# Supplementary Material

## Political Astroturfing on Twitter: How to Coordinate a Disinformation Campaign

Franziska B. Keller, David Schoch, Sebastian Stier, JungHwan Yang

*Political Communication*, 2019

## A1 Activity of NIS accounts and suspect accounts

Figure A1 shows the activity of NIS accounts from June 2012 to December 2012 together with a clustering derived from a block clustering algorithm. The algorithm identifies accounts that have very similar activity patterns – i.e., they tweet on the same days – and groups them together (Hartigan, 1972). For methodological details, see Keller, Schoch, Stier, and Yang (2017).

Accounts are grouped according to four patterns of activity (see also Keller et al., 2017): the accounts at the bottom of the plot are most distinct from the rest in that they are very active and do not necessarily cease tweeting after December 11. A manual inspection shows that they are at least partly automated and focus on tweeting news headlines. It is possible that the principals forgot or were not able to shut them down or thought they were innocuous enough to continue tweeting through the election.

The other three groups focus on retweeting other accounts: a small group is active from the beginning of the observed period until September 1 (top of figure). A larger group starts or massively increases its activity on September 1 and shuts down on December 11 (green marked accounts).

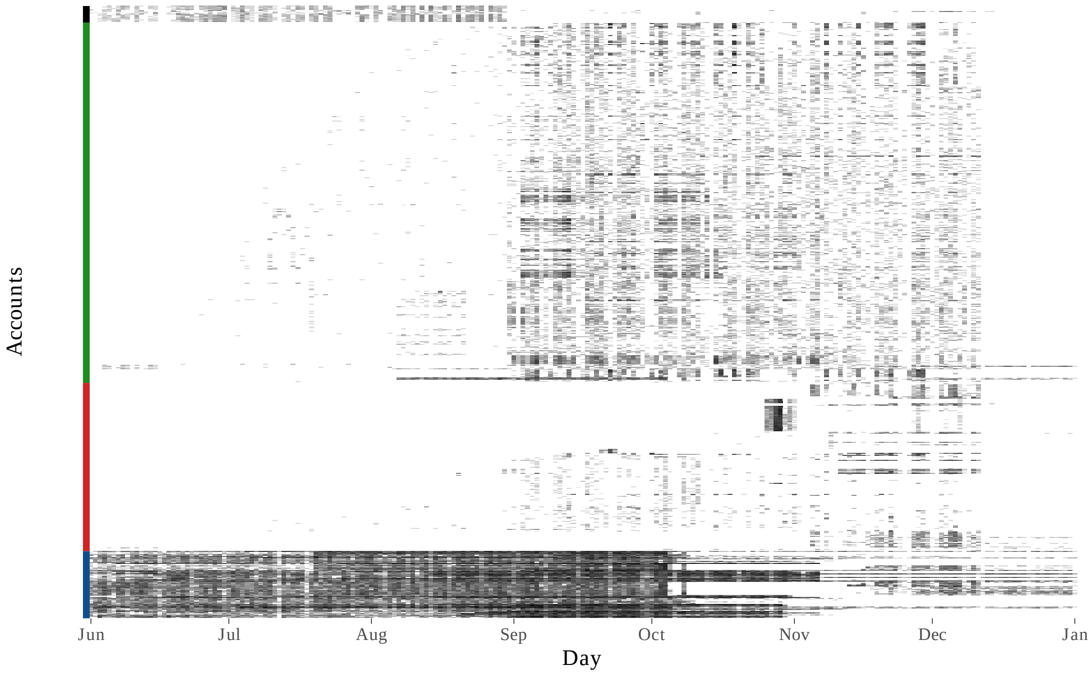

Figure A1: Activity of NIS accounts. Each line corresponds to the activity of a single account. The darker the cell of a given day, the more tweets were sent. The colored lines indicate cluster membership.

We also applied the block clustering technique to the 834 suspect accounts identified (Figure A2). Most suspect accounts have similar activity patterns as the semi-automated known NIS accounts and cease to tweet in the first week of October or beginning of November (top part of the plot). Interspersed are a few accounts that also tweet from the beginning, but stop around October 1, which resembles the pattern of a small group of known NIS accounts as well. A small group of suspects follows the most common pattern among known NIS accounts, starting on September 1 and stopping on December 11. In addition, there a considerable number of suspect accounts (black cluster) that follow a pattern of a handful of known NIS accounts which start tweeting at the end of the first week of November, and either stop or continue on December 11.

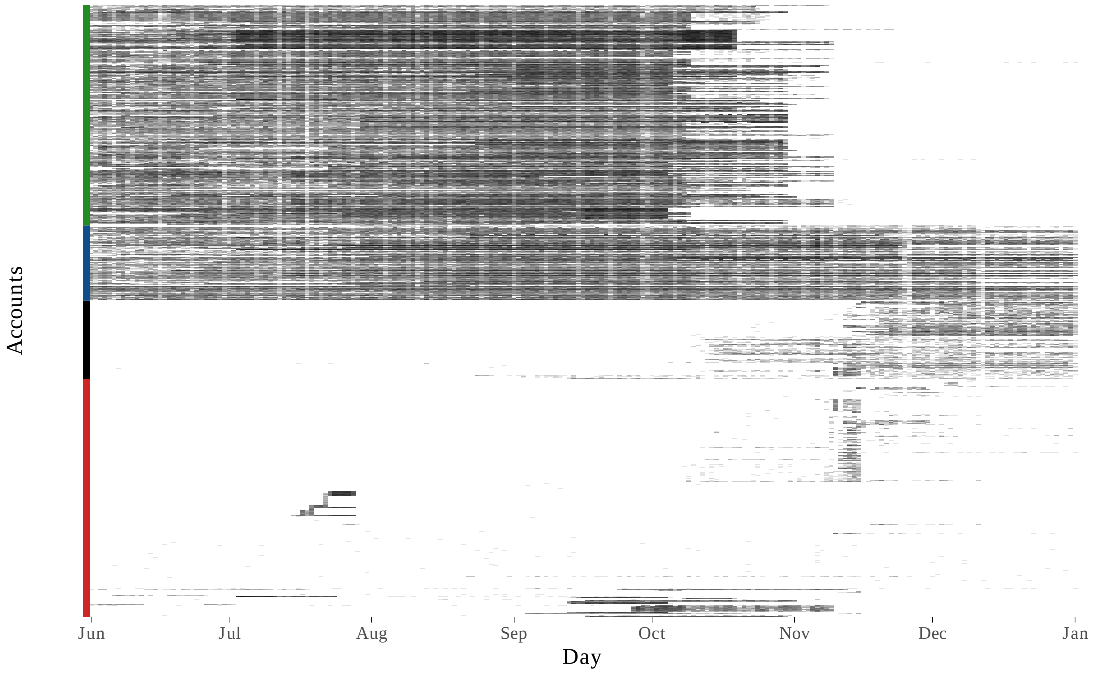

Figure A2: Activity of suspect accounts. Each line corresponds to the activity of a single account. The darker the cell of a given day, the more tweets were sent. The colored lines indicate cluster membership.

