Community detection in Slovene public spending

Skupnosti v slovenskem javnem naročanju

Vitjan Zavrtanik, Lovro Šubelj
University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, 1000 Ljubljana
vitjan.zavrtanik@gmail.com; lovro.subelj@fri.uni-lj.si

Abstract

Government public spending is a highly complicated system with many endpoints, as government funding has to be distributed to a large number of institutions that purchase goods and services using public funds from a large number of public and private organizations. Even though there are mechanisms in place which control public spending by either entrusting control to a public institution or by prescribing legal procedures that must be observed when purchasing services, these procedures are sometimes unreliable or enable exploitation. Procedures often comprise public tenders which can be exploited by bidding private firms or public servants. We hypothesise that there are patterns in such behaviour that could perhaps be identified via community detection in the public spending transaction network by examining publicly available transaction data on public funds.
Keywords: public spending, community detection, network analysis.

Izvleček

Sistem porabe javnega denarja je kompliciran predvsem zaradi potrebe po primernem razporejanju med številne državne institucije, ki sredstva porabljajo za blago in storitve potrebne za njihovo delovanje. Zaradi ogromnega števila transakcij je sistem težko nadzorovati in čeprav obstajajo mehanizmi za nadzorovanje porabe kot so javni razpisi, obstaja možnost manipulacije postopka izvajanja razpisa s strani podjetij v privatni lasti, ki se na razpis prijavljajo, ali pa s strani javnih uslužbencev, ki razpis prijavijo. Naša hipoteza je, da v takem obnašanju obstajajo vzorci, ki bi jih lahko razbrali z detekcijo skupnosti v omrežju transakcij javnih sredstev.
Ključne besede: poraba javnih sredstev, detekcija skupnosti, analiza omrežij

1. INTRODUCTION

Allocating and distributing funds is a complicated problem. A great deal of government effort goes directly into the deciding how and when to spend public funds, a majority of which goes into the essential systems such as healthcare, education and pensions which are comprised of many public institutions and are also the primary source of income for a significant percentage of private companies. Due to the vast amount of resources such public institutions utilise, there are quite a few regulations that must be followed, some of which are defined in (Zakon o javnem narocanju (ZJN-3), 2015). One of the procedures for procuring required services and goods from companies is the public tender which enables institutions to publicly announce that they require a certain service and the companies that perform that service are able to present their offers for the service. Often the public institution is obligated to purchase the service from the lowest bidder. Other times stricter criteria are involved in choosing the winning bid on a public contract, which is often the cause of conflict as the institution requiring the service is accused of corruption by overfitting the selection process to a certain bidder.

One example of such controversial activity is the recent construction of the railroad project showcase model in which the institution providing the contract first picked the more expensive bidder based on a certain criteria and the selected company later hired a less expensive bidder that was not selected to perform a portion of the service it was contracted to do (MMC RTV SLO, 2018).

While the idea of public tenders seems good on paper, it has flaws in its enforcement and regulations. It has often been criticized for the possibility of fine tuning the tender documentation to better suit a specific bidder. There is also a possibility of offer price fixing by the participating bidders. If the managers of the bidding companies knew each other, they could easily change the offer price for a certain tender which could raise the profits. Such price fixing is illegal.

We could hypothesise that there is a higher chance that price fixing occurs if the people in these companies know each other and perhaps the companies even often work together. There is unfortunately no publicly available data that shows direct business cooperation between privately owned companies as they are not obligated to report such data.

The transaction data between public or between public and private institutions are publicly available and searchable on the Erar tool (Commission for the Prevention of Corruption, 2018) maintained by the Commission for the Prevention of Corruption. The transaction data is available for download in the *csv* file format for each fiscal year separately. This data can be parsed into a network representation where nodes are public and private institutions and the links between them are the transactions or the sums of transactions between them. This gives us a network which depicts the flow of funds from public institutions to private ones. While this could be used to analyse a variety of the system properties.

2. RELATED WORK

In (Kolar & Kolar, 2017) the data from Erar (Commission for the Prevention of Corruption, 2018) was used to rank the importance of institutions in the network and was used to simulate the robustness of the network to node and edge removals. In (Kogovšek, Sovdat, & Povšič, 2013) similar data was used, however they decided to connect owners and representatives to companies based on their affiliation and attempted to discover communities. In (Lozano, Duch, & Arenas, 2006) the authors use a method based on modularity measures (Girvan, 2002) to discover communities in a large social dataset of European projects. There are many community detection algorithms available ((Rosvall & Bergstrom, 2008), (Blondel, 2008), (Ahn, Bagrow, & Lehmann, 2010)) however for a network as large as ours algorithms with the computational complexity of $O(n^2)$, where n is the number of nodes in the network are not feasible for executing on a personal computer since the execution time needed for the method to finish is substantial. In (Šubelj, Jan, & Waltman, 2016) the authors evaluate a variety of clustering methods on citation networks.

