Analyzing networks in Hungarian Agricultural subsidies

Manna Tóth
Gergo Szekely

2019.12.08

Motivation

We both feel our societies are deeply corrupted by the greed of politicians and oligarchs. They distort free markets, steal and waste precious resources of the people they are supposed to represent.

We wanted to work with a dataset that opens possibilities for future research in this area. We visited a Hungarian anti-corruption watchdog non-profit called K-monitor.[1] With their guidance we looked at various data sources that are not yet processed and could help journalists uncover malfeasance. Unfortunately, almost all of the publicly available datasets are difficult to work with because government agencies can't but more importantly don't want to provide clean data.

Agricultural Subsidies

One area of state corruption that has received news coverage in recent years is agricultural subsidies. After the transition of 1990 Hungarian oligarchs and politicians annexed a large part of the country. After the landslide victory of Viktor Orbán in the 2010 elections the trend continued on a whole new level.[2]

Both the European Union and National Governments spend a huge chunk of their budgets on supporting farmers. This is too big of a pie for politicians to ignore not just in Hungary but throughout the whole EU. As a result, most of the money does not go to those who actually work the land but rather to businessmen with good political connections who are only in the game because of these subsidies.[3]

Research Question

We wanted to see if the data supports the claim that the European Union finances oligarchs and politicians via agriculture subsidies.

[1] https://k-monitor.hu/fooldal

[2] https://www.nytimes.com/2019/11/03/world/europe/eu-farm-subsidy-hungary.html

[3] https://index.hu/belfold/2016/04/15/igy_nez_ki_felcsut_kornyeke_bevasaroltak_orbanek_baratai/

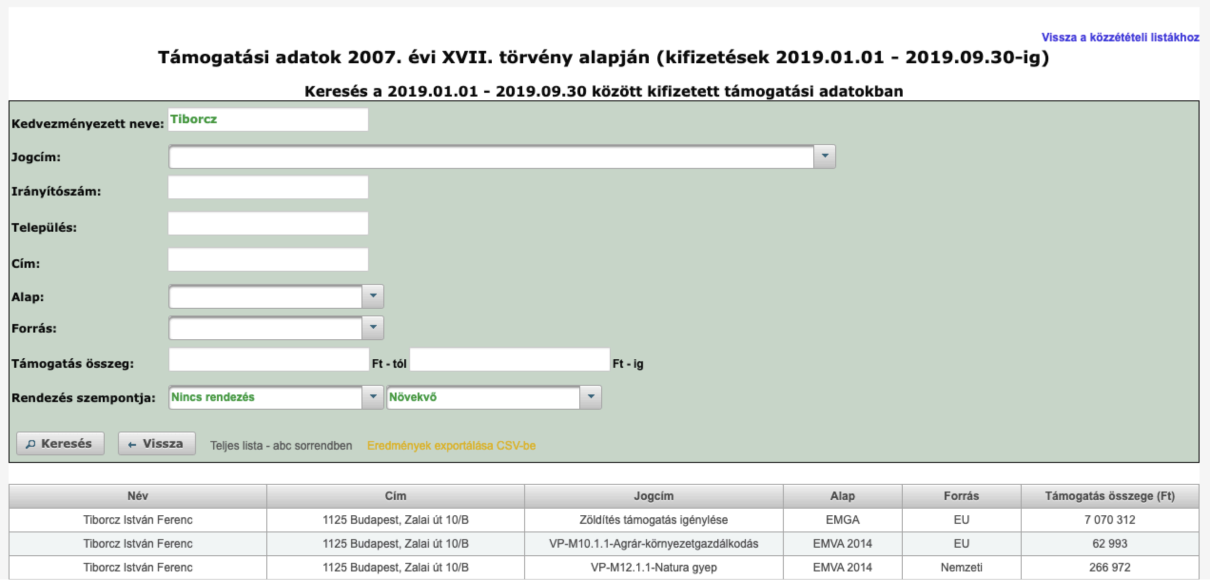
Data

Data about Hungarian agricultural subsidies are publicly available on the website of the Hungarian Treasury.[4]

Online Search Interface

The website provides a search interface but it has many limitations:

- The data can only be searched for a given year so trends cannot be explored
- Aggregations, more complex analysis cannot be performed
- The data also has a lot of typos
- Some individuals appear with different slight variations in their names
- Address formats are not standardized
- Firms and individuals are not distinguished


Vissza a közzétételi listákhoz
Keresés
Vissza
Teljes lista - abc sorrendben
Eredmények exportálása CSV-be

Támogatási adatok 2007. évi XVII. törvény alapján (kifizetések 2019.01.01 - 2019.09.30-ig)

Keresés a 2019.01.01 - 2019.09.30 között kifizetett támogatási adatokban

Kedvezményezett neve: Tiborcz

Jogcím:

Irányítószám:

Település:

Cím:

Alap:

Forrás:

Támogatás összeg: Ft - tól Ft - ig

Rendezés szempontja: Nincs rendezés Növekvő

Név	Cím	Jogcím	Alap	Forrás	Támogatás összege (Ft)
Tiborcz István Ferenc	1125 Budapest, Zalai út 10/B	Zöldítés támogatás igénylése	EMGA	EU	7 070 312
Tiborcz István Ferenc	1125 Budapest, Zalai út 10/B	VP-M10.1.1-Agrár-környezetgazdálkodás	EMVA 2014	EU	62 993
Tiborcz István Ferenc	1125 Budapest, Zalai út 10/B	VP-M12.1.1-Natura gyep	EMVA 2014	Nemzeti	266 972