## A2 Overlap and similarities between different detection methods

Table A1 shows the overlap between the different detection methods. The overlap is moderate, but it should be kept in mind that the three different methods are designed to discover three different message coordination strategies, and that the accounts do not all choose the same strategy. The semi-automated accounts tend to focus on co-tweeting, while other accounts almost only retweet other accounts. The three methods are therefore complements instead of substitutes.

Table A1: Overlap among the three identification methods: number of suspects detected. Row percentages.

| | retweet | co-tweet | co-retweet |
|---|---:|---:|---:|
| retweet | 204 (100%) | 91 (45%) | 121 (59%) |
| co-tweet | 91 (14%) | 647 (100%) | 290 (44%) |
| co-retweet | 121 (28%) | 290 (67%) | 432 (100%) |

# A3 Additional results from the text analysis

Table A2: 50 most popular keywords used among NIS accounts, suspect accounts, random sample, political sample, and opinion leaders. Proportions are caculated after preprocessing texts of 10,000 randomly sampled tweets per group.

| Rank | NIS | Prop | NIS Suspects | Prop | Random sample | Prop | Political sample | Prop | Opinion leaders | Prop |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | North Korea(북한) | 1.10 | Journalist (기자) | 0.75 | lol (ㅋㅋ) | 1.15 | GeunHye Park (박근혜) | 1.49 | GeunHye Park (박근혜) | 0.72 |
| 2 | Presidential election (대선) | 0.41 | Presidential election (대선) | 0.73 | crying emoji (ㅜㅜ) | 0.41 | JaeIn Moon (문재인) | 1.12 | Candidate (후보) | 0.51 |
| 3 | Korea (한국) | 0.38 | Saenuri (새누리) | 0.66 | Bot | 0.29 | ChulSoo Ahn (안철수) | 1.07 | ChulSoo Ahn (안철수) | 0.46 |
| 4 | Seoul (서울) | 0.36 | Candidate (후보) | 0.65 | GeunHye Park (박근혜) | 0.25 | Candidate (후보) | 0.97 | Vote (투표) | 0.45 |
| 5 | ChulSoo Ahn (안철수) | 0.36 | Seoul (서울) | 0.60 | I am (난) | 0.23 | President (대통령) | 0.60 | JaeIn Moon (문재인) | 0.42 |
| 6 | North Korea followers (종북) | 0.35 | North Korea(북한) | 0.53 | I am (내가) | 0.23 | Presidential election (대선) | 0.53 | Saenuri (새누리) | 0.33 |
| 7 | President (대통령) | 0.34 | GeunHye Park (박근혜) | 0.50 | @ | 0.21 | Citizens (국민) | 0.52 | Politics (정치) | 0.32 |
| 8 | Candidate (후보) | 0.31 | Congressman (의원) | 0.47 | Me (나) | 0.19 | Saenuri (새누리) | 0.48 | Citizens (국민) | 0.30 |
| 9 | Saenuri (새누리) | 0.31 | President (대통령) | 0.46 | lol (ㅎㅎ) | 0.19 | Vote (투표) | 0.45 | Presidential election (대선) | 0.27 |
| 10 | Journalist (기자) | 0.28 | ChulSoo Ahn (안철수) | 0.44 | JaeIn Moon (문재인) | 0.19 | Politics (정치) | 0.40 | lol (ㅋㅋ) | 0.26 |
| 11 | China (중국) | 0.25 | Minjutonghapdang (민주통합당) | 0.43 | For real (진짜) | 0.17 | Minjoodang (민주당) | 0.33 | President (대통령) | 0.25 |
| 12 | Japan (일본) | 0.25 | Korea (한국) | 0.38 | Candidate (후보) | 0.17 | Congressman (의원) | 0.22 | Korea (대한) | 0.20 |
| 13 | Citizens (국민) | 0.25 | Congress (국회) | 0.34 | ChulSoo Ahn (안철수) | 0.17 | Republic of Korea (대한민국) | 0.21 | Today (오늘) | 0.18 |
| 14 | GeunHye Park (박근혜) | 0.24 | This day (이날) | 0.32 | Today (오늘) | 0.16 | MooHyun Roh (노무현) | 0.20 | Minjoodang (민주당) | 0.17 |
| 15 | Congressman (의원) | 0.24 | Afternoon (오후) | 0.30 | Now (지금) | 0.14 | North Korea (북한) | 0.19 | Seoul (서울) | 0.16 |
| 16 | Government (정부) | 0.23 | JaeIn Moon (문재인) | 0.30 | Very (너무) | 0.13 | Congress (국회) | 0.19 | JungHee Park (박정희) | 0.15 |
| 17 | Politics (정치) | 0.23 | Primaries (경선) | 0.27 | RT | 0.13 | North Korea followers (종북) | 0.18 | Korea (한국) | 0.15 |
| 18 | U.S.A. (미국) | 0.22 | Government (정부) | 0.27 | Me too (나도) | 0.11 | Seoul (서울) | 0.18 | lol (ㅎㅎ) | 0.13 |
| 19 | Republic of Korea (대한민국) | 0.22 | Japan (일본) | 0.27 | President (대통령) | 0.11 | Korea (대한) | 0.17 | Congressman (의원) | 0.13 |