3. DATA AND METHODS

We gather all of the used data from the Erar tool (Commission for the Prevention of Corruption, 2018). The transaction data is available in *csv* format and is available on a yearly basis from the year *2003* onward. We use the transaction dataset for the year 2017 to reduce the amount of data needed to process, however this is still a list of approximately 23 million transactions between about 88000 private and public institutions. We are however not as interested in the network we can generate from this data directly but in the network that we can construct with the addition of representation and ownership data. Ownership and representation data is available through the Erar API (Commission for the Prevention of Corruption, 2018), which we must query for every institution. We can gather the data on present and past ownership and representation. We hypothesize that the cooperation between companies and institutions is not based only on the institutions themselves but on the people that represent or own these companies. Since we have the ownership and representation data of the companies we can instead construct a network of people which could give us some insight in the cooperation between private companies for which the data is not publicly available as it is likely that the people who are or were once co-owners or representatives of companies know each other. Because of this we assume that there is a certain community structure that could be extracted.

In (Kogovšek, Sovdat, & Povšič, 2013) community detection on a similarly constructed network has

been done using modularity maximization methods that have trouble detecting smaller networks due to the resolution limit ((Barabási & Pósfai, 2016), (Fortunato & Barthelemy, 2007)). This may be important for this specific problem as the communities that we are looking for could be much smaller, especially if we want to extract information of the possibility of collusion on public tenders.

Appropriate algorithms for a network of this size are Infomap (Rosvall & Bergstrom, 2008) and Louvain (Blondel, 2008) methods as they excel in speed with the computational complexity of $O(n \log(n))$, where n is the number of nodes in the network. The Louvain method is, however, also a modularity maximization algorithm and is therefore also affected by the resolution limit. We also wanted to test the performance of a further subdividing the induced graph of the larger acquired communities using perhaps less scalable community detection methods such as (Ahn, Bagrow, & Lehmann, 2010).

3.1 Data preparation

In order to construct the network, the extracted data from Erar (Commission for the Prevention of Corruption, 2018) had to be augmented with the Register of budget users downloaded from the Ministry of finance database (Uprava Republike Slovenije za javna plačila, 2018). This additional data was required as Erar's transaction data does not contain the name of the institution who issued the transactions and instead only lists the bank account from which the funds were taken. We had to extract a list of bank accounts from all the public and private institutions from Erar and match the account to the transaction data in order to get the names of the institutions transferring the funds. After constructing the list of institutions, we had to again query Erar in order to get the list of representatives and owners of the discovered business entities.

We constructed the network so that the edges are present between two people either if they work or had previously worked in the same institution, or if the institutions for which they worked are connected by a transaction. This gives us the possibility of using two types of edges in the graph which could show whether two people are affiliated only by working for the same company or due to some business activity between the two companies. The nodes could also be divided into people representing public institutions and people owning or representing private companies. It is worth mentioning that there are quite a few people who worked in both the public and private sector.

3.2 Person to person network

The constructed network consists of *157417* nodes representing the individuals working in public and private institutions. The network contains *1683451* edges. A large portion of the edges are due to the way we connected the people in the graph. We assumed that two people know each other and could collaborate if they both worked as representatives of the same firm. As a result of this, individual companies are represented as cliques.

Another assumption that we made was that if there was some business done between two companies, the representatives of these companies know each other. This assumption may be inaccurate when dealing with large institutions such as major banks, where the number of representatives and number of transactions is very high, which obviously makes the representatives of the bank highly connected due to the number of entities the bank does business with. Of course we cannot assume that a bank representative is aware of every single transaction, therefore the assumption is violated. The disproportionately large number of connections for large institution representatives makes it difficult to accurately analyse the actual connections between the representative and other individuals.

The degree distribution of the extracted network can be seen in Figure 1, where we can see that the distribution is roughly scale free. The average degree is $\langle k \rangle = 21.38$ and the maximum degree $k_{max} = 5775$ which belongs to one of the bank representatives described previously.

4. RESULTS

We used the community detection algorithms Infomap (Rosvall & Bergstrom, 2008), Louvain (Blondel, 2008), Label Propagation (Cordasco & Gargano, 2010) and METIS (Karypis & Kumar, 1998) to extract community structure data from the constructed network. Some results can be found in Table 1 where we can see that Louvain, Infomap and Label Propagation algorithms detect a large number of communities and that the average size of these communities is fairly low. This can be explained by examining our

