Demystifying Social Bots - On the Intelligence of Automated Social Media Actors
Dennis Assenmacher, Lena Clever, Lena Frischlich, Thorsten Quandt, Heike Trautmann and Christian Grimme

Supplemental material

List of Figures

S1 Distribution of open software repositories matching the respective search term over all social-media platforms investigated in this work. ................................................................. 2
S2 Geographical origin of Telegram repositories determined by evaluating the geographical information provided by the considered software collaboration platforms. ........................................... 3
S3 Geographical origin of Skype repositories determined by evaluating the geographical information provided by the considered software collaboration platforms. ........................................... 3
S4 Geographical origin of Facebook repositories determined by evaluating the geographical information provided by the considered software collaboration platforms. ........................................... 3
S5 Geographical origin of vKontakte repositories determined by evaluating the geographical information provided by the considered software collaboration platforms. ........................................... 3
S6 Geographical origin of Reddit repositories determined by evaluating the geographical information provided by the considered software collaboration platforms. ........................................... 4
S7 Geographical origin of Instagram repositories determined by evaluating the geographical information provided by the considered software collaboration platforms. ........................................... 4
S8 Geographical origin of WhatsApp repositories determined by evaluating the geographical information provided by the considered software collaboration platforms. ........................................... 4
S9 Geographical origin of Pinterest repositories determined by evaluating the geographical information provided by the considered software collaboration platforms. ........................................... 4
S10 Geographical origin of Snapchat repositories determined by evaluating the geographical information provided by the considered software collaboration platforms. ........................................... 5
S11 Data acquisition process and architecture with individual crawling instances (1), transformation of heterogeneous data into an intermediate schema (2), central persistence within an elasticsearch database (3) and individual on-demand analysis with external tools (4). ................................................................. 7
S12 Parameter optimization for LDA To find the optimal number of topics, we calculated different performance metrics, proposed in different works (Röder, Both & Hinneburg, 2015; Mimno, Wallach, Talley, Leenders & McCallum, 2011; Cao, Xia, Li, Zhang & Tang, 2009). While the two metrics on the top are minimised, both coherence measures are maximised. Additionally we strive for a solution that can be interpreted by a human. Therefore, we fix a total of 20 topics (dotted vertical line) as a suitable trade-off setting. ................................................................. 8

List of Tables

S1 Examined markets for social bot-related commodities in the clear and darknet. The analyzed selection of clearnet markets is sampled from the total identified number of $N = 97$ markets. The darknet markets represent the full list of analyzed markets. ................................................................. 2
S2 API support for social platforms, sorted by number of repositories. Platforms with no API support provide no official programming interface. Still, access by remote control of a web browser is possible but tedious. Limited API access restricts important features of the social media platform like publishing content or interacting with other accounts. API access allows access to most or all features of the platform but may be subject to rate limits. A BotAPI explicitly allows automation of accounts but simultaneously enables the platform to explicitly label them as bot accounts. We observe a positive rank correlation between the number of repositories found for a specific social platform and the corresponding level of API support ($\rho = 0.75$). ................................................................. 6
S3 LDA results: topic word distribution and representative descriptions of topics 0-9. For each topic, we selected a suitable representative from the set of five descriptions with the highest assignment probability. ................................................................. 6
S4 LDA results: topic word distribution and representative descriptions of topics 10-19. For each topic, we selected a suitable representative from the set of five descriptions with the highest assignment probability. ................................................................. 9
Figure S1. Distribution of open software repositories matching the respective search term over all social-media platforms investigated in this work.

Table S1. Examined markets for social bot-related commodities in the clear and darknet. The analyzed selection of clearnet markets is sampled from the total identified number of $N = 97$ markets. The darknet markets represent the full list of analyzed markets.

| Clearnetmarkets                      | Darknetmarkets                        |
|--------------------------------------|---------------------------------------|
| Altenen                              | 0day*                                 |
| BAE Applied Intelligence Systems     | Berlusconi Market*                    |
| Binary Revolutions Forum             | Cerberus                              |
| Black Hat Russia                     | Charlieuk*                            |
| Black Web Forum                      | Dream Market (PGP)*                   |
| Breach Forums                        | Exodu$                                |
| Cardmafia                            | FB Hack Tool                          |
| Cryptohackers                        | Fight Club*                           |
| Dark Pid                             | Free Hacking Tools                    |
| Devil Group                          | Gammagoblin*                          |
| Followlike                           | Hack Canada*                          |
| Free-Hack                            | Hackerplace                           |
| Hack a day                           | jRAT                                  |
| Hack Forums                          | Mr. Robot Shop*                       |
| Hacker’s List                        | Ostrich Hackers Hunters               |
| Hackerthreads                        | Pushing Taboo*                        |
| Hackervoice                          | Quality King*                         |
| Hitb                                 | Raniom                                |
| Mahwr                                | Rent-A-Hacker                         |
| Nulled (Forum)                       | RsClub Market                         |
| Nulled (Marketplace)                 | Stoned100                             |
| Offensive Community                  | Sourcery*                             |
| Quora                                | The French Connection*                |
| Safe Sky Hacks                       | The Hack Liar                         |
| Smart Hackerz                        | Tochka                                |
| Social Engineering                   | Torum                                 |
| Team Corrupt                         | ToYouTeam*                            |
| Toolbase                             | Trade Route                           |
| Topsocialbot                         | Valhalla                              |
| Vigilante Tech                       | Wall Street Market                    |

*Market did not contain any relevant item.

