Identifying Artificial Actors in E-Dating: A Probabilistic Segmentation Based on Interactional Pattern Analysis

Andreas Schmitz, Olga Yanenko, and Marcel Hebing

Abstract We propose different behaviour and interaction related indicators of artificial actors (bots) and show how they can be separated from natural users in a virtual dating market. A finite mixture classification model is applied on the different behavioural and interactional information to classify users into bot vs. non-bot-categories. Finally the validity of the classification model and the impact of bots on sociodemographic distributions and scientific analysis is discussed.

1 Introduction

Social networking services have extensively proliferated in the last years and thereby, they increasingly enable users to relocate various practices into the Internet. For social sciences this development is a possibility to acquire a new kind of data on social processes by directly observing agency, interactions and social networks. One example is the new approach of studying issues of mate choice processes constituted by recording and analysing interactions within online mate markets (Schmitz et al. 2009). Since those behavioural records are not originally arranged for scientific intentions, there are particular methodological aspects to be taken into account before analysing the data against a theoretical background. A considerable problem

A. Schmitz,
Chair of Sociology I, University of Bamberg, Wilhelmsplatz 3, 96045 Bamberg, Germany
e-mail: andreas.schmitz@uni-bamberg.de

O. Yanenko, (✉)
Chair for Computing in the Cultural Sciences, University of Bamberg, Feldkirchenstrasse 21, 96045 Bamberg, Germany
e-mail: olga.yanenko@uni-bamberg.de

M. Hebing
Socio-Economic Panel Study (SOEP), DIW Berlin, Mohrenstrasse 58, 10117 Berlin, Germany
e-mail: mhebing@diw.de

W. Gaul et al. (eds.), *Challenges at the Interface of Data Analysis, Computer Science, and Optimization*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-24466-7_33,

with analysing and interpreting this data arises with the presence of artificial actors or so-called *bots*. Bots are automated third-party programs trying to make users engage into contact and eventually into an over-priced and useless external product. Beyond a diminished benefit of the user and reputational losses of the particular dating-provider, data quality becomes an important issue from a scientific point of view. In this paper we develop behaviour and interaction related indicators of bot-presence and separate artificial users from humans by applying a finite mixture model on this informational vectors. The classification results suggest an amount of approximately 3.72% artificial users. Contrasting our latent class prediction with a manual screening for bot-presence yields a high validity. We demonstrate that those entities generate a disproportionately high amount of first contacts (30.5%) and significantly distort socio-demographic distributions, such as gender and age. We conclude with an outlook and a discussion about practical implications.

2 Bots in the Social Web

The term bot is the abbreviation for robot and implies different automated programs that have the ability to act autonomously to some extent (Gianvecchio et al. 2008). There is a vast variety of bots which were mostly implemented for performing tasks in different areas of the World Wide Web. The field of bot activities varies from collecting information about websites for search engine ranking generation to text editing as accomplished by Wikipedia[1] bots for instance (Fink and Liboschik 2010). These bots are created to support their operators in performing boring or recurring jobs. But there are also numerous bots, which were deployed for financial purposes. Most of these bots have a further-reaching negative impact for Internet users they are interacting with. The aims of such bots vary from copyright fraud (Poggi et al. 2007) or identity theft to gross marketing tricks. The most common operating mode for bots is to establish a direct contact with other people, usually by writing spam emails. These spam mails differ from conventional advertisement since they are formulated as personal messages and are therefore often not recognised as malicious. Human spammers exist as well, but since the manual distribution of emails is very time-consuming, it is a common practice to automate this process by implementing a bot. The progression of today's spam filters forces bot developers to strike new paths. To equip bots with some kind of personality is therefore a major strategy in creating new malicious programs. For this reason, bots are opening up new areas of social acting on the web by creating their own accounts on different Web 2.0 sites, such as online-dating platforms, and masquerade as human actors. In the social web context the spam bot problem is also referred to as Spam 2.0 (Hayati et al. 2010). Usually a Spam 2.0 bot contacts a person only once with a nice sounding text at some point telling the receiver to follow

[1] http://en.wikipedia.org/

a link to some website, to call some costly phone number (Dvorak et al. 2003) or just to change the communication channel to private email in order to collect email adresses for subsequent spamming. Particularly on dating platforms people are in search of (communication) partners and therefore prone to contacts with fake profiles.

Since bots are a real threat for other users of chat rooms and social network communities, there has been a big interest in identifying and banning them. Because of the wide range of bot types and usage scenarios many different approaches were introduced in the past. Spam bots are usually recognised by means of identifying suspect text patterns or sender addresses since no other information is available. Behavioural bots in contrast can be determined by considering the social structure of their actions. The major fact for differentiating between humans and bots is that human behaviour is not easily computable while bots always show some recurring behavioural patterns.

