personal network and you will probably find your beloved one, closest friends or parents). Conversely, some other graphlets seem to be more difficult to read during the process of interpreting social relationships through networks, and there are several possible explanations. For example, some seem too similar to others to bring a significant difference in terms of interpretation: the 4-clique plus 1 , very close to the 5-clique , or the star-like  to the star .

We use the *k-means* algorithm. This standard method aims at finding clusters of variables, that is, groups of variables which are closer to each other than to the ones from the other clusters. In our case, the variables representing the graphlets are composed of 3,694 values, the representativity values of the graphlets in each network. Thus, if two graphlets have close representativities for an important number of networks, they will be part of the same cluster of graphlets.

Since the *k-means* algorithm is not deterministic, and the number *k* of clusters has to be set as a parameter, we run 100 iterations for each value of *k* between 4 and 6. To choose the final clustering for each *k*, we use the *silhouette*, an indicator of the consistency of a cluster. The resulting clusters are presented in Figure 5.

The two upper clusters seem very strong, since they do not change with the value of *k*. The others are more mixed together. In the 4-clustering, the last cluster contains nearly all the graphlets having at least two triangles, which gathers very different structures (the clique  and the bowtie  for instance). These are dispatched when *k* grows. On the contrary, in the 6-clustering, one cluster contains only one graphlet  whose structure is not easily interpretable or distinguishable from others.

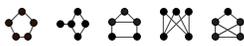
We have thus finally chosen to use the following 5 clusters to analyze our networks:

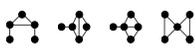
4-clustering	5-clustering	6-clustering

Figure 5: Clusters of graphlets obtained from k-means for  $k = 4, 5, 6$

**the paths**  This cluster contains the graphlets with long paths and very few induced cycle (only one triangle in the case of the hanger ). All of them have the 4-path  twice as an induced subgraph which is also the only sub-graph that they have in common.

**The star-likes**  These two graphlets have a central node tied with all the other four nodes. Having the same property, the bowtie  is however separated from them. This may indicate that being a bridge between three or more nodes (and possibly cohesive groups of nodes) is quite different from linking only two. Indeed, the two graphlets of this cluster both contain the 3-star  which is not the case of the bowtie .

**the holes**  These are the graphlets containing cycles without a chord. It is interesting to note that the 5-cycle  is close to the others while it has no combinatorial dependence from the square , that is included in all the other graphlets of this cluster. This suggests that most of the networks having five-node cycles also have squares.

**the light triangles**  These patterns contain triangles but do not have an important density. Underlining that, we notice that none of them contains more than twice the triangle , and thus, more than once the diamond , which is the second densest graphlet of size four.

**the dense**  This cluster gathers the graphlets with the highest densities, even though three of them, the closed envelope , the Sydney opera  or the pyramid  do include subgraphs having a low density.

As we have underlined here, except for the 5-cycle , all clusters of graphlets in our corpus are in relationship with a shared sub-graphlet. This suggests that they are likely to be robust on different corpuses.

## 4 Clustering the networks

Clustering data is a good way to have an idea of how it is distributed within more or less similar groups and it also enables to compare these groups according to other variables. While the distances proposed in the literature and mentioned above successfully separate groups of networks of different types, we have to deal here with a continuum of networks. We will thus have to keep in mind that a clustering algorithm will necessarily result in quite arbitrary frontiers between groups.

Before discussing a clustering method, we present the visualization technique we use to help interpret the clusters. Figure 6 shows the relative graphlet frequency of five clusters represented on a radar chart where the graphlets are put on the angle axis, grouped according

to the graphlet clustering built in the previous section to limit the visual bias attached to this graphic representation. The figure shows two versions of the frequency defined by Pržulj *et al.* (Pržulj *et al.*, 2004): raw and log-normalized. In both cases, the clusters look very similar. This can be easily explained by the comparison to the relative graphlet frequency of the whole corpus, represented in Figure 4: as already mentioned, the frequency values are very heterogeneous over all the graphlets. This is precisely why we have defined the graphlet representativity, which we will thus use to represent the clusters, in Figure 7 as well as in the forthcoming radar charts. In this visualization, the reference value, defined as 1 for the representativity, is shown by a black solid 'circle'. A dot outside this circle means the corresponding graphlet is overrepresented, a dot inside shows an underrepresentativity.

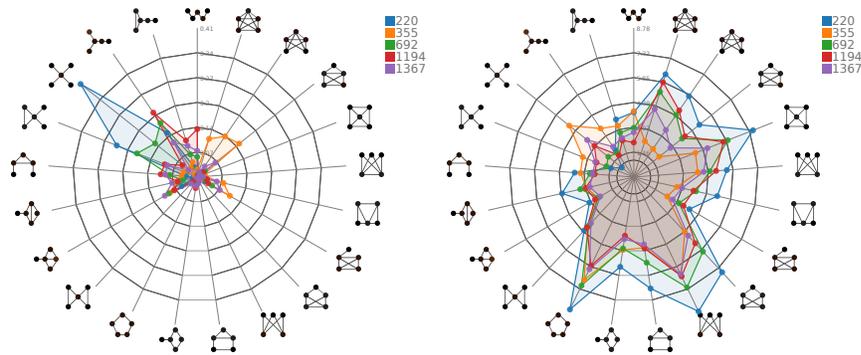


Figure 6: Displaying clusters using the relative graphlet frequency : un-normalized on the left, log-normalized on the right.