| 20 | JungEun Kim (김정은) | 0.20 | Past (지난) | 0.24 | Me (저) | 0.11 | JungHee Park (박정희) | 0.17 | I am (제가) | 0.12 |
| 21 | Military (군) | 0.20 | The North (北) | 0.23 | Vote (투표) | 0.11 | #PresidentialElection (#대선) | 0.17 | Most (가장) | 0.11 |
| 22 | Minjutonghapdang (민주통합당) | 0.18 | U.S.A. (미국) | 0.23 | crying emoji (ㅜㅜ) | 0.11 | Economy (경제) | 0.15 | Congress (국회) | 0.11 |
| 23 | Past (지난) | 0.18 | Morning (오전) | 0.22 | Korea (한국) | 0.11 | lol (ㅋㅋ) | 0.15 | Prosecutor (검찰) | 0.11 |
| 24 | JaeIn Moon (문재인) | 0.18 | Delegate (대표) | 0.22 | Now (이제) | 0.10 | Busan (부산) | 0.13 | I am (내가) | 0.11 |
| 25 | Korea (대한) | 0.18 | Politics (정치) | 0.22 | Citizens (국민) | 0.10 | Minjootonghapdang (민주통합당) | 0.13 | Nation (국가) | 0.10 |
| 26 | Congress (국회) | 0.18 | Stated (밝혔다) | 0.21 | Seoul (서울) | 0.09 | http | 0.13 | Now (지금) | 0.10 |
| 27 | The North (北) | 0.17 | Seoul=Yeonhap (서울=연합뉴스) | 0.21 | I am (나는) | 0.09 | Korea (한국) | 0.13 | MooHyun Roh (노무현) | 0.10 |
| 28 | Nation (국가) | 0.17 | China (중국) | 0.21 | Us (우리) | 0.09 | Today (오늘) | 0.13 | Very (너무) | 0.10 |
| 29 | Economy (경제) | 0.17 | Citizens (국민) | 0.20 | For real (정말) | 0.08 | Nation (국가) | 0.11 | I am (저는) | 0.10 |
| 30 | Afternoon (오후) | 0.15 | Police (경찰) | 0.20 | Politics (정치) | 0.08 | Us (우리) | 0.11 | Us (우리) | 0.10 |
| 31 | This day (이날) | 0.15 | Military (군) | 0.19 | Japan (일본) | 0.08 | Joint nomination (단일화) | 0.10 | Economy (경제) | 0.10 |
| 32 | Dokdo (독도) | 0.14 | Yeolin (열린) | 0.19 | ^ | 0.08 | Military (군) | 0.10 | 이제 | 0.10 |
| 33 | Our (우리) | 0.14 | Korea (대한) | 0.18 | I am (제가) | 0.08 | #GeunHyePark (#박근혜) | 0.10 | Together (함께) | 0.10 |
| 34 | Police (경찰) | 0.13 | Seoul=Newsis (서울=뉴시스) | 0.18 | What (뭐) | 0.08 | Government (정부) | 0.10 | For real (정말) | 0.10 |
| 35 | Recently (최근) | 0.13 | Charge (혐의) | 0.17 | 2 | 0.08 | 1 | 0.10 | Same (같은) | 0.10 |
| 36 | Delegate (대표) | 0.12 | Recently (최근) | 0.16 | Again (다시) | 0.07 | Now (이제) | 0.10 | Japan (일본) | 0.10 |
| 37 | #kocon | 0.12 | Prosecutor (검찰) | 0.16 | Presidential election (대선) | 0.07 | MyungBak Lee (이명박) | 0.09 | Again (다시) | 0.09 |
| 38 | JungIl Kim(김정일) | 0.12 | Economy (경제) | 0.16 | Good (좋은) | 0.07 | Camp (캠프) | 0.09 | *postposition* (겁니다) | 0.09 |
| 39 | Morning (오전) | 0.12 | Dokdo (독도) | 0.15 | ~ | 0.07 | Now (지금) | 0.09 | Military (군) | 0.09 |
| 40 | Charge (혐의) | 0.12 | Busan (부산) | 0.15 | Honey (오빠) | 0.07 | This time (이번) | 0.09 | Government (정부) | 0.09 |
| 41 | Primaries (경선) | 0.11 | MyongBak Lee (이명박) | 0.14 | I am (저는) | 0.07 | People (사람이) | 0.09 | 1 | 0.09 |
| 42 | Stated (밝혔다) | 0.10 | Nation (국가) | 0.14 | Tomorrow (내일) | 0.07 | Prosecutor (검찰) | 0.09 | Everyone (모두) | 0.09 |
| 43 | Missile (미사일) | 0.10 | JungEun Kim (김정은) | 0.13 | See (보고) | 0.06 | For real (정말) | 0.09 | North Korea(북한) | 0.09 |
| 44 | Busan (부산) | 0.10 | Minjoodang (민주당) | 0.13 | 3 | 0.06 | JungHee Lee (이정희) | 0.09 | People (사람이) | 0.09 |
| 45 | RT | 0.10 | Jeju (제주) | 0.11 | Like that (그렇게) | 0.06 | Primaries (경선) | 0.08 | Republic of Korea (대한민국) | 0.08 |
| 46 | Yeolin (열린) | 0.10 | First (첫) | 0.11 | Korea (대한) | 0.06 | Matador (비방) | 0.08 | Differnt (다른) | 0.08 |
| 47 | Jeju (제주) | 0.09 | Independent (무소속) | 0.11 | Saenuri (새누리) | 0.06 | Delegate (대표) | 0.08 | North Korea followers (종북) | 0.08 |
| 48 | MyungBak Lee | 0.09 | Minjoo (민주) | 0.11 | How (어떻게) | 0.06 | Together (함께) | 0.08 | What (무슨) | 0.08 |
| 49 | lol (ㅋㅋ) | 0.09 | Center (가운데) | 0.10 | Kr | 0.06 | Very (가장) | 0.08 | Big (큰) | 0.08 |
| 50 | Most (가장) | 0.09 | Running (출마) | 0.10 | The | 0.06 | Same (같은) | 0.08 | Immediately (바로) | 0.08 |

Table A3: 30 most retweeted Twitter accounts by NIS accounts, suspects, regular users, political sample, and opinion leaders. nis subscript indicates that the retweeted account is part of the NIS campaign.