2.1 Bot Detection

Since spam bots are the main problem in the area of social web communities, content analysis is used for their identification. Therefore lists with keywords that are known to be used by bots are compared with messages. This method is not very reliable as bots regularly update their vocabulary depending on the blacklists. Another point is that bots often add random characters to the keywords that look like typing errors and on the one hand let them seem more human and on the other hand avoid matching a word from the blacklist (Gianvecchio et al. 2008). In some cases there is no possibility to analyse the text contents, for example due to privacy reasons.

The detection of bots by analysing their behavioural patterns is a common practice for identifying bots in online games like World of Warcraft[2] (Chen et al. 2006, 2009). Similar techniques can be used in social network scenarios as well. Gianvecchio et al. (2008) proposed a method for detecting bots in Internet chat rooms. They collected interactional data on the Yahoo! chat system[3] which then was manually labelled as human, bot or ambiguous. Additionally, two indicators, namely inter-message delay and message size, were defined to help identifying the bots. The classification was made by a machine learning approach, where half of the available data was used for training the classifiers and the other half for testing them. However, bot detection methods introduced in the past rely either on pure content analysis or behavioural patterns. Thus, we argue that interactional patterns should be taken into account as well and conjecture that they provide better evidence as they include further information.

[2] http://eu.battle.net/wow/

[3] http://messenger.yahoo.com/

2.2 Bot-Induced Problems with Data Quality and Theory-Testing

But what kind of problems can bots cause in our field of application? From a commercial perspective the presence of artificial actors represents a risk of lowering the quality of the product and thereby the customer retention. But if there is a difference between natural and artificial actors some analytical problems arise as well. First, the total amount of the sample might be overrated to the extent of the presence of bots. Second, the distribution of relevant characteristics or marker parameters as for example sex (female), age (young), mating preferences (short term), and outer appearance (very attractive) might be biased. This can be expected as programmers rely on their stereotypes when coding a bot with an expected maximum of success probability. The same holds true for manual spammers. Third, the behavioural and interactional patterns, being the core advantage of process-generated data (Schmitz et al. 2009), might be biased. As for example young attractive women are usually selective with regard to their contact and answer behaviour, whereas artificial actors with similar profile information do probably not reveal this behaviour, as it would be irrational to be confined to only a few male victims of fraud. This problem of misrepresentative interactional patterns becomes worse if we take into account that a bot usually contacts a lot of users, leading to an aggregate pattern where usually choosy persons show up as extreme outgoing ones. Thus, results of previous research on e-dating using web-generated process data have to be interpreted with caution.

3 Method

3.1 Sample

The raw data is provided as anonymised dump files from an operative SQL database of the cooperative maintaining company. It consists of real time click streams, which first have to be edited and then converted to flat file formats readable for common statistical packages. This process-generated interaction data (who sends whom a message and at what time) is integrated with profile information (what do users tell other users about themselves on their profile pages) via a unique user-id. The sample used in this paper consists of 32,365 active users who generated 683,312 messages from January 1st 2009 to April 14th 2010, forming 362,067 dyads.

3.2 Behavioural and Interactional Indicators

Based on the information of the sender's and his contact's user-ids in combination with the time stamp of the message, we identify interactional dyads. In a next

step, we deduce information about their behavioural and interactional patterns. We combine this information with indicators about the content of the messages to identify messages containing email or web addresses and to count the number of characters in the messages. This information is aggregated on the level of the individual user. We define two types of indicators: behaviour and interaction related indicators. Behaviour related indicators measure how users send messages. We expect artificial actors to differ from humans in some key properties:

- *First Contact Sending Rate (log)* [FCSR]: Bots need to produce a huge amount of messages in order to enhance the intended reaction rate. As the total number of first contacts depends on the duration of membership on the platform, we calculate a logarithmised time-dependent rate.
- *High-Speed Sending-Ratio* [HSSR]: Bots are able to send messages in a faster sequence than humans are. We calculate a ratio, comparing the amount of messages sent in less then 20 seconds divided by the total amount of messages sent.
- *Message Length Standard Deviation (log)* [MLSD]: Most bots send standardised messages. Even if they are able to exchange the name of their contact, messages should not differ in length. We use the standard deviation of characters in the messages of one user as an indicator for the variation of message content.
- *Inspection vs. Contacted-Ratio (log)* [IVCR]: There is no reason for bots not to contact a user after visiting his profile. As it is not necessary to visit a user's profile before contacting him, artificial actors should initiate more contacts then visiting profiles.
- *Web-* and *Email-Address-Ratio* [WAR/EAR]: Bots are programmed to initiate a contact outside the platform, so they send further contact information. The challenge is to identify this contact information as the programmers try to codify them to avoid spam filter. For example, you will find (at) instead of an @-sign. This indicator is defined as the amount of messages containing such expressions, divided by the total amount of messages.