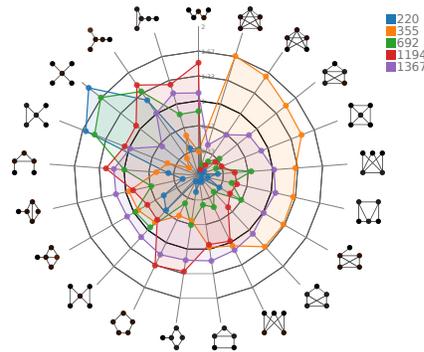
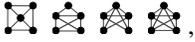
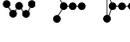


Figure 7: Displaying clusters using the graphlet representativity (clusters obtained from the relative graphlet frequency).

## 4.1 Clustering using state-of-the-art metrics

The clusters represented in Figures 6 and 7 have been obtained with a k-means performed on the values of the relative graphlet frequency. They put into evidence mainly three distinctive types of networks, with overrepresentation of dense graphlets (orange cluster, with 355 networks), of star-like graphlets (blue and green clusters, 220 and 692 networks respectively<sup>2</sup>), or paths graphlets (red cluster, 1194). However the distribution of the networks among these clusters is very heterogeneous: the largest cluster (purple, 1367), gathering more than one third of the whole corpus, gives very poor interpretative elements, apart from the under-representativity of the clique  and of the star . The unclear distinction between the blue (220) and green (692) cluster is also a bad point of this clustering.

We then discuss here other clusterings, based on various distances mentioned above. First, we have performed an aggregative clustering (the one implemented in Python's sklearn library) on distance matrices with both the relative graphlet frequency distance (Pržulj et al., 2004) or the graphlet degree distribution agreement (Pržulj, 2007). The two resulting clusterings were unsatisfactory: representativity values very close to 1, heterogeneous distribution of networks among the clusters, and no more precise information than on the previous clustering.

Aggregative clustering based on the Graphlet Correlation Distance (Yaveroğlu et al., 2015a), represented on Figure 8, is quite better. Two clusters (blue with 524 networks and red with 130) show high values of representativity for the densest graphlets , and low values (very low, in the case of the red 130-cluster) for the paths  and for the star-like graphlets . On the contrary, one (orange, 1710) cluster has high representativity values for paths and star-like, while low values for the dense graphlets. Another tendency is represented by a (purple, 487) cluster, with particularly high values for two graphlets containing a hole , and for the paths graphlets , while very low values for the dense graphlets.

This clustering does not exhibit a wide variety of cases: two clusters (blue, 524 and red, 130) are strongly similar, a third one (orange, 1710) has quite opposite values. Moreover, one cluster (green, 843) does not provide really useful representativity values (close to the null circle on the figure).

We also have tried other clustering methods (k-means and k-medoids) on the distance matrices for the various graphlet distances considered here, with no success. Let us recall that these metrics have proven to be very efficient to discriminate networks coming from different sources. Here, on the contrary, we work on a corpus of very similar networks, since they come from the same source (Facebook ego networks).

---

<sup>2</sup>In the wole of the paper we will refer to the clusters using both their colors and their sizes, to optimize the readability for either color-aware or color-blind readers

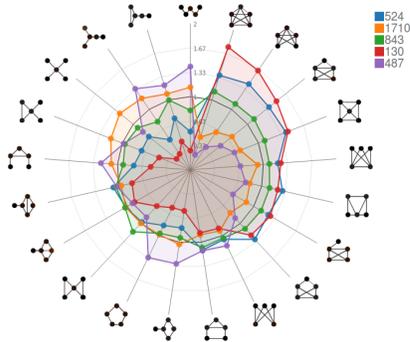


Figure 8: The clusters obtained from the graphlet correlation matrix with 11 orbits.

## 4.2 Clustering using representativity

The same way we have used the representativity to easily visualize the results of graphlet counting, we now present how it can be useful to compute better clusterings of the networks of our corpus. Let us however note that an aggregative clustering did not produce better results than previously presented. In this case, we have computed the distance matrix with the Euclidean distance between the representativity vectors of the networks.

Now the representativity values for the graphlets of size 5 make 21-dimension vectors for each network. We clustered them using the k-means algorithm, with the same 100-runs silhouette method as in Section 3 for clustering the graphlets, with tested values for  $k$ , the number of clusters, ranging from 3 to 10.