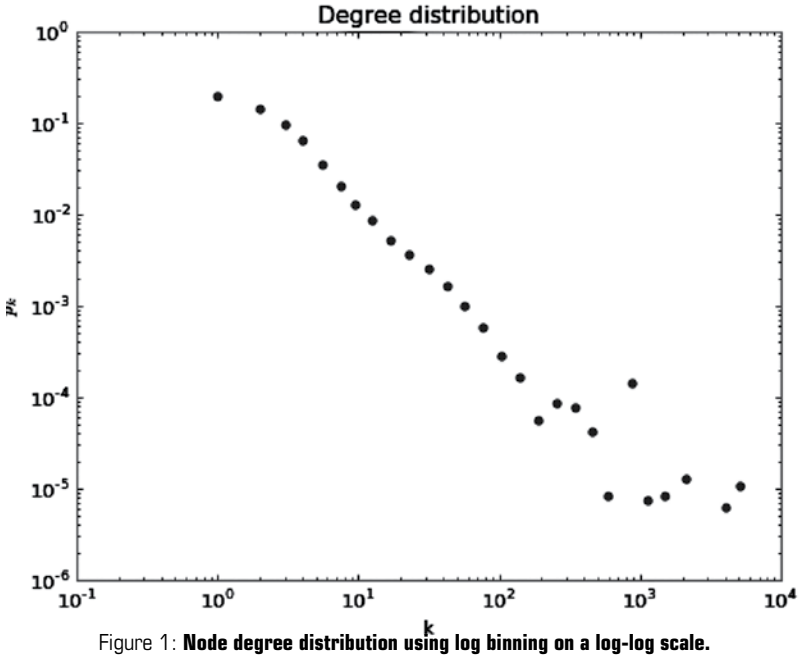

Figure 1: **Node degree distribution using log binning on a log-log scale.**

data where we can see that there are many nodes that are connected only to a very low number of other nodes which could cause the algorithms to exclude them from other communities. These low degree nodes often represent individual contractors that appear in the data because they probably did some work for the government in the fiscal year *2017*, but are not connected to any other institution.

The METIS method is a k-way graph partitioning algorithm and is dependent on the partition number parameter *K* as it will partition the graph more or less uniformly into *K* clusters. Contrary to other methods it's resolution is arbitrarily high and can be used to partition the graph into many small clusters. This seems like a favourable quality for our application as we want to observe small clusters of people and their connections in the network. In practice however the algorithm depends so much on *K* that the resulting clusters make no sense if *K* is set too high or too low. In (Šubelj, Jan, & Waltman, 2016) good results were achieved by using METIS in conjunction with other community detection algorithms. We combined the Louvain and METIS methods so that the communities are first detected with Louvain and are then further subdivided using METIS as some of the communities detected by Louvain are very large. In Table 1 we can see that the combination of methods does indeed subdivide larger communities detected by Louvain however the quality of these subdivisions again depends on *K* and on the communities that are being subdivided as there are some large communities that cannot be partitioned in a sensible way without losing information about the network structure.

In Table 1 we can also see that the sizes of maximum communities are quite high. This is not unexpected since there are individual nodes with a very high degree as described in Subsection 3.2 and it is also frequently the case that these nodes are connected to each other forming a strong community structure. The distribution of the community size is also quite different depending on the method that is used as we can see in Figure 3, Figure 4, Figure 5 and Figure 6. We can see that the Louvain+METIS method results in very small clusters however this depends on the choice of the K parameter. The other methods typically results in a much higher probability of large communities.

Table 1: **Number of communities detected, average, maximum and minimum size of the detected communities for each of the utilized community detection methods.**

Method	$K_{community}$	K_{avg}	K_{max}	K_{min}
Louvain	43692	3.6	23832	1
Infomap	47251	3.33	3672	1
Label prop.	43819	3.59	22298	1
METIS	10000	15.74	17	11
Louvain+METIS	47223	3.32	45	1

Table 2: **Number of communities detected (Kcommunity), average (Kavg), maximum (Kmax) and minimum (Kmin) size of the detected communities for METIS executions with different partition number parameters (K).**

K	$K_{community}$	K_{avg}	K_{max}	K_{min}
100	100	1574.3	1620	1459
1000	1000	157.44	162	141
10000	10000	15.74	17	11
20000	19996	7.87	11	

4.1 Evaluation

Due to no known community structure of the data, we have no data to compare it with. This makes evaluation of the results difficult. We can of course manually look at the results and interpret the quality of the discovered communities however this is time consuming and not quantifiable.

Nevertheless, this approach seems to be useful if we visualize a discovered community that contains a certain individual. Many times it finds an informative representation of the business network of a certain person. It shows us which people that person is in business with and with which people that person has interacted by working in the same company. As