Figure 1: Online search interface

Download and Process Raw data

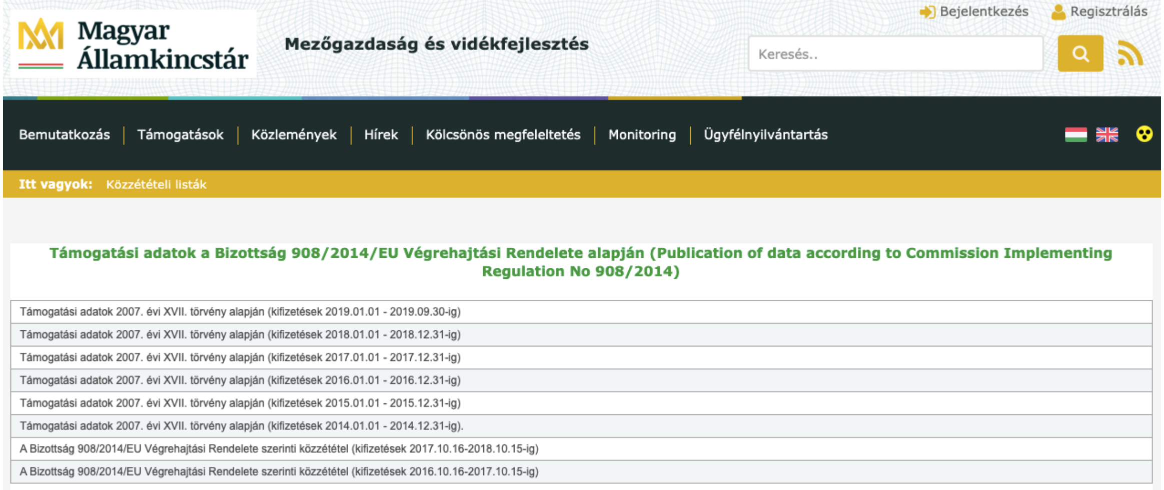
The website of the Hungarian Treasury allows users to download the raw data behind the search interface in CSV format. The site also allows scraping the data.[5]

Each file contains subsidies given to winners over one year. Each yearly CSV is between 47 and 75 MBs. Data covers the years between 2014 and 2019. This is quite a bit of information to process on a personal computer so to do basic data preparation and cleaning we used a very efficient programming language called Q.[6]

[4] https://www.mvh.allamkincstar.gov.hu/kozzeteteli-listak1

[5] https://www.mvh.allamkincstar.gov.hu/robots.txt

[6] https://code.kx.com/v2/

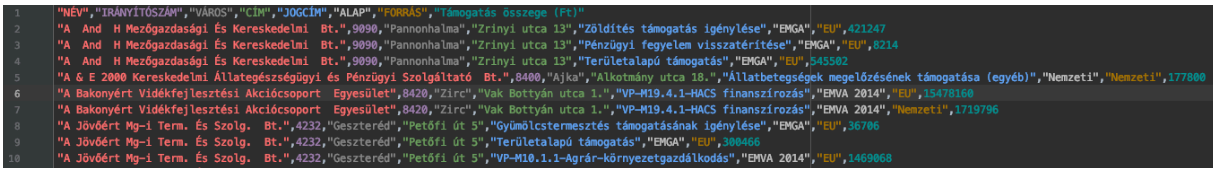


Támogatási adatok a Bizottság 908/2014/EU Végrehajtási Rendelete alapján (Publication of data according to Commission Implementing Regulation No 908/2014)

Támogatási adatok 2007. évi XVII. törvény alapján (kifizetések 2019.01.01 - 2019.09.30-ig)
Támogatási adatok 2007. évi XVII. törvény alapján (kifizetések 2018.01.01 - 2018.12.31-ig)
Támogatási adatok 2007. évi XVII. törvény alapján (kifizetések 2017.01.01 - 2017.12.31-ig)
Támogatási adatok 2007. évi XVII. törvény alapján (kifizetések 2016.01.01 - 2016.12.31-ig)
Támogatási adatok 2007. évi XVII. törvény alapján (kifizetések 2015.01.01 - 2015.12.31-ig)
Támogatási adatok 2007. évi XVII. törvény alapján (kifizetések 2014.01.01 - 2014.12.31-ig).
A Bizottság 908/2014/EU Végrehajtási Rendelete szerinti közzététel (kifizetések 2017.10.16-2018.10.15-ig)
A Bizottság 908/2014/EU Végrehajtási Rendelete szerinti közzététel (kifizetések 2016.10.16-2017.10.15-ig)