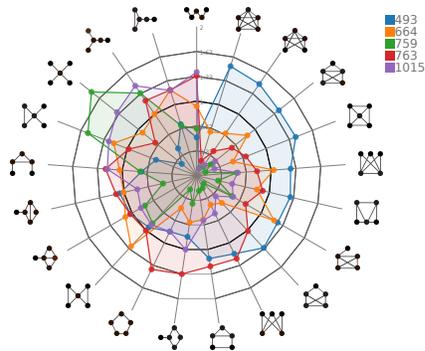


Figure 9: The clusters obtained from the graphlet representativity

In the resulting clustering, depicted in Figure 9, all clusters are significant and show properties of their own, while the networks of the corpus are rather well distributed. This clustering is thus more satisfying than in the previous attempts. This section is devoted to a description of the characteristics of the five clusters, along with a visualization of a typical example of

Table 1: Structural measures per cluster

	dense	bowties	stars	holes	paths
density	0.18	0.12	0.08	0.11	0.07
clustering coefficient	0.7	0.64	0.46	0.52	0.47
diameter	5.63	6.26	6.3	6.47	7.06
betweenness centralization	0.15	0.2	0.35	0.17	0.23
number of nodes	84	81	89	98	94
number of links	625	365	291	530	327
number of nodes in main connected component	58.1	58.2	70.8	75.5	74.1
number of Louvain communities of size > 5	4.01	4.4	5.09	4.74	5.16
modularity	0.39	0.52	0.56	0.46	0.58

network for each of them, which is the closest to its centroid (the vector whose 21 coordinates are average among all the elements of the cluster). For each cluster, next to the radar chart, a traditional node-link drawing of the illustrative centroid network is shown, where the size of the nodes is set according to the betweenness centrality and the color according to the degree. The radar chart shows the representativity values of both the centroid network (in black) and its cluster (in its former color).

Note that the names given to these five clusters of networks are quite similar to the names we have given to the five clusters of graphlets in the previous section. There is definitely some correlation between them that will be explicated in Section 5 but these should however not be confused. Table 1 gives the mean values of some classical structural measures for the five clusters. Note that among these measures, the result of the non-deterministic Louvain community detection algorithm may vary. The last two rows of the table are computed on a single run (the modularity is computed from the result of Louvain).

The description given below is based both on the representativity values, the structural properties, and the observation of many examples of networks picked randomly among the clusters.

**Dense cluster - 493 (13% of the corpus)** This cluster gathers many networks with high representativity for the dense graphlets , while low values for paths  or star-like graphlets . On average, the networks are among the smallest of the corpus in terms of number of nodes and, quite logically, the densest. A correlation between the representativity of the dense graphlets and the density of the networks is indeed not a surprise (but we will see for other clusters that there is no equivalence). The very low representativity of the star-like graphlets  suggests that these networks either do not have brokers (alters who create a bridge between communities), or if they do, then these brokers are not very central, as can be confirmed by the low mean value of Freeman's betweenness centralization. Accordingly, this group also has the lowest average values of

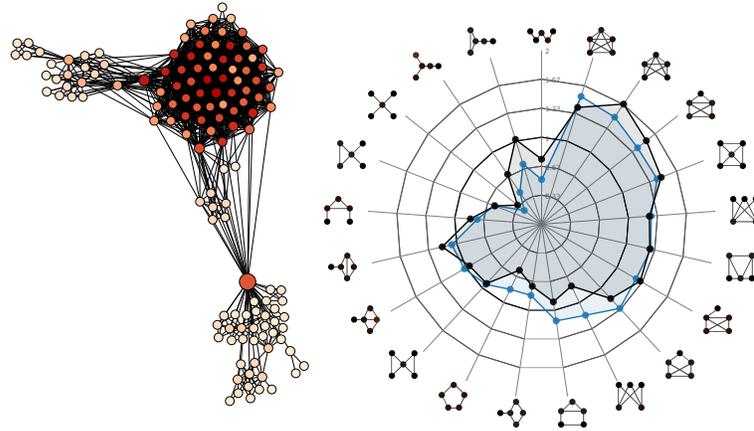


Figure 10: **dense centroid** It is composed by many cliques and quasi-cliques inside its main group of alters (at the top of the figure). The broker between the main group and the one at the bottom has many neighbors in both groups, but when taken in one same group, they are likely to be tied to each other. It explains why there are so few graphlets from the star-likes group.

diameter, number of Louvain communities and modularity.

**Bowties cluster - 664 (18% of the corpus)** We have chosen this group's name because of the bowtie graphlet , whose representativity value is clearly distinct from the other clusters'. These networks are composed by different groups of alters tied by brokers. The relatively high density and clustering coefficient of the networks in this cluster, despite low representativity values for dense graphlets, indeed suggests the presence of at least small groups of connected nodes. Now these brokers are not necessarily so called alter-egos, alters that would share most of ego's friends: these brokers most likely connect no more than two groups each, contrarily to the forthcoming Star cluster. This interpretation is reinforced by the fact that the star  is lightly underrepresented. Among the star-like graphlets  that appear, the two peripheral nodes probably occur most of the time in the same group but not tied to each over, or as nearly isolated nodes in the network. Evidences of pairwise connections between groups of nodes may also be given by two other overrepresented graphlets  . The clique-plus-one  is particularly interesting with an overrepresentativity highest than any other dense graphlet for this cluster.

We can also mention that the holes      have very low representativity values, which might be related to the presence of many brokers.