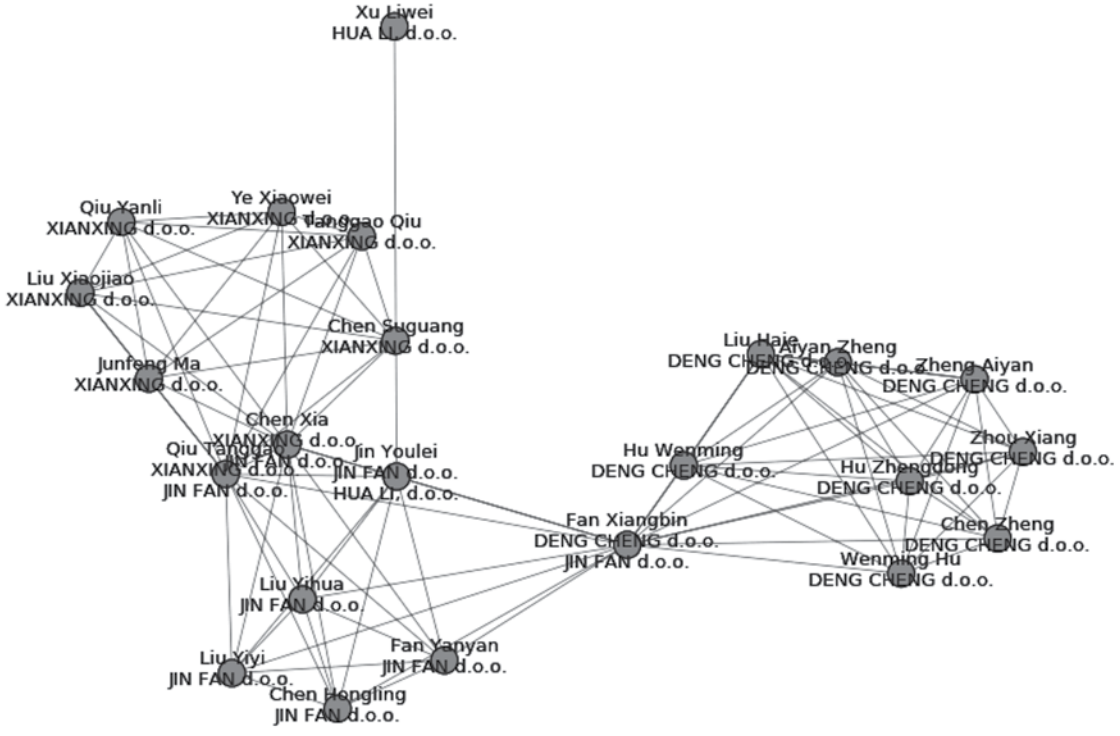

Figure 2: **Personal community with a single detected community.**

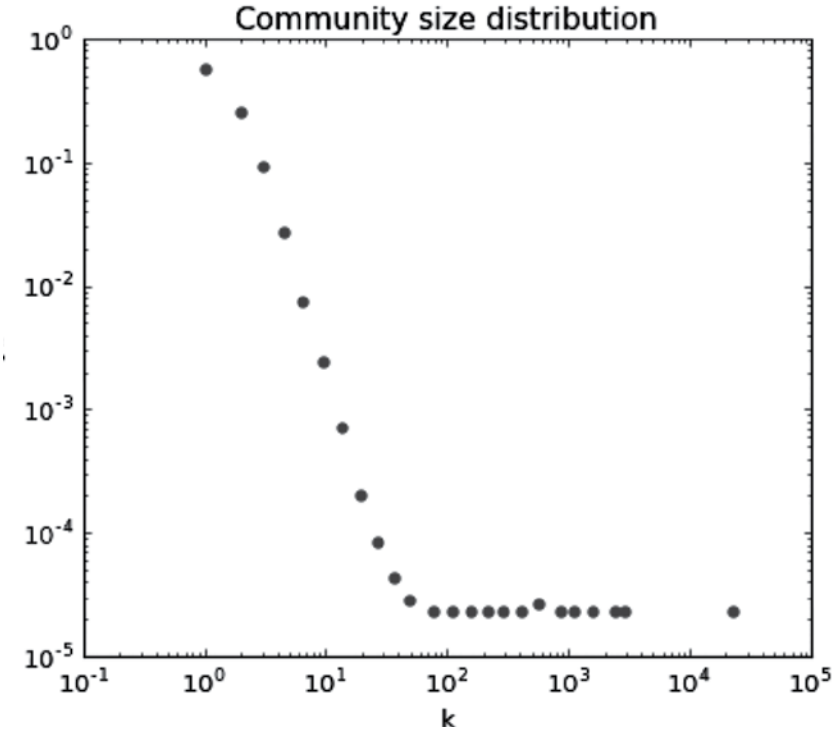

Figure 3: **Louvain method community size distribution.**

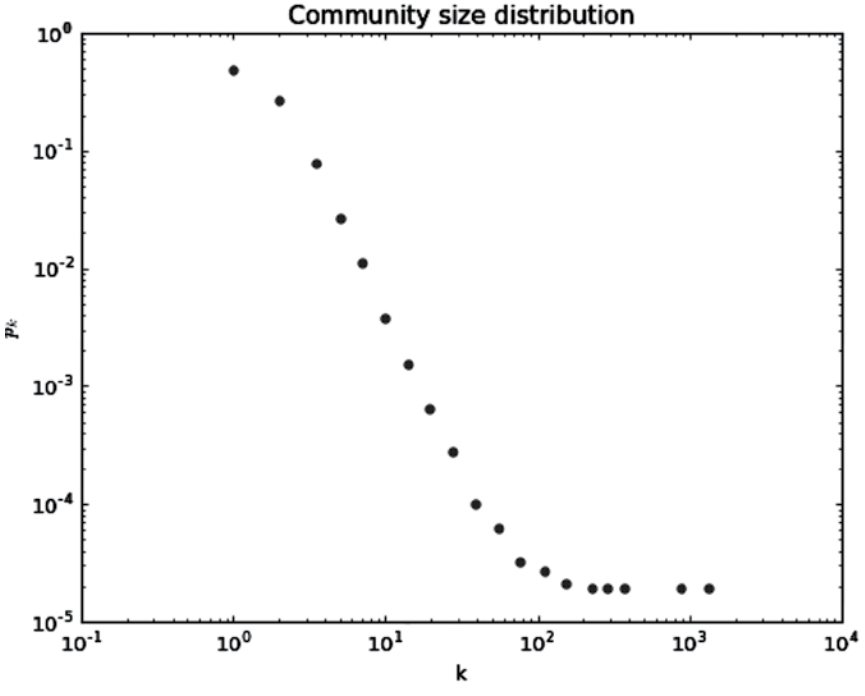

Figure 4: **Label propagation method community size distribution.**

mentioned in Subsection 3.2, there is limited usability of this approach for individuals that turn out to be hubs in our network, as the data is rarely accurate since we assume that if business is being conducted between two entities, the representatives must know each other.

We found that a visualization of a person's community is frequently sensible if we visualize the detected community of a certain node and the neighbours of all the members of the detected community. Visualizing neighbours gives us additional information which is frequently needed since the detected community is regularly composed of people within a single company. The neighbour approach sometimes fails to work for visualizing the surrounding community as the companies are often well separated from the rest of the network. An example of this can be seen in Figure 2.

A significant problem that we are running into and currently have no way of fixing is our inability to distinguish between individuals with the same name. We assume that the majority of individuals have a name that is unique enough that there are no other owners or representatives of companies with the same name however we can never be sure. This of course is not always the case which is why we have nodes in our data that are vastly more connected than they should be due to the fact that it represents multiple people, for example the node that represents *Janez Novak* is connected to several hundred nodes only by affiliation with 17 different entities which of course were not founded by a single *Janez Novak*. The visualization for this graph is therefore not informative at all and the subgraph is also composed of several large communities. There are currently no possible ways of mending this issue as the data required, such as personal identification numbers, are not publicly available.

The visualization of the communities is quite difficult since we want to display the names of the individuals as well as the companies with which they are affiliated. In the event of a larger community the visualization often gets filled with text and would require an alternative solution to visualizing this data. An example of a poorly readable visualization can be seen in Figure 7.