In addition, we define interaction related indicators, which take into account that the human reaction towards artificial actors differs from the reaction towards human actors and thereby reveals a lot about the character of the contacting user.

- *Mean Conversation Length (log)* [MCL]: If two users hold up a conversation for a longer time, it is an indication for reciprocal human interaction. We calculate the mean of the number of interactions in the conversations initiated by the users.
- *Response-Received-Ratio* [RR]: In contrast, bots will less often receive answers on their messages as humans identify bots in most cases by common sense. This indicator measures the ratio of contacts initiated by one user receiving an answer in relation to the total number of initiated contacts by the same user.
- *Blocked User* [BU]: Users sometimes complain about artificial and commercial actors, which eventually leads to a block of the particular profile.

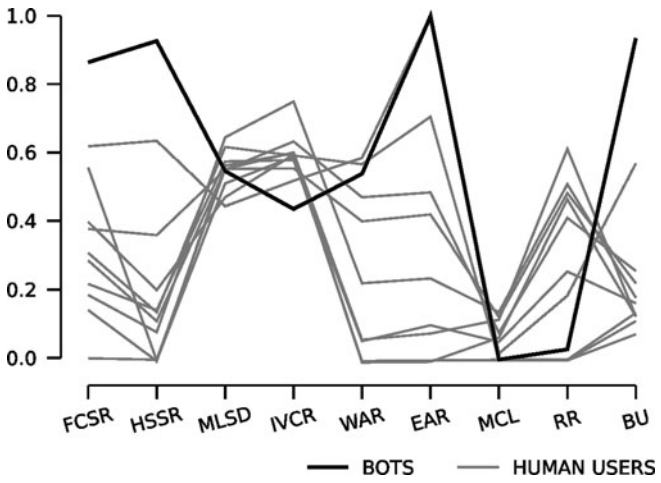

Fig. 1 Profile-Plot of 13 Latent Classes (Class 9 highlighted)

3.3 A Finite Mixture Analysis of Bot Indicators

Usually different information on behaviour is used univariately in bot-detection approaches. In order to distinguish between natural and artificial actors, we apply a finite mixture model with mixed indicators (Vermunt and Magidson 2002) on the described behavioural and interactional indicators. Since the finite mixture model is independent from the scaling (e.g. categorical, ordinal, continuous, Poisson, etc.) of the variables used it is an appropriate choice for categorising online-daters based on the indicators described in the previous section. Thus, the mixed indicators model is a general latent variable approach that specifies a model with manifest variables of different scaling and C categorical latent variables. Those C unobserved categorical variables can be identified by maximum likelihood estimation under the assumption of multivariate normal distributions given the latent profiles. The joint density of the observed indicators, $f(y)$, is a function of the prior probability of belonging to latent class π_k with class-specific mean vectors and covariance matrices $f_k(y_{ij}|\theta_{jk})$, given the model parameters θ.

$$f(y_i|\theta) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} f_k(y_{ij}|\theta_{jk})$$

The observed indicators in our model, the different information derived from the click stream data, are simultaneously modelled by this technique. The models are estimated with the statistic software Latent Gold 4.5.

According to BIC and CAIC-information criteria the optimal solution yields 13 latent classes. Figure 1 shows the profile plot of the optimal solution. Consider class 9 which covers 3.72% of all users in the sample. This class strongly differs from the

other classes by means of its high First Contact Sending Rate and its considerable High-Speed Sending-Ratio. Furthermore, the Inspection vs. Contacted-Ratio shows that actors belonging to this class are less interested in the profile information of contacted users. In addition, this class has a very high amount of messages containing email addresses and hyperlinks. The average length of interaction with other users is very short and they can be characterised by a low Response-Received-Ratio. Finally, this class consists of users who are with a very high probability (.94) blocked by the provider. Thus, class 9 differs clearly from classes 1–8 by means of their behavioural and interactional patterns. But class 10, containing 3.22% of all actors, shows a distinctive pattern as well. Users in this class reveal the second highest First Contact Sending Rate and the second fastest High-Speed Sending-Ratio. They have the lowest standard deviation of their text messages and thereby use the same text several times in their communication. According to the second lowest Inspection vs. Contacted-Ratio, they contact most of the profiles they visited but not all of them, pointing out to some degree of selectiveness. Similar to class 9, the probability of sending an email or web address is very high in class 10. On average their interaction courses are very short and their Response-Received-Ratio is adverse. Finally, users of class 10 have the second highest probability of being blocked by the provider (.57). We interpret actors belonging to class 10 as spammer, people trying to scam users in respect to financial purposes. Class 10 shows similar behavioural patterns as bots do, though due to their manual restrictions they cannot contact as many users as fast as bots can.