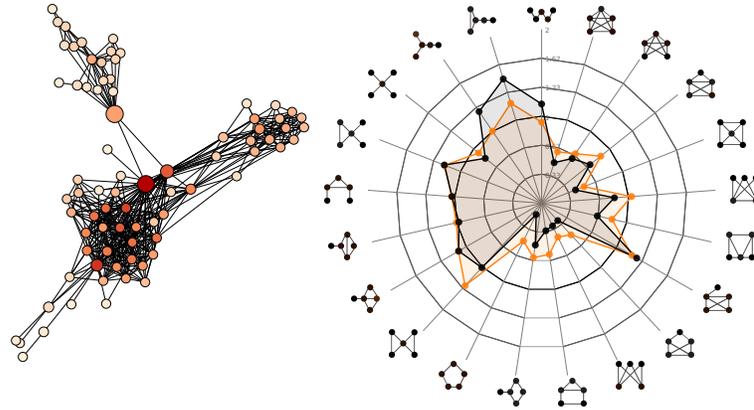


Figure 11: **bowties centroid** This network offers an interesting situation. There are three groups as well as three main brokers that are connected together, forming a triangle. The two in the bottom are involved in many bowties , either including the mentioned triangle or not. Since the groups of nodes are not very dense, many path graphlets are found that probably cross the whole network.

**Star cluster - 759 (20% of the corpus)** The most important representatives of star-like graphlets  are found in this cluster indicating that it is composed of many networks where one or several alters create bridges between others. This star group, unexpectedly, has the lowest representativity of the bowtie graphlet , which shows again that there is no specific relationship between this graphlet and the star-likes even though all of them contain an alter in a position of broker. It means that there is quite a difference between connecting two groups like the bowtie  does and connecting three or more groups as with the star-likes  . This cluster contains the networks with the highest mean Freeman betweenness centralization, which is common for star-shaped networks, and also a high mean modularity value. This once again underlines similarities between the properties of a network and of its most represented graphlets, as was seen earlier for the Dense cluster. The Star cluster has, by the way, one of the lowest representativity values for the dense graphlet, while the star graphlet itself has a low density, and the mean density value of the networks in this cluster is nearly the lowest.

**Holes cluster - 763 (21% of the corpus)** Here are the egos with overrepresented hole graphlets     within their network. The holes are very unlikely in general, occurring much less than other graphlets, as can easily be seen on Figure 4 with the relative frequencies. This cluster may thus put into evidence a social structure that usually remains unnoticed. All the more so that a very similar cluster is produced when using the Graphlet Correlation Distance (Figure 8, purple 487 cluster). There may be two structural

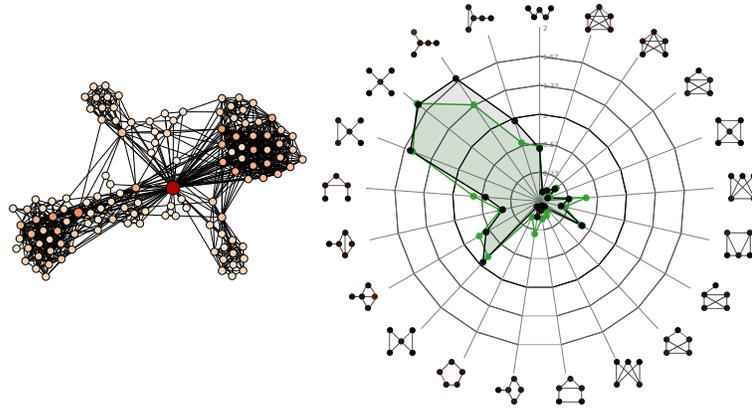


Figure 12: **star-like centroid** This network has a very central alter, that can be considered as a so-called alter-ego of the respondent. It is the center of many of the the star-like graphlets that appear. The presence of other alters that link the groups one-by-one explain why so many fork graphlets  are to be found within that network.

explanations for the appearance of a hole: in a large sparse group of nodes, or through several groups. The idea with the latter case is that if several ties exist between two or more groups, the nodes of the square  or of the 5-node cycle  can be picked in several of these groups, with edges as bridges between them. With a sparse group or several groups connected together, it is also natural to find so many paths graphlets , since the four-length path  is a shared sub-graphlet.

**Sparse cluster - 1015 (28% of the corpus)** Here we have the cluster with the most important representativities of paths graphlets  with also a high representativity of star-like graphlets . Not surprisingly, these graphs have the highest diameter in average since many of them are composed of several groups that form chains of node knots. Quite logically, the networks in this cluster are poor in terms of graphlets of high density, since nodes from different groups are poorly interconnected. Note that this cluster shares, with the Holes (763) cluster, high representativity values for paths graphlets, but they differ on the star-like graphlets (more representative here) and obviously on the holes  (underrepresented here).