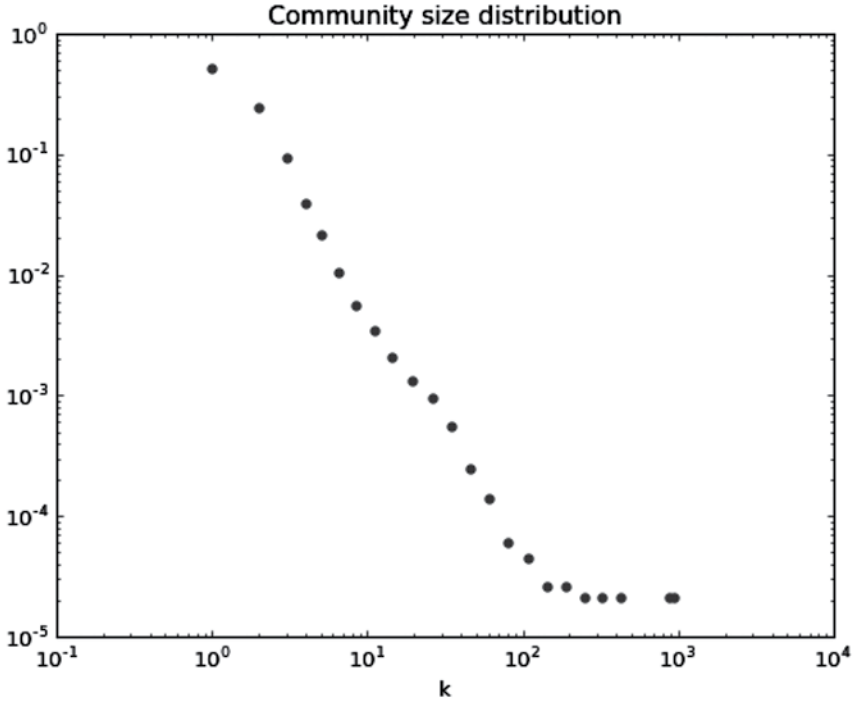

Figure 5: **Infomap method community size distribution.**

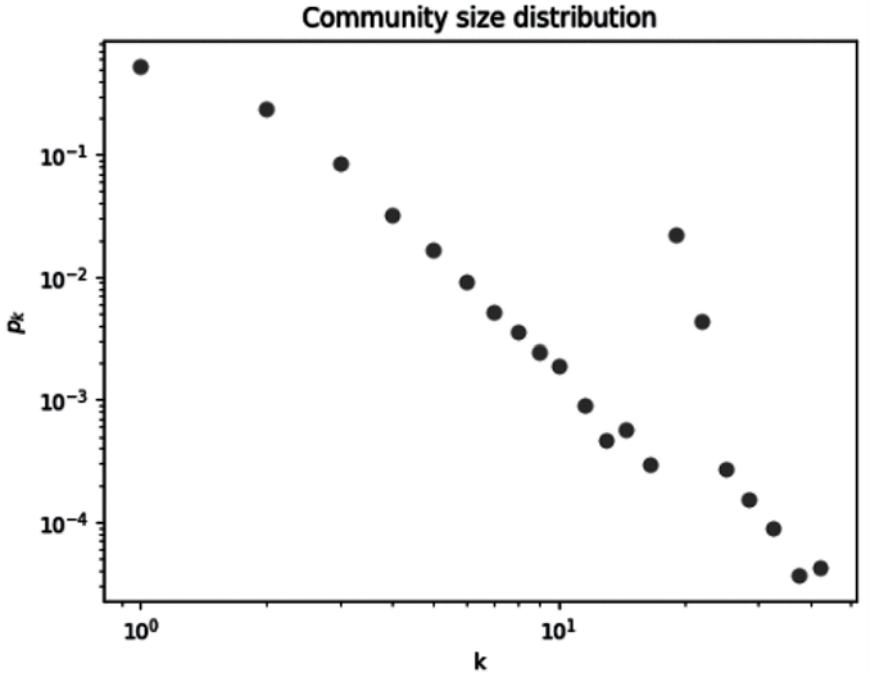

Figure 6: **Louvain + METIS method community size distribution.**

4.2 Visualization

As we can see in Figure 7, the visualization of the communities is lacking in clarity due to the overlapping text and uninformative node colours. We significantly reduced the font size of the information displayed for each node and reduced the amount of information that is displayed by text. In Figure 7 we see that the both the entity name and the institution it is affiliated with are written over the node, in certain individuals the number of affiliated institutions is very high resulting in a block of text that is hard to read.

We decided that it is better to remove most of the affiliation information from the visualization and instead use node colours to show which people belong to the same institution. We previously used colour coding to display which discovered community a certain node belongs to. This information is lost from

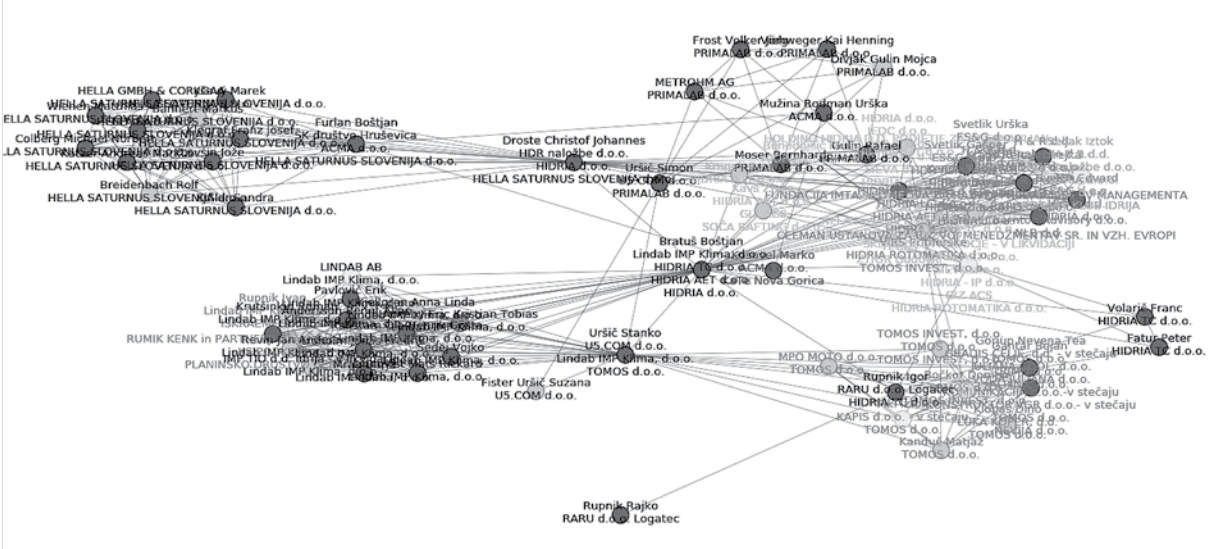

Figure 7: **Multiple detected communities. The figure showcases the problematic visualization when depicting larger networks.**

the visualization however we feel that the information is not as relevant to the observer as a single visualization contains all the nodes of a single community and their neighbours and is therefore in part still contained within the subgraph. The colour coded information is of course shown within the legend adjacent to the network visualization.

The edges in the graph can signal that the nodes are connected by affiliation with the same company, a transaction between two institutions or by both. We coloured the different types of links so that the nature of the connection between two nodes is more apparent.

The community displayed in Figure 2 is better visualized in Figure 8 where only the names of the individuals are displayed as text and the company names are listed in the legend.