Figure 2: Download page of Hungarian agricultural subsidies

The character encoding of the raw files was Latin-2 so we had to convert the files to UTF-8.



```
"NÉV","IRÁNYÍTÓSZÁM","VÁROS","CÍM","JOGCÍM","ALAP","FORRÁS","Támogatás összege (Ft)"
"A  And  H Mezőgazdasági És Kereskedelmi  Bt.",9090,"Pannonhalma","Zrinyi utca 13","Zöldítés támogatás igénylése","EMGA","EU",421247
"A  And  H Mezőgazdasági És Kereskedelmi  Bt.",9090,"Pannonhalma","Zrinyi utca 13","Pénzügyi fegyelem visszatérítése","EMGA","EU",8214
"A  And  H Mezőgazdasági És Kereskedelmi  Bt.",9090,"Pannonhalma","Zrinyi utca 13","Területalapú támogatás","EMGA","EU",545502
"A & E 2000 Kereskedelmi Állategészségügyi és Pénzügyi Szolgáltató  Bt.",8400,"Ajka","Alkotmány utca 18.","Állatbetegségek megelőzésének támogatása (egyéb)","Nemzeti","Nemzeti",177800
"A Bakonyért Vidékfejlesztési Akciócsoport  Egyesület",8420,"Zirc","Vak Bottyán utca 1.","VP-M19.4.1-HACS finanszírozás","EMVA 2014","EU",15478160
"A Bakonyért Vidékfejlesztési Akciócsoport  Egyesület",8420,"Zirc","Vak Bottyán utca 1.","VP-M19.4.1-HACS finanszírozás","EMVA 2014","Nemzeti",1719796
"A Jövőért Mg-i Term. És Szolg.  Bt.",4232,"Geszteréd","Petőfi út 5","Gyümölcstermesztés támogatásának igénylése","EMGA","EU",36706
"A Jövőért Mg-i Term. És Szolg.  Bt.",4232,"Geszteréd","Petőfi út 5","Területalapú támogatás","EMGA","EU",300466
"A Jövőért Mg-i Term. És Szolg.  Bt.",4232,"Geszteréd","Petőfi út 5","VP-M10.1.1-Agrár-környezetgazdálkodás","EMVA 2014","EU",1469068
```

Figure 3: sample of the raw CSV

We manually prepared a keyword list from the dataset that includes dozens of phrases like “firm”, “LTD”, “Corporation”, etc… This seems to split the data quite to individuals and firms so we could add a Boolean variable to the data to store this.

All of the code that we used for the data preparation and analysis is available on GitHub.[7]

[7] https://github.com/gergo/hungarianAgricultureSubsidies

Interesting Findings

Once the data is loaded into memory the Q programming language allows simple SQL-like queries that highlight potential malfeasance. Such analysis is not part of this exercise but we found this part very exciting.

Address sharing

We were surprised to see that no firm shares an address with an individual as the below query returns no records.

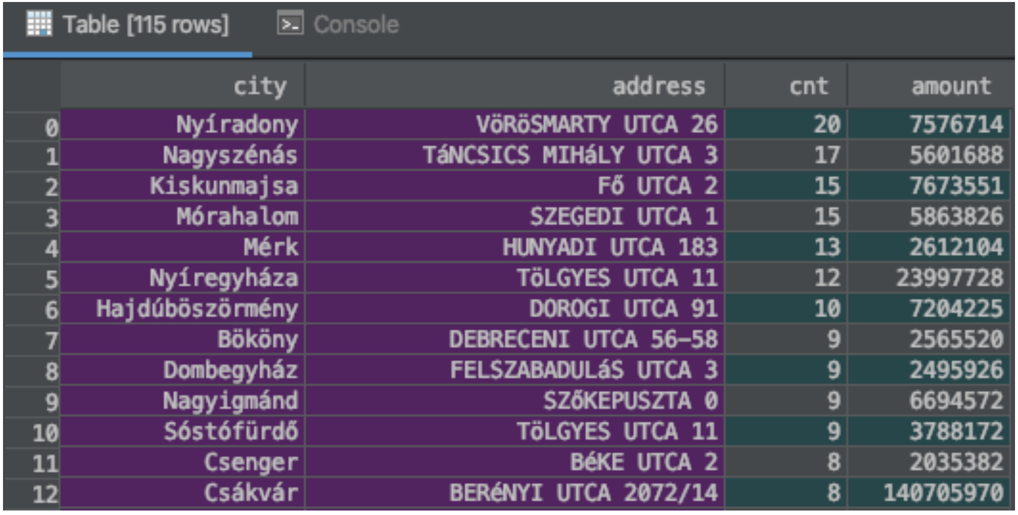
```
// Are there individuals and firms that share address?
(`zip`city`address xkey .agrar.firms) ij `zip`city`address xkey .agrar.ppl
```

Co-Location

It was obvious from the raw data that some people share an address. Let's see what addresses are shared by the most people.

```
// which households contain the most winners (along with the amounts)
select from (`cnt xdesc select nm: enlist name, cnt: count i,sum amount by
city,address from select sum amount by name,city,address from .agrar.ppl where
address<>`) where cnt>5
```

There are 115 cases where more than 5 distinct individuals sharing the same address won money. We did not try to see what was behind these but these look quite unrealistic.