The observation of these five clusters of ego networks stresses the relationship between the values of structural indicators at the networks level and at the graphlets level. It is indeed clear concerning the densest graphlets, overrepresented within the densest networks, for the star-like graphlets with modularity or Freeman centralization, and for the paths graphlets and the diameter. Moreover, the graphlet representativities are both correlated with the

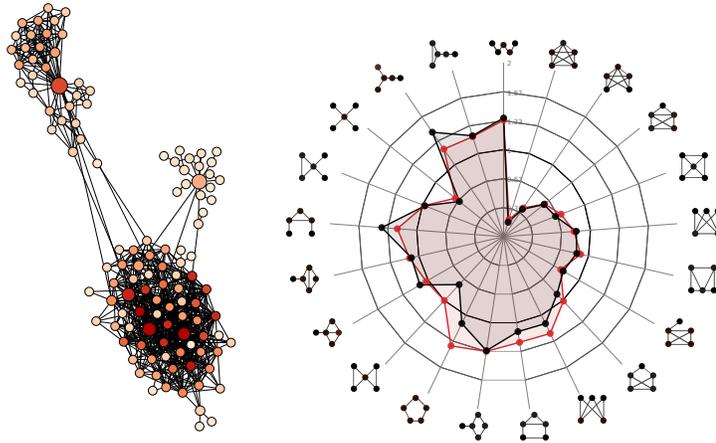


Figure 13: **holes centroid** This network has several groups, a main one which is quite sparse, as can be guessed from the heterogeneous betweenness and degrees among the nodes, and two smaller groups that are both organized around central alters. The largest group, much less dense than it appears, actually contains holes and the clique  is highly under-represented in the network. This makes this network very different – structurally speaking– from the one in Figure 10 despite visual similarities. Also, several edges going between the main group and the two others are likely to be part of any hole graphlet.

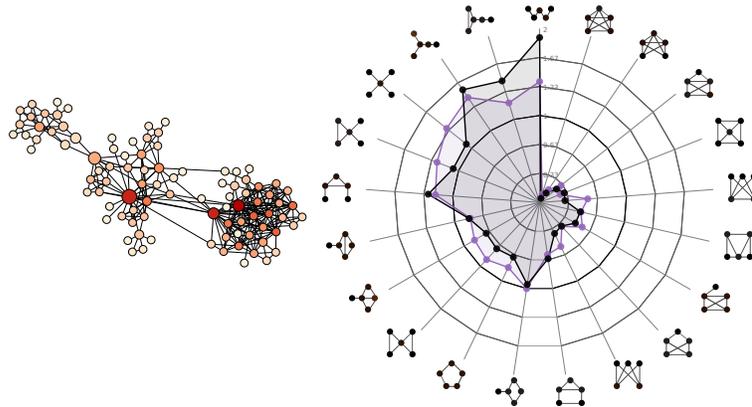


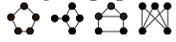
Figure 14: **sparse centroid** This network is built like paths linking small groups of alters. These being sparse, there are a many opportunities of finding, as induced subgraphs, paths that cross them. As there are no central alters that reach several groups, the star-like graphlets as well as the bowtie are a bit less represented in this specific network than on average in the rest of the cluster of networks.

inner structure of groups of alters and with the ties between those groups since they also can be distributed among several of them.

### 4.3 Hints on socio-demographic data

We have provided a clustering of the networks based on the graphlet representativity, along with interpretative elements in terms of network structure. We now add supplementary data from our survey, in order to show that graphlets may be a useful element in a sociological analysis based on a corpus of ego networks. We have combined two variables, the age and the relationship status of our respondents, with the clusters obtained in the previous section. The results are presented in Tables 2 and 3.

We find two classical results. Young persons have denser networks (42% of dense networks' egos have less than 25 years), which is due to the number of different social groups that they usually have met in their lifetime (Bidart and Lavenu, 2005; Kalmijn, 2012) (and this is even truer for students). Moreover, groups of young adults appear to have much more Facebook friendship ties (university cohorts for instance).

The Holes cluster of networks is, with the Dense cluster, the one with the youngest population. We make the hypothesis that a high representativity of holes graphlets  in a network may often be explained by a high homophily among the alters despite a relatively low density: friends of ego may for instance have similar ages or social backgrounds. Young people are more likely to be associated with other persons of their age and cultural environment, which should lead to either big groups of alters with a rather low density or unexpected ties between several denser groups through alters that are not necessarily very central in the network.

The other unsurprising result is the prominence of married persons in the star-like cluster, that confirms previous findings (Bidart and Lavenu, 2005; Kalmijn, 2012; Backstrom and Kleinberg, 2014). Since the central node of the star-like graphlets  is in relation with at least three persons that do not know each other, it surely has a central position within the whole network. Moreover, there are chances that in the many such graphlets appearing in this cluster, the central position is to be matched with the same few nodes (or even one node) in each network, and that one of them is ego's lover.

The discrepancy between the populations of the two tables comes from the fact that some people did not answer all the questions of the survey or did not inform their relationship status on Facebook. It is the occasion to question the variability of the willingness to share private information with a platform like Facebook versus with sociologists. For instance, some respondents gave Facebook aberrant values for their age while giving more credible answers to our survey. Our survey did not ask for the romantic status but it is noteworthy that the ratio between the number of persons who gave it to Facebook and the number of those who gave us their age, varies significantly from one cluster of networks to another: it is 0.61 and 0.66 in the Dense cluster and the Holes cluster respectively, while 0.72 in average. This might indicate that young people (more present in these two clusters) are less eager to give their relationship status to Facebook. In the meantime, the Star cluster is the one with the

Table 2: Age distribution among each cluster of networks (those described in Section 4.2). The sum of each column is 1 up to rounding errors. The age is the one declared by the respondents in the survey application.

age	dense	fork	star-like	holes	bowtie	all
18-25	0.42	0.26	0.11	0.42	0.26	0.28
26-40	0.32	0.46	0.56	0.31	0.45	0.43
41-60	0.22	0.25	0.3	0.24	0.27	0.26
60+	0.03	0.03	0.03	0.03	0.02	0.03
pop.	322	614	487	494	401	2318

Table 3: Distribution of the relationship status among each cluster of networks (those described in Section 4.2). The sum of each column is 1 up to rounding errors. The relationship is the one declared on the Facebook page of the respondents, when available.