4.3 Connection type

The edges in our network can be a result of two factors. An edge can either be present due to a direct collaboration between two entities signalled by a transaction between them, or due to two individuals representing the same organization. This gives us 3 distinct edge types, since edges can be present due to transaction, affiliation or both. We were especially interested in the latter as an edge of this type would mean that the individuals presented by the nodes are were at one time affiliated with the same institution and that there is a possibility that they are now handling transactions of public funds.

There are very few cases where people are involved with both public and private institutions and are doing business with the public institution that employed them. Edges of this type mostly appear in transactions between public institutions. In the cases where one of the connected nodes is a representative of a private company, business is mostly conducted with a representative of the local community of the area where the company operates. In the majority of such cases, the representative of the public institution was previously at the same private company as the service provider. These sort of transactions could be legitimate since local communities are usually small and it is possible that there are no other companies in the area that offer the same kind of services. The fact that these people are connected by previous employment in the private sector should still be taken into consideration when reviewing these transactions.

Our network has only 64 such edges. One of the possible reasons why the occurrence of these edges is so low is because we only use the transaction data

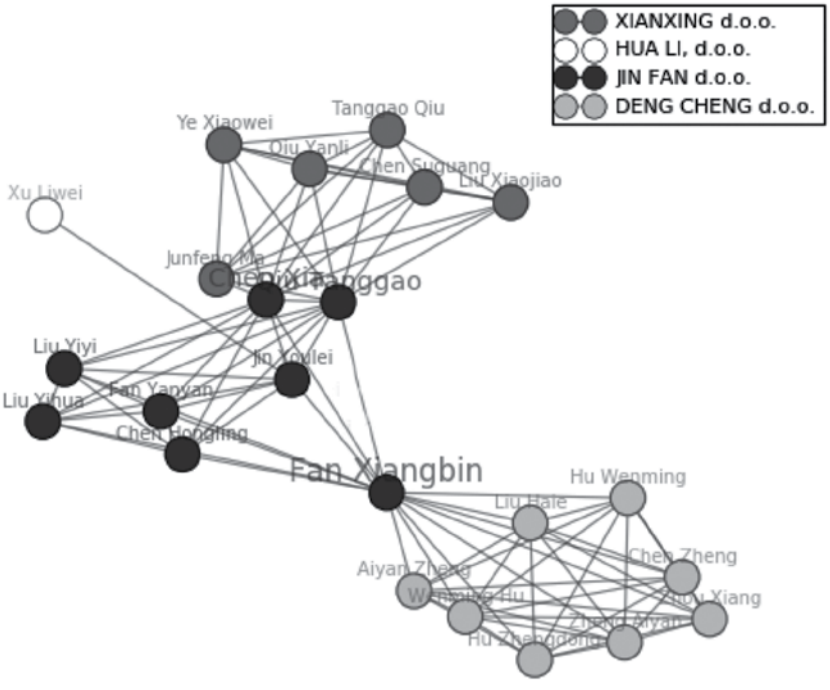

Figure 8: **Personal community with a single detected community with the updated visualization.**

for 2017. If we were to add the data from the previous years, the number of these edges would probably be higher. We also lack data on which we could build additional connections between individuals. In order to build an accurate representation, we would require data that would enable us to link people on a social level as it would be foolish to assume that such transactions only happen between individuals that were previously in business with each other. An example of the required data would be social information such as family members or friends. We also lack transactional information between private entities which would, without a doubt, be very informative, however such information is not of public nature.

The affiliation and transaction edges are much more common. In our network there are *1495876* affiliation connections and *255896* transactional connections. The large number of affiliation links is not surprising since each company is represented as a clique and therefore has the maximal number of edges between members of the same company.

5. DISCUSSION

We have gathered the transaction, ownership and representation data from Erar's (Commission for the Prevention of Corruption, 2018) database and converted it into a network of individuals active in the Slovene public spending system. We tested several community detection approaches with the goal of discovering small densely connected communities of people connected by either affiliation through employment at the same company or by a public transaction. The examined community detection methods return vastly different communities. Methods such as Infomap (Rosvall & Bergstrom, 2008), Louvain (Blondel, 2008) and Label Propagation (Cordasco & Gargano, 2010) can return very large communities that can be subdivided using algorithms such as METIS (Karypis & Kumar, 1998), however the resulting subdivision is often poor as further division is sometimes not appropriate as some individuals simply do business with many others which is why their community is proportionally larger.

It is difficult to evaluate how well each community detection algorithm performs on the network as we do not know what the actual communities are. We can check results for different individuals and see whether the returned community makes sense, but we cannot confidently state that a certain algorithm outperforms the others.

We are still facing issues with proper visualization as we have a lot of data that needs to be displayed in text such as names of individuals and companies and there is simply no space for a proper visualization when displaying a graph with more than 50 nodes or even less if the nodes are densely connected which they often are.

Our hope for this work was to discover significant smaller, tightly connected communities of individuals working in the entrepreneurial space of Slovenia, that are connected to public institutions. We were interested in seeing whether such communities exist and how well they are connected to individual public establishments. We also wanted to examine whether such communities could potentially collude to influence the results of public tenders offered by a specific organisation. We wanted to see whether there are representatives of public establishments that are also part of these communities.