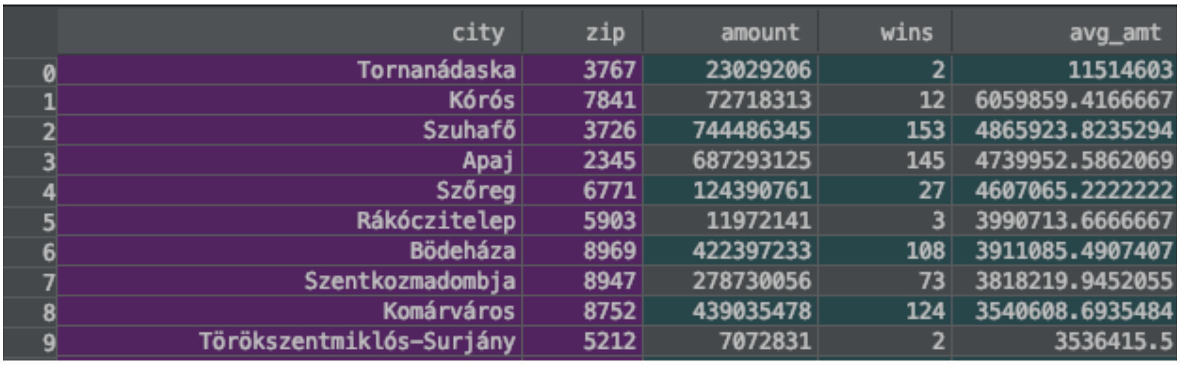


	city	address	cnt	amount
0	Nyíradony	VöRöSMARTY UTCA 26	20	7576714
1	Nagyszénás	TáNCSICS MIHáLY UTCA 3	17	5601688
2	Kiskunmajsa	Fő UTCA 2	15	7673551
3	Mórahalom	SZEGEDI UTCA 1	15	5863826
4	Mérk	HUNYADI UTCA 183	13	2612104
5	Nyíregyháza	TöLGYES UTCA 11	12	23997728
6	Hajdúböszörmény	DOROGI UTCA 91	10	7204225
7	Bököny	DEBRECENI UTCA 56-58	9	2565520
8	Dombegyház	FELSZABADULáS UTCA 3	9	2495926
9	Nagyigmánd	SZŐKEPUSZTA 0	9	6694572
10	Sóstófürdő	TöLGYES UTCA 11	9	3788172
11	Csenger	BéKE UTCA 2	8	2035382
12	Csákvár	BERéNYI UTCA 2072/14	8	140705970

Figure 4: Co-locating winners

Biggest winners

```
// Residents of which town won the largest amount of subsidies
`avg_amt xdesc update avg_amt: amount%wins from select sum amount, wins: count i by
city,zip from .agrar.ppl
```



	city	zip	amount	wins	avg_amt
0	Tornanádaska	3767	23029206	2	11514603
1	Kórós	7841	72718313	12	6059859.4166667
2	Szuhafő	3726	744486345	153	4865923.8235294
3	Apaj	2345	687293125	145	4739952.5862069
4	Szőreg	6771	124390761	27	4607065.2222222
5	Rákóczitelep	5903	11972141	3	3990713.6666667
6	Bödeháza	8969	422397233	108	3911085.4907407
7	Szentkozmadombja	8947	278730056	73	3818219.9452055
8	Komárváros	8752	439035478	124	3540608.6935484
9	Törökszentmiklós-Surjány	5212	7072831	2	3536415.5

Networks

We focused on networks of individuals, not firms. We mapped people with identical names and addresses to create aggregate data points where we summed up all of the subsidies they received over the covered time period. These are our nodes. Our edges are relationships between the nodes based on the similarity of names and addresses.

Address Score

One component of the edge strength is the similarity of addresses. If they match exactly we gave them 10 points. This covers individuals who share the same address or people who appear in the raw data with multiple names. Sometimes the last part of the address field had different variations ("... UTCA 2/A" vs. "… UTCA 2A"). If the addresses didn't math exactly we removed the last part and checked if the remainder matched. This gave a low score of 1 point to those edges where the address had a typo and also for individuals who live in the same street. We think this introduces some false positives but that is not a real problem because in a later step we filter those records where the overall score is at least 3. If this was the only connection then they will not appear in the final network. Also, in the country side in small villages most people know each other so there might indeed be some weak relationship between people who live in the same street.

Name Score

The other component of the strength of an edge came from the similarity of names. We calculate the similarity score and weight it by the frequency of the names. This way if two names have a partial match but they are infrequent we give them higher score than for the same kind of match if it appears between very frequent names.

The unweighted score gives 10 points if the names match exactly. Our algorithm gives 9 points if one name is the prefix of the other. This covers cases where the wife took the name of the husband ("Mészáros Lőrinc" and „Mészáros Lőrincné"). It also covers cases where someone has a second given name but it does not appear in all records. Then we check if the first part of the name matches (family name). This yields 5 points according to our algorithm. As the last step we remove all legal given names by downloading it from the Hungarian Academy of Sciences site and applying a modified version of the Levenshtein distance on what remains[8,9].

Levenshtein distance is the smallest number of character replacements that are needed to convert one string to another. We didn't check all possibilities; instead we stopped the algorithm after 3 steps. This is enough to find typos and slight variations of names but still keeps the search space traversable for the hundreds of thousands of edges that we have to check. The algorithm gives low score to partial matches so this will be filtered out later when we drop very weak connections.

Full Network

One of the networks that we analyse is based on the whole dataset. Unfortunately, processing all data points would be too much for a PC. Many of the data points are very low amounts won by individuals so to arrive at a manageable amount of data we apply 2 cutoff points. First, we drop individual wins that are smaller than 500 000 HUF (1670 USD). We get about 700 000 records after applying this rule. We also drop entries from the aggregate table where the total

[8] http://www.nytud.mta.hu/oszt/nyelvmuvelo/utonevek/

[9] https://en.wikipedia.org/wiki/Levenshtein_distance

amount of money won by the individual is less than 1 000 000 HUF. The aggregate dataset has 100 000 records. To find all possible relationships we would have to do a self-join on this dataset. This would mean doing about 5 billion name and address comparisons. It is doable but was too much for the scope of this analysis. Also, almost all of the relationships would have been very weak. On the full network we apply a simple heuristic: check relationships of people who have the same zip code. This seems much more realistic and yields a smaller network of 7 000 000 edges. Once we drop weak relationships (anything with a score smaller than 3) we end up with 90 000 edges.