Facebook relationship	dense	fork	star-like	holes	bowtie	all
single	0.42	0.27	0.14	0.4	0.31	0.29
couple	0.27	0.39	0.36	0.34	0.32	0.35
married	0.31	0.34	0.5	0.26	0.37	0.36
pop.	197	450	414	328	278	1667

highest value for this ratio, so it may just indicate that married people (more present in this cluster) are more likely to advertize it than single people (of course more present among young users).

## 5 Shifting to 4-node graphlets?

We have obtained in Section 3 five clusters of rather similar graphlets, and in Section 4 five clusters of networks grouped according to the distributions of the graphlet representativity values. Figure 15 sums up the combination of these two clusterings, presenting our corpus of networks in a very sketchy visualization. A much simpler way of simplifying the analysis would have been to count only the 6 graphlets of size 4 instead of clustering the 21 graphlets of size 5. All the more so that as already mentioned, the time complexity of the computation of the graphlet enumeration made us choose to put an upper limit to the size of the networks in our corpus.

Figure 16 shows the result of a clustering of the networks of our larger corpus of 12,242 networks. As with the previous clustering, we obtain a (blue, 2044) cluster with high representativities of the dense graphlets and a (purple, 2718) cluster with overrepresented paths  and stars . The (orange, 1852) cluster is more likely to correspond to the Bowties cluster since it has a bit more of cliques  and hanger  than other clusters. Un-

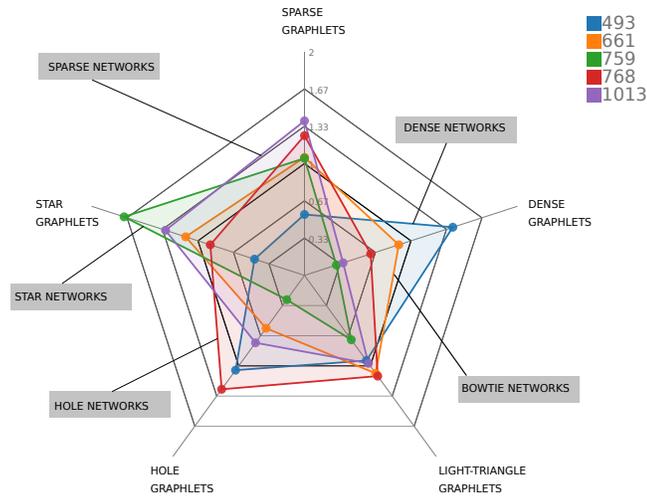


Figure 15: Network clusters depending on the clusters of graphlets

fortunately, this cluster as well as the one with the highest representativity of squares  (red, 1739) are clearly more difficult to read here than in the case of 5-node graphlets since they do not have representativities very far from 1. The specificity of the bowtie graphlet  compared to the star-like graphlets , discussed in the previous section, is an example of structural properties that can hardly be caught with the size-4 graphlets.

To be able to formally compare these two clusterings, we have used the Rand index, a measure of how two clusterings are close to each other. Its result is between 0 (for two identical clusterings) and 1 (for two clusterings with no agreement). In the comparison of the clustering made through 21-dimensional vectors from the graphlets of size 5, and the clustering made through 6-dimensional vectors from the graphlets of size 4, we can compute the Rand index limited to the smaller networks that appear in both cases. The score of 0.79 we obtain indicates that there is a high similarity between them. Of course some networks were moved from a cluster to another but their cores are preserved.

The clustering based on graphlets of size 5 is more interesting in the sense that the various clusters are easier to interpret and to understand, with much richer structural configurations. But the relatively good value of the Rand index suggests an empirical trade-off between insight and performance: once clusters have been computed the expensive way with 5-node graphlets for networks of limited sizes, the larger networks may be attributed to clusters according to the faster computation based on the 4-node graphlets.