We discovered that the network is indeed mostly constructed out of small connected communities however these communities are often large enough to cause issues with our visualization. There are also a few individuals who are very well connected and are therefore a part of larger communities that are very hard to properly visualize.

It is very hard to conclude anything about possible collusion between two representatives of public and private institutions since we lack the information to further connect business owners to public representatives. The only way to connect them in the context of our data is if both parties were once members of the same institutions and are now responsible for transactions between certain private and public institutions. As mentioned in Subsection 4.3, such connections are very rare and we would require data that is not of public nature to accurately identify representatives where the risk of collusion is higher.

An analysis of this sort could be much better if we had access to additional data and not just the transactional data of public companies. A government institution such as the Commission for the Prevention of Corruption would be much better suited to perform such research as more data is more readily available to them.

6. REFERENCES

[1] Ahn, Y.-Y., Bagrow, J. P., & Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. *Nature, 466*(7307), 761.

[2] Barabási, A.-L., & Pósfai, M. (2016). *Network Science.* Cambridge University Press. Pridobljeno iz http://barabasi.com/networksciencebook/

[3] Blondel, V. D.-L. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*. Pridobljeno iz https://arxiv.org/abs/0803.0476

[4] Commission for the Prevention of Corruption. (2018). *Erar API*. Pridobljeno iz https://erar.si/doc/: https://erar.si/doc/

[5] Commission for the Prevention of Corruption. (2018). *Erar, aplikacija za prikaz porabe javnega denarja v Republiki Sloveniji*. Pridobljeno iz https://erar.si/

[6] Cordasco, G., & Gargano, L. (2010). Community detection via semi-synchronous label propagation algorithms. *IEEE International Workshop on Business Applications of Social Network Analysis (BASNA)*, 1–8.

[7] Fortunato, S., & Barthelemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences, 104*(1), 36–41.

[8] Girvan, M. a. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences, 99*(12), 7821–7826.

[9] Karypis, G., & Kumar, V. (1998). A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on scientific Computing, 20*(1), 359–392.

[10] Kogovšek, R., Sovdat, B., & Povšič, R. (2013). Analysis of Slovene Company Ownership Network. *ASI 13/14 project*.

[11] Kolar, A., & Kolar, L. (2017). Resilience Analysis of the Slovene Economy. *ASI 17/18 project*.

[12] Lozano, S., Duch, J., & Arenas, A. (2006). Community detection in a large social dataset of european projects. Pridobljeno iz https://archive.siam.org/meetings/sdm06/workproceed/Link%20Analysis/17FP6-SIAM.pdf

[13] MMC RTV SLO. (2018). Je za izbiro dražje makete kriv Excel ali dogovarjanje med ponudniki? Pridobljeno iz https://www.rtvslo.si/gospodarstvo/je-za-izbiro-drazje-makete-kriv-excel-ali-dogovarjanje-med-ponudniki/448670

[14] Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences, 105*(4), 1118–1123.

[15] Šubelj, L., Jan, V. E., & Waltman, L. (2016). Clustering scientific publications based on citation relations: A systematic comparison of different methods. *PloS one, 11*(4), 0154404. Pridobljeno iz https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0154404

[16] Uprava Republike Slovenije za javna plačila. (2018). Seznami registra proračunskih uporabnikov. Pridobljeno iz https://www.ujp.gov.si/dokumenti/dokument.asp?id=127

[17] *Zakon o javnem narocanju (ZJN-3).* (2015). Ljubljana: Uradni list RS, num. 91, pg. 10201. Pridobljeno iz https://www.uradni-list.si/glasilo-uradni-list-rs/vsebina/2015-01-3570?sop=2015-01-3570

▪

Vitjan Zavrtanik končuje magistrski program računalništva in infromatike na Fakulteti za računalništvo in informatiko Univerze v Ljubljani. Zanimajo ga področja strojnega učenja, analize podatkov in računalniškega vida. Trenutno končuje magistrsko delo na temo semantične segmentacije slik. V preteklosti je med drugim sodeloval tudi v ekipi za distribuirano upravljanje s podatki v CERN-u na projektu Rucio.

▪

Lovro Šubelj je docent na Fakulteti za računalništvo in informatiko Univerze v Ljubljani. Diplomiral je leta 2008 na Fakulteti za matematiko in fiziko in Fakulteti za računalništvo in informatiko ter doktoriral leta 2013 na temo analize velikih omrežij. Je avtor ali soavtor več kot petdeset znanstvenih prispevkov in patentov ter urednik prestižnih mednarodnih znanstvenih revij. Njegovo preteklo delo je bilo izbrano kot izjemen znanstveni dosežek v Sloveniji ter predstavljeno na uglednih mednarodnih univerzah kot sta Stanford in UCSD. Sodeloval je že pri številnih uspešno zaključenih raziskovalnih in razvojnih projektih v sodelovanju s podjetji Petrol, Celtra, Optilab, Iskratel in drugimi.