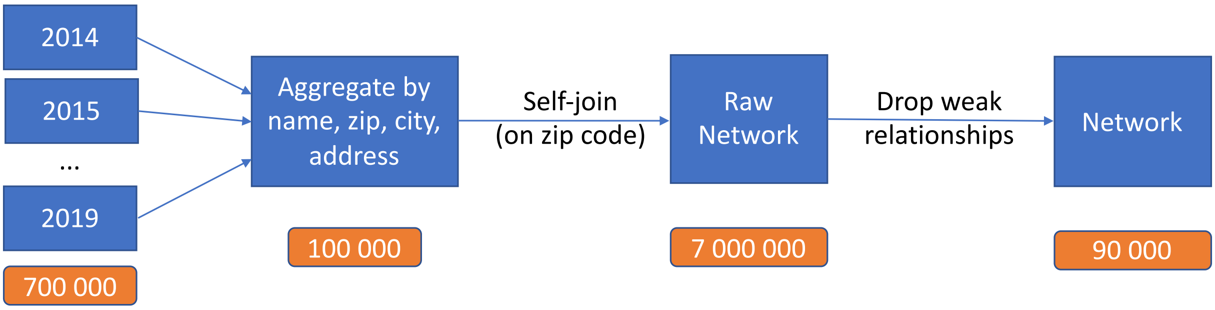


Figure 5: Full network on the whole dataset

Felcsút Network

The other network we analyze is a politically interesting area in Hungary, the home town of Viktor Orbán and the surrounding areas. Felcsút is a village close to the capital with about 2000 inhabitants. The area has seen an incredible inflow of capital since Orbán took over the country following a landslide victory in 2010. His childhood friend Lőrinc Mészáros got all of his wealth via state contracts and became the richest Hungarian man by 2018 [10]. The family members of Orbán and Mészáros also experienced incredible increase in their assets.

Figure 6: Viktor Orbán's house and a tax-payer funded, under-utilized soccer stadium across the street

When constructing the network, we filtered on zip codes of Felcsút and the surrounding villages. This data set is small enough to fit into memory and do a full join analysis in less than a minute.

[10] https://bbj.hu/business/meszaros-becomes-richest-hungarian_159451

Results

Full Network

The graph[11] below shows the distribution of scores (the strength of relationships by our calculation). The values between 0-5 indicate very weak relationships, these scores can be only the results of similar or same names. Values between 6-10 can show weak relationship: neighbors or distant relatives. Scores higher than 16 indicate strong family ties (people living in the same household, marriages, siblings or father-children).

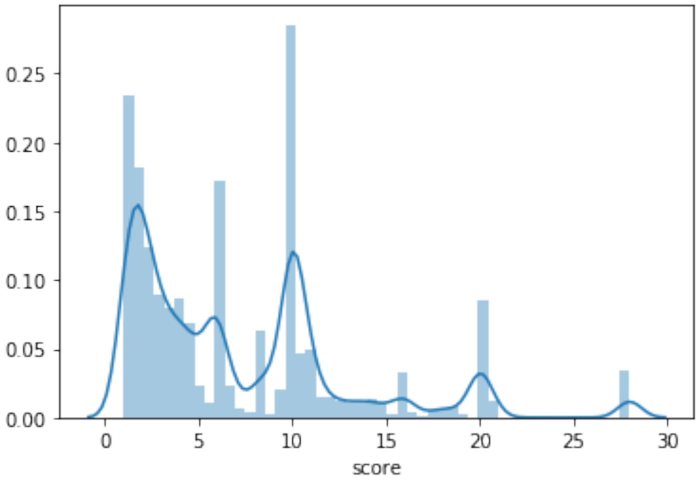
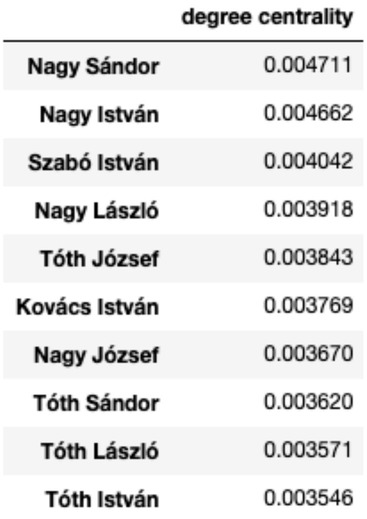


Figure 7: Distribution of scores

The graph of the nodes with the 10 largest degree centrality[12] shows that common Hungarian surnames (Kovács, Nagy, Szabó, Tóth) and first names (Sándor, István, László, József) could biased our results in its present form for as far as the whole dataset is concerned.



	degree centrality
Nagy Sándor	0.004711
Nagy István	0.004662
Szabó István	0.004042
Nagy László	0.003918
Tóth József	0.003843
Kovács István	0.003769
Nagy József	0.003670
Tóth Sándor	0.003620
Tóth László	0.003571
Tóth István	0.003546

Figure 8: nodes with the 10 largest degree centrality in the whole dataset

[11] An updated, improved dataset was used for the paper, so the data and the graphs are slightly different from the ones used for the presentation.