| Rank | NIS | % | NIS suspects | % | Regular users | % | Political sample | % | Opinion leaders | % |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | yark991 | 3.30% | kangjaechon | 0.83% | pyein2 | 0.10% | pyein2 | 1.45% | welovehani | 0.55% |
| 2 | pyein2 | 2.45% | nk_humanrights | 0.66% | oisoo | 0.10% | julka1024 | 0.79% | mettayoon | 0.53% |
| 3 | koreaspiritnana | 2.16% | darmduck | 0.57% | mettayoon | 0.06% | mettayoon | 0.71% | odoomark | 0.41% |
| 4 | kangjaechon | 1.84% | pyein2 | 0.54% | soohjc | 0.06% | machokkk | 0.65% | funronga | 0.38% |
| 5 | nk_humanrights | 1.61% | hosu7722 | 0.51% | julka1024 | 0.05% | supersniper1219 | 0.52% | du0280 | 0.38% |
| 6 | hosu7722 | 1.49% | junghoonyoon | 0.45% | funronga | 0.05% | at_pgh | 0.51% | mindgood | 0.37% |
| 7 | yoonjujin | 1.38% | koreaspiritnana | 0.43% | sbs3927 | 0.05% | sexycat881 | 0.49% | biguse | 0.34% |
| 8 | junghoonyoon | 1.32% | yoonjujin | 0.38% | machokkk | 0.05% | anbjxotn | 0.47% | unheim | 0.31% |
| 9 | freedomnorth | 1.00% | yark991 | 0.23% | unheim | 0.04% | yoonjujin | 0.45% | dogsul | 0.29% |
| 10 | bangmo | 0.95% | bangmo | 0.14% | yark991 | 0.04% | andyou13 | 0.45% | patriamea | 0.25% |
| 11 | nudlenudle$_{nis}$ | 0.86% | hanirm | 0.13% | idol_kadura_bot | 0.04% | funronga | 0.45% | jhohmylaw | 0.25% |
| 12 | darmduck | 0.82% | hotgoodid | 0.12% | supersniper1219 | 0.04% | unheim | 0.44% | oisoo | 0.24% |
| 13 | jogisic$_{nis}$ | 0.74% | hslee3601 | 0.06% | yoonjujin | 0.04% | scartoon1 | 0.42% | madhyuk | 0.23% |
| 14 | shore0987$_{nis}$ | 0.63% | zmfpfm$_{nis}$ | 0.06% | at_pgh | 0.04% | junghoonyoon | 0.42% | nodolbal | 0.23% |
| 15 | hanirm | 0.52% | ourholykorea | 0.05% | du0280 | 0.03% | du0280 | 0.41% | sewoosil | 0.21% |
| 16 | ourholykorea | 0.47% | nudlenudle$_{nis}$ | 0.05% | welovehani | 0.03% | koreaspiritnana | 0.39% | truthtrail | 0.20% |
| 17 | hansuyon19 | 0.44% | humordelivery89$_{nis}$ | 0.04% | sexycat881 | 0.03% | kooceo | 0.37% | viewnnews | 0.20% |
| 18 | wlcodh | 0.42% | oisoo | 0.04% | andyou13 | 0.03% | oisoo | 0.35% | histopian | 0.20% |
| 19 | zmfpfm$_{nis}$ | 0.38% | taesan4$_{nis}$ | 0.04% | hanaag1006 | 0.03% | twitteist | 0.34% | actormoon | 0.19% |
| 20 | sisament | 0.37% | dlarkw$_{nis}$ | 0.04% | koreaspiritnana | 0.03% | patriamea | 0.31% | saveourmbc | 0.19% |
| 21 | yunheesung | 0.35% | shore0987$_{nis}$ | 0.04% | patriamea | 0.03% | welovehani | 0.31% | newstapa | 0.18% |
| 22 | taesan4$_{nis}$ | 0.32% | 1004pansoo | 0.03% | anbjxotn | 0.03% | iamegg3 | 0.29% | pyein2 | 0.18% |
| 23 | seokyoungduk | 0.32% | omajueda$_{nis}$ | 0.03% | thekey16 | 0.03% | neccyber1390 | 0.28% | kyunghyang | 0.18% |
| 24 | kiminhye0 | 0.31% | rhanzhd$_{nis}$ | 0.03% | haeminsunim | 0.03% | jbkim8888 | 0.27% | kyung0 | 0.18% |
| 25 | naya2816 | 0.29% | enhance81$_{nis}$ | 0.03% | junghoonyoon | 0.03% | yark991 | 0.25% | tak0518 | 0.17% |
| 26 | hslee3601 | 0.28% | kimjun283$_{nis}$ | 0.02% | scartoon1 | 0.03% | actormoon | 0.25% | ohmynews_korea | 0.17% |
| 27 | enhance81$_{nis}$ | 0.27% | jogisic$_{nis}$ | 0.02% | artist_kadura_b | 0.03% | darmduck | 0.25% | sungsooh | 0.17% |
| 28 | kangminy99 | 0.27% | tellatoz$_{nis}$ | 0.02% | woqjf012 | 0.03% | maestrok1 | 0.25% | impeter701 | 0.16% |
| 29 | humordelivery89$_{nis}$ | 0.26% | coffe_kim$_{nis}$ | 0.02% | ibgdrgn | 0.03% | wontwowin | 0.25% | baltong3 | 0.16% |
| 30 | kimjungaaa | 0.26% | nobigdeal00$_{nis}$ | 0.02% | neccyber1390 | 0.03% | ruready88 | 0.23% | mediamongu | 0.16% |

# A4 Robustness Checks

This section is devoted to several robustness checks concerning our detection strategies and to better characterize the behavior of NIS accounts.

## A4.1 Varying thresholds and time windows when detecting suspect accounts

In the main text, we flagged accounts as suspicious if they retweet NIS accounts in more than 50% of their total retweets, or if they are on the same co-tweet and co-retweet network component as a known NIS account, where the co-tweets or co-retweets have to occur within one minute. The following figures will show that selecting reasonably smaller or larger thresholds would not change the main results.

Figure A3 illustrates the fraction of co-(re)tweets captured if we choose a different time window than the one-minute interval selected. Specifically, it shows the number of unique pairs of co-(re)tweeting NIS accounts captured as a fraction of all unique pairs of co-(re)tweeting NIS accounts in our dataset.

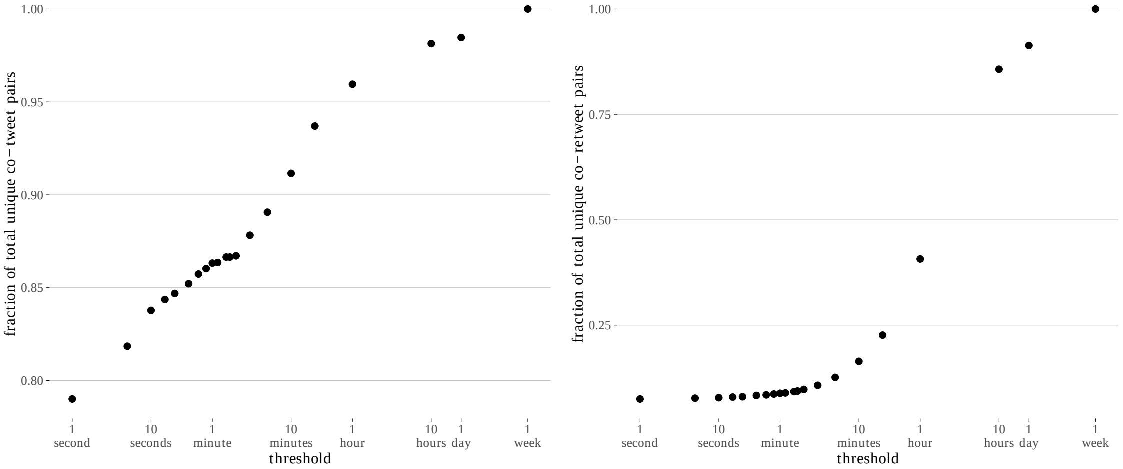

Figure A3: Fraction of unique pairs of NIS account co-tweeting (left panel) and co-retweeting (right panel) within a specific time window.