## 6 Conclusion, future work and applications

After recalling previously defined metrics over networks based on graphlet counting, we have proposed an efficient measure that we call graphlet representativity, to study ego net-

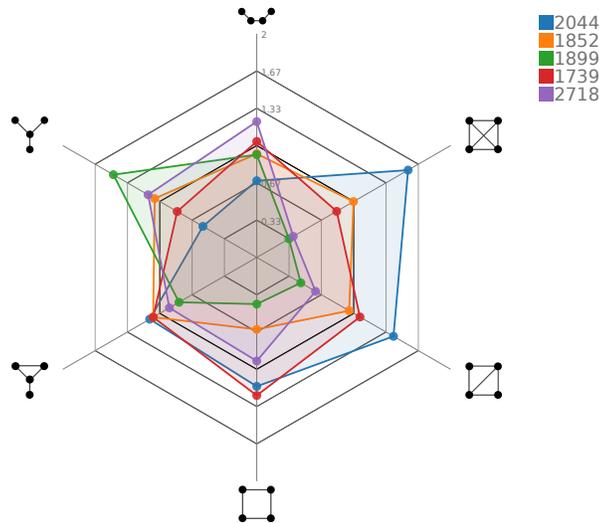


Figure 16: Clusters obtained by a k-means algorithm with the representativity data of the graphlets of four nodes.

works from a possibly large corpus. With this measure we have first produced a clustering of the thirty graphlets of size 5, into five groups of graphlets (paths, star-like, holes, light triangles, dense). Both these graphlet clusters and the representativity itself provide a useful visualization scheme with radar charts, helping the process of designing a clustering of the ego networks of our corpus. We have then been able to exhibit five clusters of networks with shapes that are consistent with the structure of their most represented graphlets (dense, bowties, stars, holes, sparse), putting into evidence non-trivial structural properties such as the distinction between bowties graphlets (for more local centrality) and star-like graphlets (for more global centrality), or overrepresentativity of (generally unlikely) holes in some networks. This shows the relevance of counting all the graphlets rather than selecting only some of them, despite the computational cost. As for computation, we have shown that the enumeration of the 4-node graphlets is a powerful tool even if it is not as rich as the enumeration of 5-node. We have thus proposed a progressive analytical workflow, joining the richness of the latter and the efficiency of the former.

Now from a sociological point of view, the work starts here: provided with a new structural measure and five categories of ego networks based on it, one can investigate many questions about the nature of Facebook ties, the structure of personal networks, and the relationships between the two. A first lead to address these would be to explore, for each ego, the subpart of the friends network induced by the alters who comment on ego's posts. A friends network can be seen as a base map upon which the activity is performed by only a fraction of alters who play the actual Facebook game in ego's social life. The graphlet representativity might be used to compare the structure of the networks of Facebook friends and of Facebook commenters of ego's. One would of course consider the commenters networks as a separate corpus, by recomputing the reference graphlet frequencies for the computation of the repre-

representativity values, and then compare, for each ego, their two networks. With, or without the structures of the commenters networks, one might investigate the social meaning of some characteristic structures: what types of activity and interaction take place among networks with overrepresented dense graphlets? Are path-like graphlets only an evidence of a lack of engagement in the platform? Do networks from the Bowties or Star clusters contain distinct cohesive subgroups with separate public conversations with ego? This question is also related to qualifying the social meaning of subgroups of alters among the ego networks, which can be addressed with the answers to the short survey about ties between ego and at least five of their most active alters (which are qualified as family, coworkers, lover, etc.), and with the cooccurrences of alter comments on ego's posts, of tags, etc. This may also be a way to investigate whether, as was hypothesized in Section 4.3, networks with an overrepresented amount of holes graphlets indicate a high level of social homophily.

Working on subgroups may also be done by considering a corpus of 'communities' in the heuristic sense of groups returned by the so-called Louvain community detection algorithm. The networks in our corpus have a mean number of about 6 such groups, which would make a corpus of about 60 000 small networks upon which to compute reference graphlet frequency values, then graphlet representativities. Ego networks would thus be analyzed according to the types of groups they are composed of. Another method to untangle the ego networks is to investigate structural roles of alters, which may be done by keeping tracks of the alters which appear in each position of each graphlet while counting. These positions are what is called orbits in the graphlet literature. One could compute the graphlet representativity of all the orbits in the ego networks of our corpus, which would exhibit social roles among the networks. These structural roles are of course to be related to the interaction data of the alters with the egos, and with the information the respondents have given about their relationships with some selected friends. The special case of the structural positions of lovers is a first issue that can easily be addressed. Furthermore, to all the issues mentioned above, the findings may vary according to socio-demographic variables such as ego's profession, age and city of residence, which have been collected in the survey. Now more generally, all these directions are indications of analyses can be performed on other corpuses of personal networks, regardless of how they are collected, including traditional surveys with manual recording by the respondents themselves.

Of course our approach is applicable to other types of networks, including non-ego networks, including networks of non-social ties, as, of course, protein-protein interaction, but also functional connectivity among brain regions, citation links between documents, etc. As soon as one needs to compare and categorize many networks of same kind, the approach presented here is relevant. Moreover, on networks of various origins, one could investigate how the graphlet representativity compares to methods presented in Section 2.1 to categorize them.

Regarding the technical aspects of graphlet counting, there may also be followups to the present work. In particular the relationships between graphlets of size  $k$  and of size  $k + 1$ , are of course a key issue in the combinatorial and probabilistic aspects of the problem. Comparing the relative representativities of dependent graphlets might be a way to investigate and measure these dependancies.