[12] Codes of betweenness and closeseness centrality are in the script, but due to the large size of dataset a standard MacBoor Air's capacity was not sufficient to calculate these statistics.

Strong and weak ties

We tried to test Granovetter's strong and weak tie theory in our dataset for the class presentation.

We calculated the neighborhood overlap based on a binary variable created by the scores where score values under 3 were defined as no relationship (0) and scores greater than 3 meant that there is a relationship (1). We defined a tie weak if its neighborhood overlap value was smaller than 0.6 and values with 0.6 or higher were counted as strong ties. Interestingly, - as the graph shows below – there are higher-scored relationships in the weak tie category than the strong-tie category. The average score was 4.44 among the weak ties and 4.19 among the strong ties. These results were the opposite of what we expected; we assumed that stronger ties will come with higher scores.

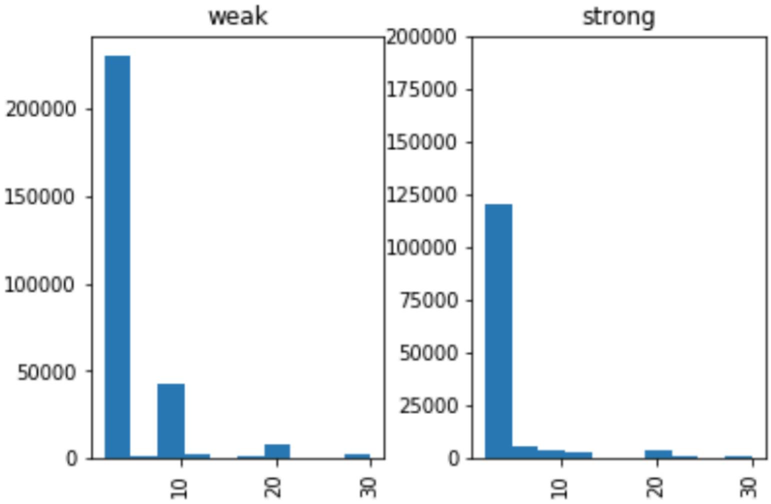


Figure 9: Distribution of scores by weak and strong ties

We have tested our data from the other way around: the value of neighborhood overlaps by scores. We have divided the data into two score categories, scores lower than 16 were defined as low and scores 16 or higher were taken as high. As the graph below shows, lower scored values have more relationships (on average 0.43) than higher-scored values (on average 0.34).

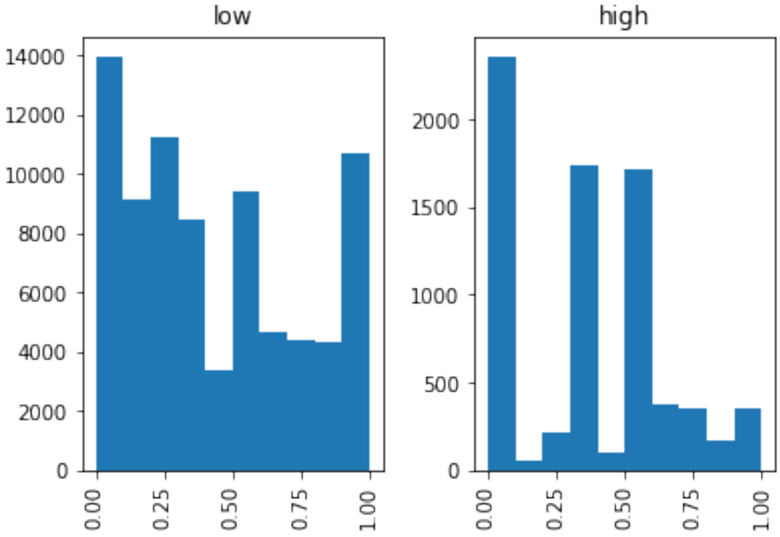


Figure 10: Distribution of neighbourhood overlaps by low and high scores

Conclusion

We believe that the reason for the difference between the strength of ties and the scores can be that our network is defined based on different principles than other social networks we studied. At Granovetter's (1973) the strengths of ties depend on 1) time spend together, 2) emotional intensity, 3) intimacy and 4) reciprocal services. Granovetter's strong and weak ties are mostly measured by asking people (nodes) to mark the strength of their relationship with certain other people (nodes). We defined the network we are analyzing based on quantitative measures, which are mostly related to the name and address of people. Thus, the nature of our network is quite different. Furthermore, in our opinion the indicators we used are more suitable to detect Granovetter's strong ties as we can mostly identify family relationships. Weak ties are impossible to catch by our methodology. So, we can argue that relationships in our dataset – which are not biased by common family and given names – mostly show the strength of Granovetter's strong ties.

Felcsút

There are 82 nodes and 89 edges of network of Felcsút, with and average 2,19 degrees. The distribution of the network of Felcsút is similar to the distribution of the whole country, it seems that there are fewer lower scored relationship as the minimum of the data is higher here. The density of this network is also higher than the one of the whole country's: it is 0.027 in Felcsút and 0.00008 in the whole dataset. (These differences can be related to the fact that Felcsút is a small village and the number of nodes differ significantly in the two datasets).