While the great majority of co-tweets happens within one minute, only 20% of all co-retweets occur in this time window. As argued in the main text, we have good reasons for nevertheless choosing one minute as threshold for co-retweets. First, message coordination should happen within a short time window if the goal is to amplify the specific tweet.

Retweets occurring hours or days later may still increase the impact, but they do not help starting a viral campaign. In addition, many accounts that co-retweet a message may do so simply because they follow the same account and (independently) find the message worth retweeting. We therefore do not consider co-retweets which are days apart as suspicious as co-retweets that occur within the same minute.

Figure A4 shows the number of detected suspect accounts for different thresholds. The number and identity of the suspects remains largely unchanged until the time window approaches 9-10 minutes. At that point, we start including an increasingly large number of regular users – many of which may well have nothing to do with the campaign at all, but simply happen to follow and occasionally retweet the same accounts, or post one common tweet that many other users tweet as well.

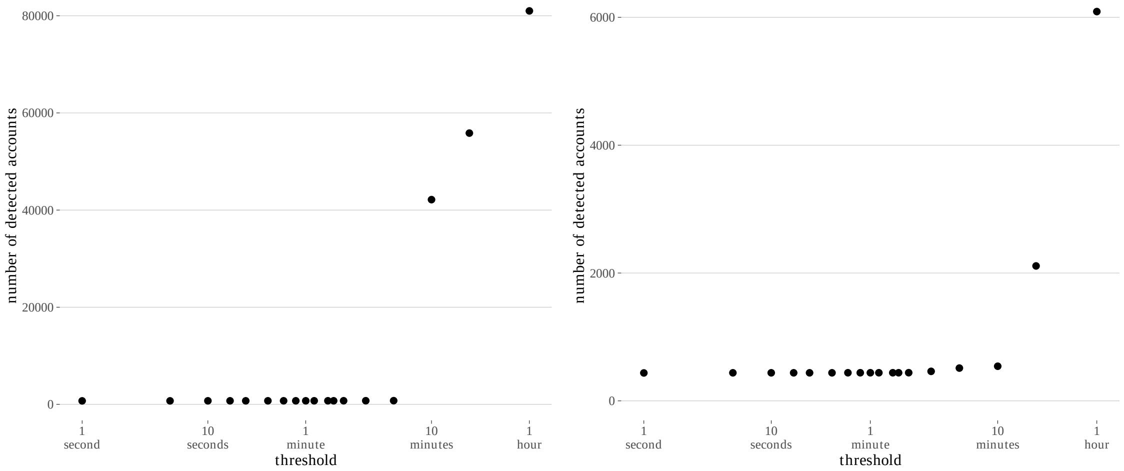

Figure A4: Number of detected suspects for co-tweeting (left panel) and co-retweeting (right panel) within a specific time window.

Turning to the detection via a large proportion of retweeting of known NIS accounts, Figure A5 shows the number of flagged accounts if the threshold of 50% is altered. There are two reasons for choosing this threshold. First, it is a quite natural choice since it implies that an account is suspicious if it retweets NIS accounts more frequently than others. A more practical reason is illustrated in the figure. While the number of suspicious accounts does only increase slightly up to 50%, it starts growing faster afterwards, indicating that the set of suspect accounts then includes too many false positives.

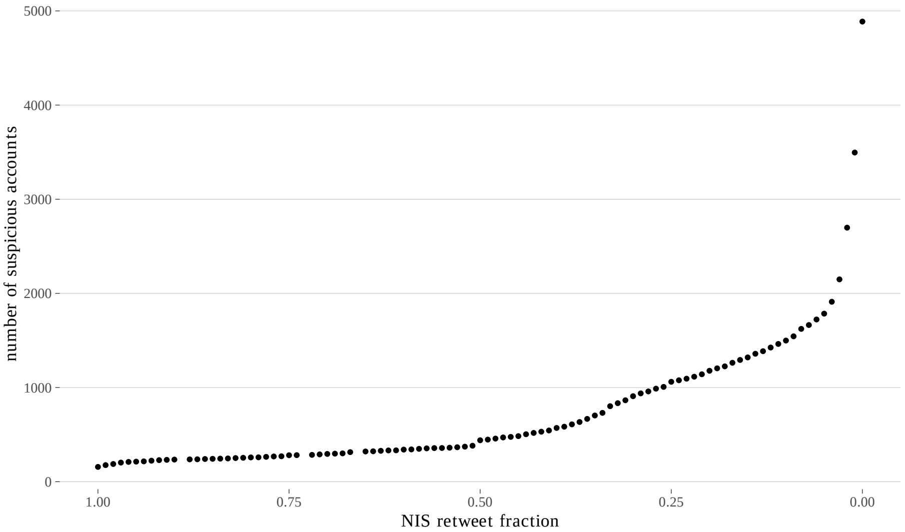

Figure A5: Number of suspicious accounts detected if retweet threshold is altered. Note that the x-axis is reversed.

## A4.2 Detection accuracy

Our outlined detection strategy via the retweet and co-(re)tweet networks does not lend itself for a rigorous accuracy analysis since there are no training steps involved to calibrate parameters. Nevertheless, we employed the following simple test to measure the performance of our detection strategy. We split the set of known NIS accounts into a training set (80%) and a test set (20%). We apply our detection steps with the training set as known accounts with the goal to recover the remaining test accounts. In other words: we examine whether the known NIS accounts in the test set are located in the same connected components as the known NIS accounts in the training set, or if they meet a 50%-threshold of retweeting known NIS accounts. This procedure is repeated 20 times with randomized training and test sets. On average, 85% of the test accounts are detected again. Note though, that the set of known NIS accounts includes many accounts that tweet less than 50 times in the whole period. These accounts are hard to detect with any strategy but also have a negligible impact on the campaign as a whole.

## A4.3 Network statistics

The following tables show basic statistic for the NIS networks examined in the main text and compares them to those of the equivalent networks constructed among a random set of users of equal size. The values for the random sample rows are averaged over 5,000 draws, one time for the complete dataset and once for the dataset restricted to political keywords. The political keywords include the names of the candidates and key political actors, event-related keywords, as well as other general political keywords defined by Song, Kim, and Jeong (2014).