To compensate for the disproportional large share of Clearnet markets, we examined a random 30 in detail.
Figure S2. Geographical origin of Telegram repositories determined by evaluating the geographical information provided by the considered software collaboration platforms.

Figure S3. Geographical origin of Skype repositories determined by evaluating the geographical information provided by the considered software collaboration platforms.

Figure S4. Geographical origin of Facebook repositories determined by evaluating the geographical information provided by the considered software collaboration platforms.
Figure S5. Geographical origin of vKontakte repositories determined by evaluating the geographical information provided by the considered software collaboration platforms.

Figure S6. Geographical origin of Reddit repositories determined by evaluating the geographical information provided by the considered software collaboration platforms.

Figure S7. Geographical origin of Instagram repositories determined by evaluating the geographical information provided by the considered software collaboration platforms.
Figure S8. Geographical origin of WhatsApp repositories determined by evaluating the geographical information provided by the considered software collaboration platforms.

Figure S9. Geographical origin of Pinterest repositories determined by evaluating the geographical information provided by the considered software collaboration platforms.

Figure S10. Geographical origin of Snapchat repositories determined by evaluating the geographical information provided by the considered software collaboration platforms.
Table S2. API support for social platforms, sorted by number of repositories. Platforms with no API support provide no official programming interface. Still, access by remote control of a web browser is possible but tedious. Limited API access restricts important features of the social media platform like publishing content or interacting with other accounts. API access allows access to most or all features of the platform but may be subject to rate limits. A BotAPI explicitly allows automation of accounts but simultaneously enables the platform to explicitly label them as bot accounts. We observe a positive rank correlation between the number of repositories found for a specific social platform and the corresponding level of API support ($\rho = 0.75$).

| Social Platform | no API | limited API | API | BotAPI |
|-----------------|--------|-------------|-----|--------|
| Telegram        | ×      |             |     |        |
| Twitter         | ×      |             |     |        |
| Facebook        | ×      |             |     |        |
| Reddit          | ×      |             |     |        |
| Skype           | ×      |             |     |        |
| Instagram       | ×      |             |     |        |
| Youtube         | ×      |             |     |        |
| Whatsapp        | ×      |             |     | ×      |
| Linkedin        |        | ×           |     |        |
| Tumblr          | ×      |             |     |        |
| vKontakte       | ×      |             |     |        |
| Snapchat        | ×      |             |     |        |
| Pinterest       | ×      |             |     |        |

Table S3. LDA results: topic word distribution and representative descriptions of topics 0-9. For each topic, we selected a suitable representative from the set of five descriptions with the highest assignment probability.
Figure S11. Data acquisition process and architecture with individual crawling instances (1), transformation of heterogeneous data into an intermediate schema (2), central persistence within an elasticsearch database (3) and individual on-demand analysis with external tools (4).
Figure S12. Parameter optimization for LDA To find the optimal number of topics, we calculated different performance metrics, proposed in different works (Röder, Both & Hinneburg, 2015; Mimno, Wallach, Talley, Leenders & McCallum, 2011; Cao, Xia, Li, Zhang & Tang, 2009). While the two metrics on the top are minimised, both coherence measures are maximised. Additionally we strive for a solution that can be interpreted by a human. Therefore, we fix a total of 20 topics (dotted vertical line) as a suitable trade-off setting.
|   | representative repository description                                                                 | word distribution |
|---|----------------------------------------------------------------------------------------------------------------|-------------------|
| 10| Pinterest API to interact with Pinterest. This bot need use your authenticate information like email/password. |                   |
| 11| A simple example of how one could possibly make a Twitter bot that mirror’s somebody else’s speech. Since twitter bots are against Twitter’s TOS, I would never actually make one and you shouldn’t either. |                   |
| 12| A terminal bot app that allows the user to use the various features of the Twitter API.                      |                   |
| 13| This is the Alpha Team FB Messenger Bot service for handling Enque Appointment Requests on Facebook messenger. |                   |
| 14| A Telegram bot to add ”Liking” functionality similar to GroupMe for group chats.                            |                   |
| 15| Telegram bot using the Python API that gets films rating from IMDb and metacritic.                           |                   |
| 16| A Twitter bot which replies with Air Quality Index of a city on tweeting #airqualityin followed by name of the city |                   |
| 17| Telegram bot for telegram, that notifying you about upcoming events, corrections imminent and searching through users |                   |
| 18| This is a twitter bot that uses a list of nouns and verbs and tweets out every few hours in the form of ”Verb the Noun”. |                   |
| 19| A Markov chain bot that pulls from Trump’s Twitter as well as other sources.                                  |                   |

Table S4. LDA results: topic word distribution and representative descriptions of topics 10-19. For each topic, we selected a suitable representative from the set of five descriptions with the highest assignment probability.
References
Cao, J., Xia, T., Li, J., Zhang, Y. & Tang, S. (2009). A density-based method for adaptive lda model selection. *Neurocomputing, 72*(7-9), 1775–1781.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M. & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 262–272). Association for Computational Linguistics.

Röder, M., Both, A. & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399–408). ACM.