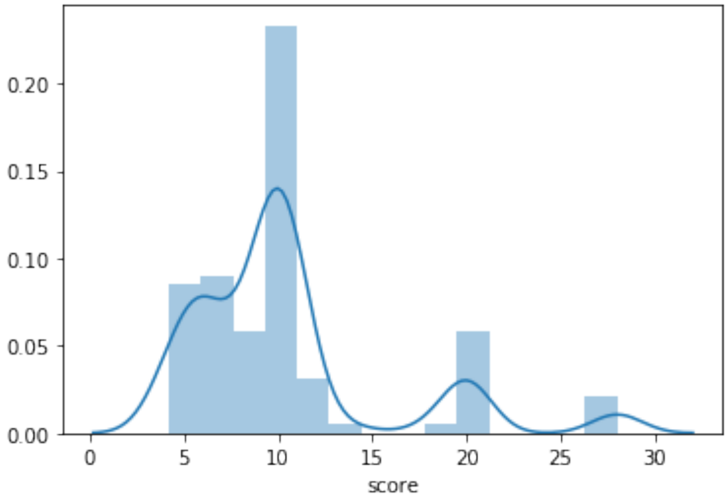


Figure 11 Distribution of scores in Felcsút

The graph [13]of the network of Felcsút also suggest that it is a dense network, a great majority of the nodes are connected to the whole network. There are some within networks (the triangles and the cubes) and the greatest within network – marked with orange at the left side of the graph -– is the one with the Mészáros's are associated with, and almost all the nodes with the highest centrality measures are also part of this network. In this paper we do not have a chance to further investigate this network, but we assume that these nodes could be the most interesting for investigative journalist or corruption researchers since it slightly

[13] The graph is attached as pdf as well, for better visibility

shows a different pattern from the rest of the village and it seems to have a quite important role (and part) in the agricultural subsidies of the village.

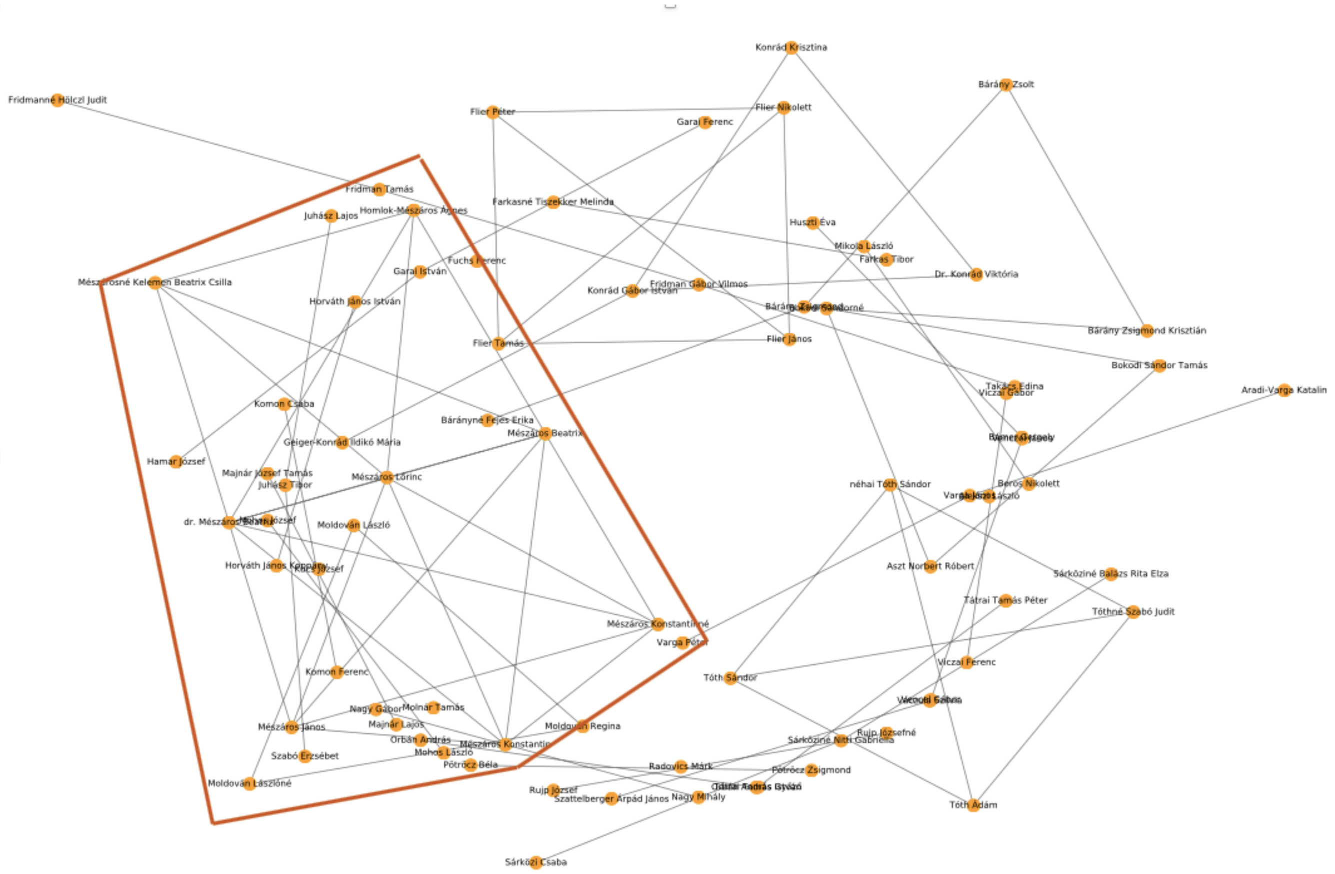
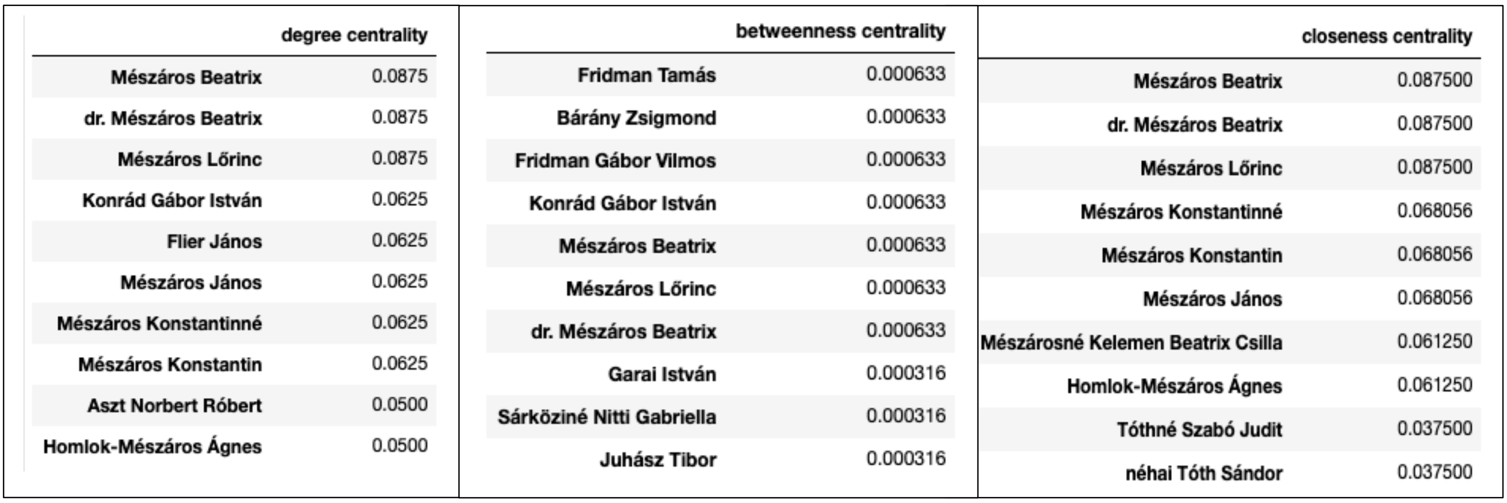