Table A4: Statistics for retweet networks

| Type | unique pairs | density |
|---|---:|---:|
| random samples (all) | 14 | 0.0001 |
| random samples (pol. tweets) | 23 | 0.002 |
| NIS | 12412 | 0.0187 |

Table A5: Statistics for co-tweet networks

| Type | unique pairs | density |
|---|---:|---:|
| random samples (all) | <1 | <0.0001 |
| random samples (pol. tweets) | 2.4 | 0.0005 |
| NIS | 1093 | 0.121 |

Table A6: Statistics for co-retweet networks

| Type | unique pairs | density |
|---|---:|---:|
| random samples (all) | <1 | <0.0001 |
| random samples (pol. tweets) | 2.6 | 0.0008 |
| NIS | 4654 | 0.023 |

Irrespective of whether we look at the retweet, cotweet or co-retweet network, the NIS accounts exhibit a much denser coordination pattern than the random users.

## A4.4 Impact analysis

Figure A6 reproduces our findings on the impact of NIS accounts in terms of retweets received, using only tweets with political keywords. Even the political tweets of NIS accounts do rarely get retweeted and make up a negligible fraction of the retweets in our dataset.

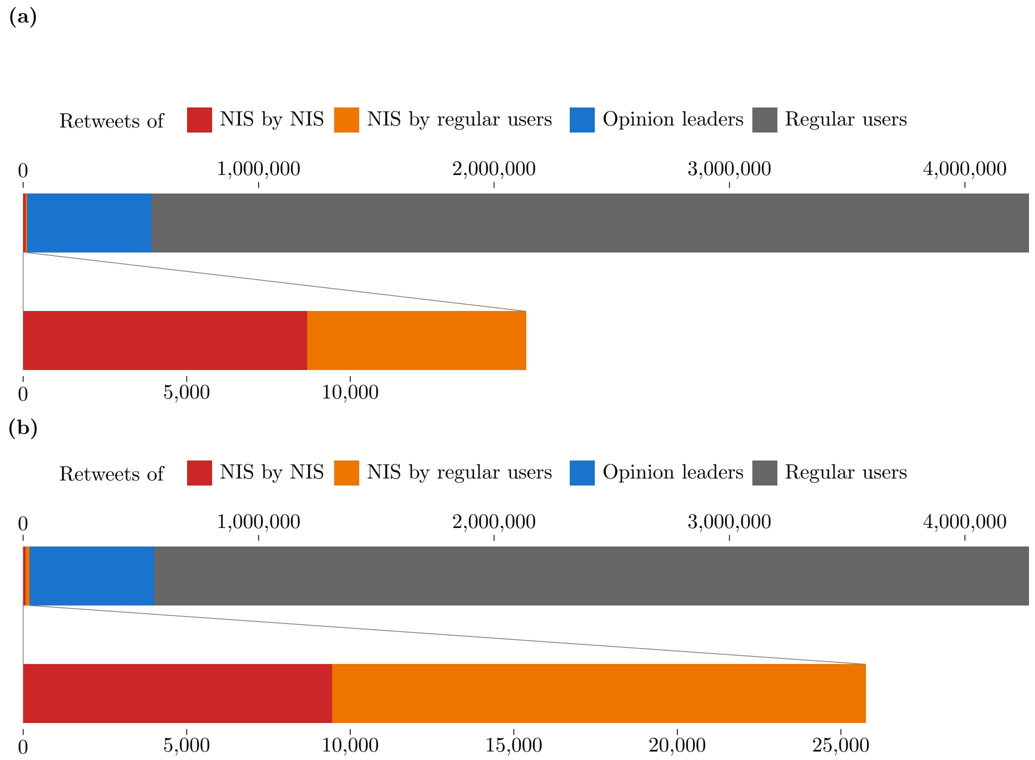

Figure A6: Received political retweets by different groups of users. (a) only includes known NIS accounts and (b) includes suspect users in the group of NIS accounts.

We also examine the impact of NIS accounts by examining whether the NIS was successful in boosting the impact of like-minded accounts by retweeting them: do regular users follow the lead and retweet those accounts as well? If this is the case, we should expect two patterns: we should see an increase of retweets of like-minded accounts in our sample of random and political tweets, in particular after the NIS campaign ramps up on September 1, and possibly a drop after the campaign got revealed. We may also observe that the retweet patterns of other Twitter users are similar to that of the NIS accounts. To test this idea, we compare in Figure A7 how often these like-minded accounts were retweeted by NIS accounts, NIS suspects, and opinion leaders, and how often the retweets of the like-minded accounts appeared in a random and a political sample.

Figure A7 does not show any signs of a retweet propagation by NIS. Although a fair amount of retweets of those like-minded accounts were found in other samples, the patterns of retweeting are more similar to each other than to the known NIS accounts. We also see no noticeable lasting change in the number of retweets of these like-minded accounts in the random and political samples after the the sudden shutdown of the campaign. If the

NIS was indeed successfully boosting these accounts, then the disappearance of such support should presumably reduce the number of retweets these supported accounts receive from other users. The only drop in retweets received is – unsurprisingly – among the NIS and NIS suspects themselves, which, having been shut down, cannot retweet anymore. There is a steady increase of retweets received over time, but the most noteable increase occurs just before September 1, and there is a general increase of tweets over time in our dataset anyway.