## References

- Ali, W., Rito, T., Reinert, G., Sun, F., and Deane, C. M. (2014). Alignment-free protein interaction network comparison. *Bioinformatics*, 30(17):i430–i437.
- Backstrom, L. and Kleinberg, J. (2014). Romantic partnerships and the dispersion of social ties: a network analysis of relationship status on facebook. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 831–841. ACM.
- Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509–512.
- Barnes, J. A. (1954). Class and committees in a norwegian island parish. *Human Relations*, 7(1):39–58.
- Bidart, C. and Lavenu, D. (2005). Evolutions of personal networks and life events. *Social networks*, 27(4):359–376.
- Bornholdt, S. and Schuster, H. G. (2006). *Handbook of graphs and networks: from the genome to the internet*. John Wiley & Sons.
- Bott, E. (1957). *Family and social network: Roles, norms and external relationships in ordinary urban families*. Routledge.
- Brandes, U., Robins, G., McCranie, A., and Wasserman, S. (2013). What is network science? *Network Science*, 1(01):1–15.
- Brooks, B., Hogan, B., Ellison, N., Lampe, C., and Vitak, J. (2014). Assessing structural correlates to social capital in Facebook ego networks. *Social Networks*, 38:1–15.
- Cunningham, P., Harrigan, M., Wu, G., and O’Callaghan, D. (2013). Characterizing ego-networks using motifs. *Network Science*, 1(02):170–190.
- Easley, D. and Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.
- Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999). On power-law relationships of the internet topology. In *ACM SIGCOMM computer communication review*, volume 29, pages 251–262. ACM.
- Faust, K. (2010). A puzzle concerning triads in social networks: Graph constraints and the triad census. *Social Networks*, 32(3):221–233.
- Freeman, L. (2004). The development of social network analysis. *A Study in the Sociology of Science*, 1.
- Friggeri, A., Chelius, G., and Fleury, E. (2011). Triangles to capture social cohesion. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 258–265. IEEE.

- Hocevar, T. and Demsar, J. (2014). A combinatorial approach to graphlet counting. *Bioinformatics*, 30(4):559–565.
- Hogan, B. (2018). Social Media Giveth, Social Media Taketh Away: Facebook, Friendships, and APIs. *International Journal of Communication*, (12). preprint.
- Holland, P. W. and Leinhardt, S. (1976). Local structure in social networks. *Sociological methodology*, 7:1–45.
- Hočevár, T. and Demšar, J. (2017). Combinatorial algorithm for counting small induced graphs and orbits. *PLOS ONE*, 12(2):e0171428.
- Kalmijn, M. (2012). Longitudinal analyses of the effects of age, marriage, and parenthood on social contacts and support. *Advances in Life Course Research*, 17(4):177–190.
- Milo, R. (2002). Network Motifs: Simple Building Blocks of Complex Networks. *Science*, 298(5594):824–827.
- Moreno, J. L. (1934). *Who shall survive*, volume 58. JSTOR.
- Nasim, M., Charbey, R., Prieur, C., and Brandes, U. (2016). Investigating Link Inference in Partially Observable Networks: Friendship Ties and Interaction. *IEEE Transactions on Computational Social Systems*, 3(3):113–119.
- Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.
- Park, N., Lee, S., and Kim, J. H. (2012). Individuals’ personal network characteristics and patterns of Facebook use: A social network approach. *Computers in Human Behavior*, 28(5):1700–1707.
- Pinar, A., Seshadhri, C., and Vishal, V. (2017). Escape: efficiently counting all 5-vertex subgraphs. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1431–1440. International World Wide Web Conferences Steering Committee.
- Pržulj, N. (2007). Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183.
- Pržulj, N., Corneil, D. G., and Jurisica, I. (2004). Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515.
- Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007). An introduction to exponential random graph ( $p^*$ ) models for social networks. *Social Networks*, 29(2):173–191.
- Scott, J. (2017). *Social network analysis*. Sage.
- Simmel, G. (1908). Sociology: investigations on the forms of sociation. *Duncker & Humblot, Berlin Germany*.
- Spiliotopoulos, T. and Oakley, I. (2013). Understanding motivations for Facebook use: Usage metrics, network structure, and privacy. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3287–3296. ACM.

- Stoica, A. and Prieur, C. (2009). Structure of neighborhoods in a large social network. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, pages 26–33. IEEE.
- Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–2.
- Wellman, B. (2007). The network is personal: Introduction to a special issue of Social Networks. *Social Networks*, 29(3):349–356.
- Wernicke, S. (2006). Efficient detection of network motifs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(4).
- Wernicke, S. and Rasche, F. (2006). Fanmod: a tool for fast network motif detection. *Bioinformatics*, 22(9):1152–1153.
- Yaveroglu, Ö. N., Malod-Dognin, N., Davis, D., Levnajic, Z., Janjic, V., Karapandza, R., Stojmirovic, A., and Pržulj, N. (2015a). Revealing the Hidden Language of Complex Networks. *Scientific Reports*, 4(1).
- Yaveroglu, Ö. N., Milenković, T., and Pržulj, N. (2015b). Proper evaluation of alignment-free network comparison methods. *Bioinformatics*, 31(16):2697–2704.
- Zhao, Z., Wang, G., Butt, A. R., Khan, M., Kumar, V. A., and Marathe, M. V. (2012). Sahad: Subgraph analysis in massive networks using hadoop. In *Parallel & Distributed Processing Symposium (IPDPS), 2012 IEEE 26th International*, pages 390–401. IEEE.