Figure 12 Network of Felcsút

Three types of centrality measures were calculated for Felcsút: degree, betweenness and closeness centralities. The graph below shows the nodes with the 10 largest values.



	degree centrality
Mészáros Beatrix	0.0875
dr. Mészáros Beatrix	0.0875
Mészáros Lőrinc	0.0875
Konrád Gábor István	0.0625
Flier János	0.0625
Mészáros János	0.0625
Mészáros Konstantinné	0.0625
Mészáros Konstantin	0.0625
Aszt Norbert Róbert	0.0500
Homlok-Mészáros Ágnes	0.0500

	betweenness centrality
Fridman Tamás	0.000633
Bárány Zsigmond	0.000633
Fridman Gábor Vilmos	0.000633
Konrád Gábor István	0.000633
Mészáros Beatrix	0.000633
Mészáros Lőrinc	0.000633
dr. Mészáros Beatrix	0.000633
Garai István	0.000316
Sárköziné Nitti Gabriella	0.000316
Juhász Tibor	0.000316

	closeness centrality
Mészáros Beatrix	0.087500
dr. Mészáros Beatrix	0.087500
Mészáros Lőrinc	0.087500
Mészáros Konstantinné	0.068056
Mészáros Konstantin	0.068056
Mészáros János	0.068056
Mészárosné Kelemen Beatrix Csilla	0.061250
Homlok-Mészáros Ágnes	0.061250
Tóthné Szabó Judit	0.037500
néhai Tóth Sándor	0.037500

Figure 13 nodes with the 10 largest degree, betweenness and closeness centralities in Felcsút

We can see the dominance of Mészáros family among all the centrality measures. They have the highest number of relationships (degree centralities) and they can get to the closest way to other network members of Felcsút (closeness centrality). It shows how important their role is in the network of agricultural subsidies of the village. We can find other people among the ones with the highest betweenness centrality, it can show that there might be other important brokers between the different within networks of Felcsút. These other people might be “simple” people from Felcsút, who are not involved into politics, but they are more popular or have more local ties.

Conclusion

To sum up, we can conclude that it is possible to catch certain types of Granovetter’s strong ties - which are associated with similarities in address and name - by using certain computations. This methodology can be useful in case of corruption research where data availability is usually an issue.

Our examination is not broad enough to certify corruption, but we can say that based on our preliminary results and basic descriptive statistics, the importance of Mészáros family in Felcsút and the network of Felcsút comparing to the whole country is definitely worth further investigating.

Further research

We put a lot of time and effort to the project to achieve its current state. We defined a number of improvement ideas:

- join agricultural subsidies with election data to see what influence municipal leadership has on wins
- refine ZIP code matching: big cities have multiple zip codes
- join data with population density to see how much money is won per resident in each city
- create an interactive gui so non-technical people can play around with the data. Ideally this would be a javascript-based pivot table[14] with Q or Python serving the data
- find a way to handle data bias due to common surnames and given names.

[14] https://pivottable.js.org/examples/index.html