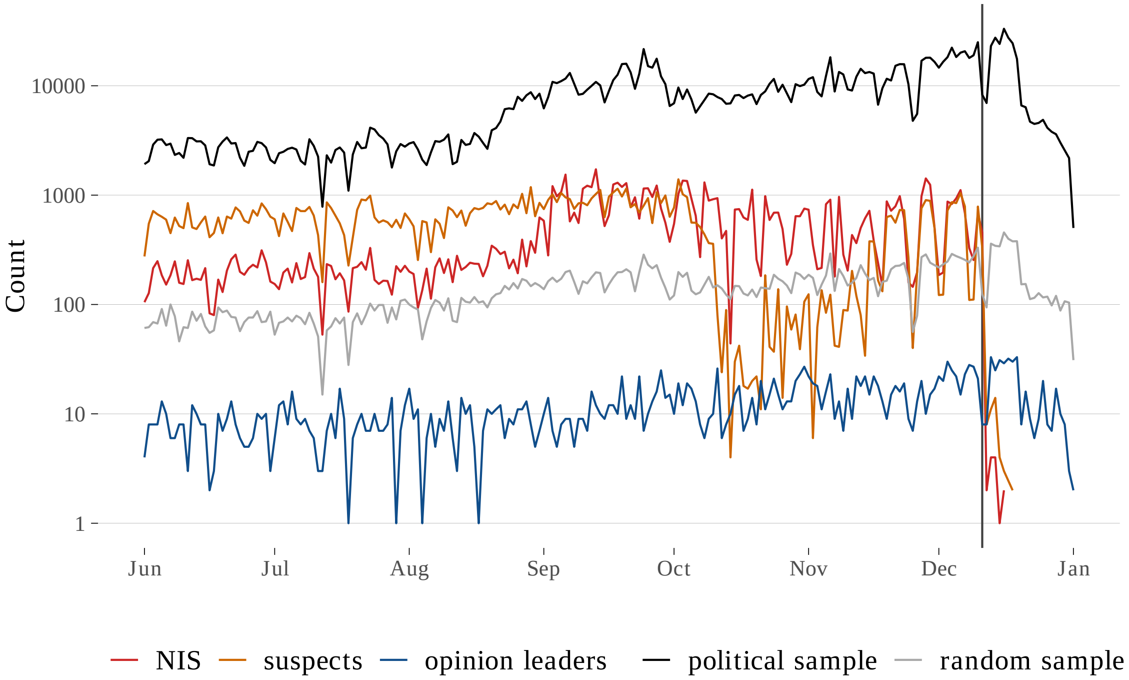

Figure A7: Retweet counts of the Twitter accounts retweeted by NIS. The counts are log-transformed values.

# A5 Dataset statistics

Table A7: Statistics of the dataset

| Type | tweets | retweets | retweets/tweets |
|---|---:|---:|---:|
| All | 86,980,130 | 15,745,090 | 18% |
| NIS | 194,190 | 93,519 | 48% |
| political | 5,840,159 | 4,272,576 | 73% |
| NIS political | 77,798 | 43,601 | 56% |

# A6 Comparison with other detection methods

We are not aware of existing methods that allow for the detection of disinformation. There is, however, a growing literature on bot detection in computer science. Yet, it should be noted that there are considerable conceptual differences between (the detection of) bots and astroturfing accounts. There are plenty of bots that are not part of a secret influence campaign, but simply aggregate news on a specific topic, or inform followers of the current weather, for instance. Many social science studies purporting to study and identify bots are thus in fact interested only in a subset of bots: those involved in astroturfing campaigns. On the other hand, we also know that most astroturfing campaigns do not just employ automated accounts, but instead use a mix of automation and posting by humans – sometimes in the same account (Grimme, Assenmacher, & Adam, 2018). Researchers using bot detection algorithms to study astroturfing therefore succumb to a conceptual mismatch between what they want to study and what they actually measure.

Our relational detection method is based on metrics derived from group-based behavior of accounts, irrespective of whether the behavior is automated or not, and therefore is designed to measure the actual concept of interest, astroturfing campaigns. Still, we will compare the two approaches here – also to illustrate the problems with using bot detection to identify astroturfing campaigns.

The Botometer (formerly BotOrNot), created by a team of academics from the Indiana University, is most frequently used by social scientists studying Twitter.[1] The algorithm uses a plethora of account-level and context information (such as device used, or geolocation) to assign each account examined a probability that it is automated. Since we lack many of the variables needed and cannot get this data for the already deleted NIS accounts, we cannot compare our detection method to the Botometer.

[1] https://botometer.iuni.iu.edu

We can, however, evaluate our findings against a commonly used method for detecting highly automated accounts. Howard and Kollanyi (2016) have argued in a series of publications that accounts posting more than 50 tweets per day on average should be considered highly automated, i.e., social bots. In a 10% sample of all tweets, this converts to a threshold of 5 tweets per day. Of the 702 NIS accounts in our dataset, only 94 would thus be flagged as highly automated. A manual inspection reveals that this roughly captures the group of NIS accounts spreading newspaper headlines only, but misses the other accounts – despite the fact that they also rely on automation to post retweets. The approach fares better with the suspect accounts, where 511 of 834 meet the threshold. However, it also identifies 14,522 additional accounts as highly automated. These may well be social bots, but they are unlikely to be part of the NIS campaign. We take this as evidence that our relational approach is preferable to methods focusing on automation by individual accounts in isolation.

## A7 Complete co-(re)tweet networks

Figures A8 and A9 show the complete co-(re)tweet networks of the dataset with a one minute threshold.

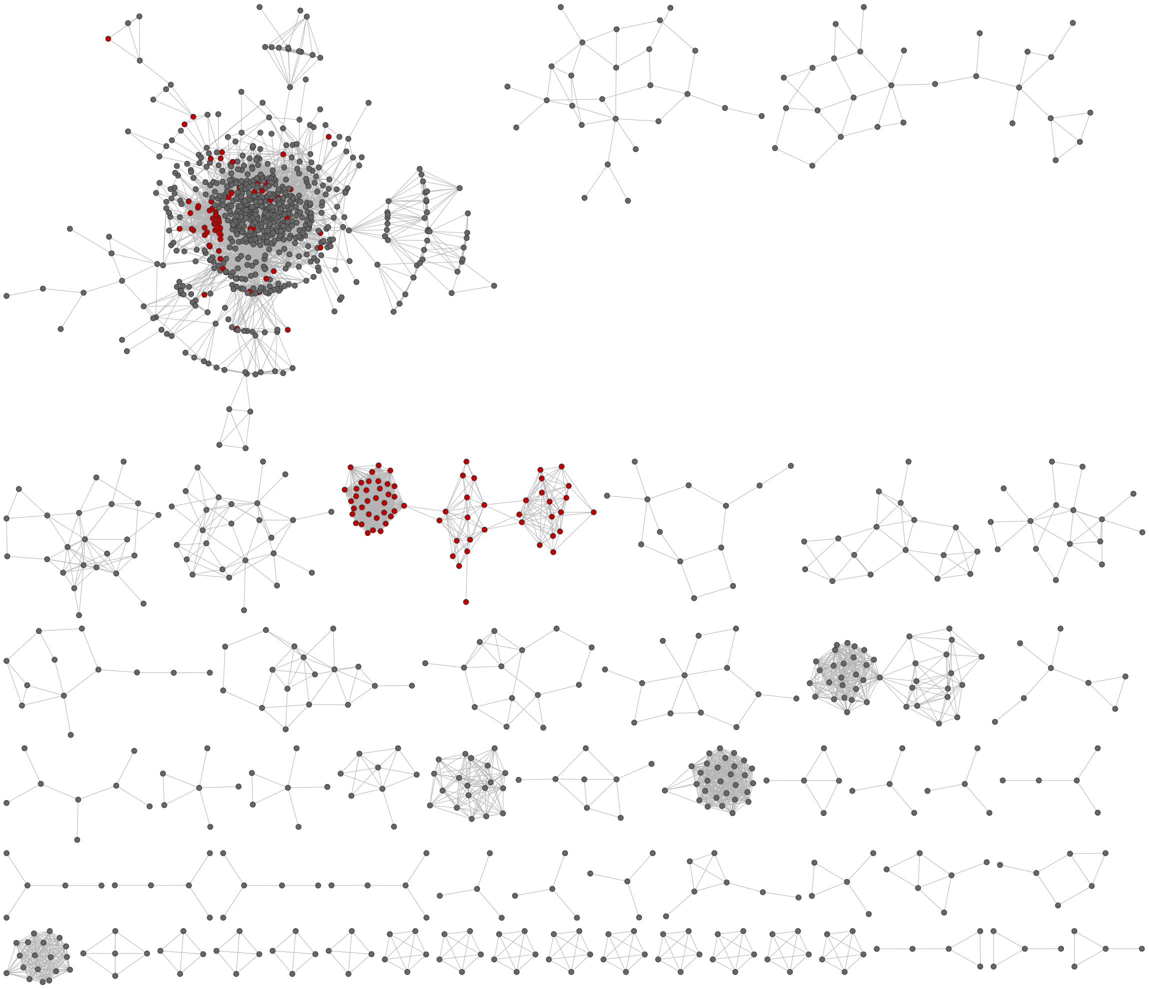

Figure A8: Complete co-tweet network with one minute threshold. Only components with more than five nodes are shown to reduce clutter. Known NIS accounts are shown in red.

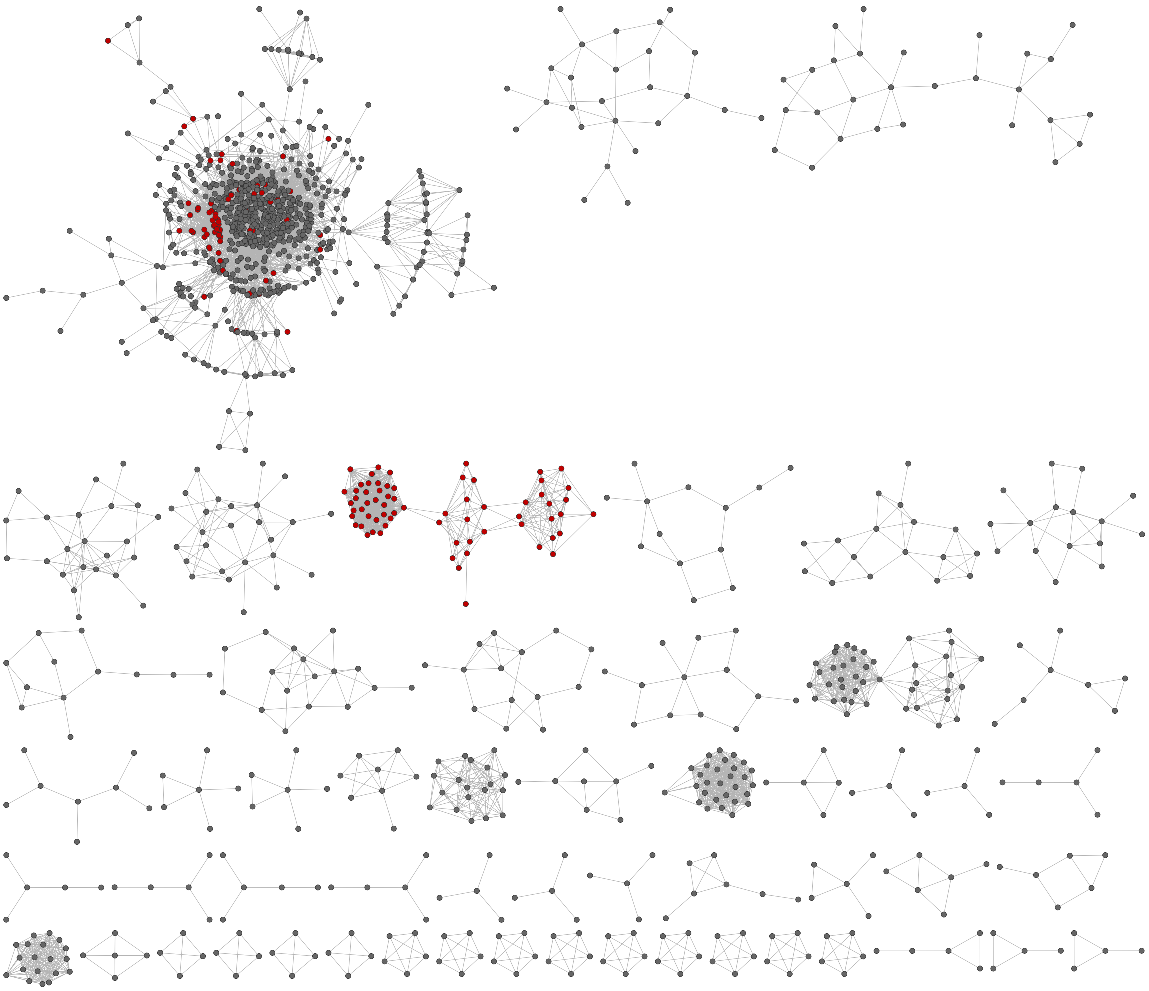

Figure A9: Complete co-retweet network with one minute threshold. Only components with more than five nodes are shown to reduce clutter. Known NIS accounts are shown in red.

# References

Grimme, C., Assenmacher, D., & Adam, L. (2018). Changing perspectives: Is it sufficient to detect social bots? In *International Conference on Social Computing and Social Media* (pp. 445–461).

Hartigan, J. A. (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association*, *67*(337), 123–129. Retrieved 22.12.2018, from http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1972.10481214

Howard, P. N., & Kollanyi, B. (2016). *Bots, #Strongerin, and #Brexit: Computational propaganda during the UK-EU Referendum.* Retrieved from http://politicalbots.org/wp-content/uploads/2016/06/COMPROP-2016-1.pdf

Keller, F., Schoch, D., Stier, S., & Yang, J. (2017). How to manipulate social media: Analyzing political astroturfing using ground truth data from South Korea. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media* (pp. 564–567). Menlo Park, CA: The AAAI Press.

Song, M., Kim, M. C., & Jeong, Y. K. (2014). Analyzing the political landscape of 2012 Korean presidential election in Twitter. *IEEE Intelligent Systems*, *29*(2), 18–26.