In a next step, we validated our bot analysis by contrasting our prediction with a direct visual check of the text messages for bot presence[4]. Table 1 shows that 98.23% of latent class 9 revealed bot behaviour in their text messages. In total 54.95% of all bots are present in class 9[5]. Table 2 compares bots and human users

Table 1 Manual check for bots

	Humans	Bots	
Class 9	1.77%	98.23%	100.00%
Classes 1–8, 10–13	95.16%	4.84%	100.00%

Table 2 Impact of bot presence

	Women	Age (mean)	BMI (mean)	First contacts
Humans	48.38%	37.23	23.89	69.55%
Bots	99.50%	28.74	20.58	30.45%
Total	53.97%	36.19	23.67	100.00%

[4]Due to privacy reasons the visual message check was done by the online-dating provider.

[5]A chi-square test on predicted vs. actual bots yielded a Cramer's V of 0.7156 ($\chi^2 = 1,000$, $p = 0.00$).

by means of their socio-demographic characteristics. Bots are nearly always female, about 10 years younger than average and show a favourable body-mass-index. This class of artificial actors generates 30.5% of all first contact attempts.

4 Conclusion

When analysing social web data, it is necessary to realise that there are several new issues and challenges of data quality that come into play. We discussed the presence of artificial actors called bots foiling the potential of research designs utilising web-generated process data of human interactions. Previous research on online mating behaviour for example did not consider that bots can dramatically reduce the quality of observed interactional records by suggesting a larger sample and by inducing a huge amount of artificial contact patterns. Based on assumptions of bot behaviour, we were able to construct behaviour and interaction related indicators based on web-based process-generated data. We applied a finite mixture model on the different indicators in order to encircle the artificial actors from different directions. The model identified a group that conformed to our theoretic conditions. A group of 3.72% of the actors showed a pattern where all variables indicated bot behaviour. Another class revealed similar behaviour and was interpreted as human spammers. A comparison of predicted and visually identified bots showed a high validity of our model. Bots differ clearly from human actors by means of their socio-demographic profiles. This is considerably problematic as bots produce about one third of all contact attempts. Further research is needed to improve the detection model and to analyse the impact of bots on substantial models such as regression statistics on human choosing behaviour. Furthermore, the proposed classification approach can be useful whenever spam behaviour of bots or human spammers occurs and log-file data on behaviour and interaction is recorded. For example social web platforms as well as instant messenger systems could be improved by removing unwanted bots. Future research has to deal with the distinction between such unwanted and requested bots on the one hand and the differentiation between human spammers and artificial behaviour on the other hand. The latent variable approach was useful to get first insights into the nature of artifical web behaviour so that future research can rest on these findings. Since the goal of our research is the assignment of actors to a fixed number of classes (actually bot vs. non-bot) future work will focus on the improvement of our results by applying different machine learning techniques, such as classification trees.

References

Chen K, Jiang J, Huang P, Chu H, Lei C, Chen W (2006) Identifying MMORPG bots: A traffic analysis approach. In: Proceedings of the ACM SIGCHI International Conference on Advances in Computer Entertainment Technology, ACM, New York, vol 266:4

Chen K, Liao A, Pao HK, Chu H (2009) Game bot detection based on avatar trajectory. In: Stevens SM, Saldamarco SJ (eds) Proceedings of the 7th international Conference on Entertainment Computing, Springer, Berlin, Heidelberg, Lecture Notes In Computer Science, vol 5309, pp 94–105

Dvorak JC, Pirillo C, Taylor W (2003) Online! The Book. Prentice Hall, Upper Saddle River, NJ

Fink RD, Liboschik T (2010) Bots–Nicht-menschliche Mitglieder der Wikipedia-Gemeinschaft. Working Paper. Online: http://www.wiso.tu-dortmund.de/wiso/ts/Medienpool/AP-28-Fink-Liboschik-Wikipedia-Bots.pdf (accessed: 13. February 2011).

Gianvecchio S, Xie M, Wu Z, Wang H (2008) Measurement and Classification of Humans and Bots in Internet Chat. USENIX Security Symposium pp 155–170

Hayati P, Potdar V, Talevski A, Firoozeh N, Sarenche S, Yeganeh EA (2010) Definition of spam 2.0: New spamming boom. In: IEEE Digital Ecosystem and Technologies, Dubai, UAE

Poggi N, Berral JL, Moreno T, Gavalda R, Torres J (2007) Automatic detection and banning of content stealing bots for e-commerce. In: NIPS Workshop on Machine Learning in Adversarial Environments for Computer Security., Whistler, Canada, pp 7–8

Schmitz A, Skopek J, Schulz F, Klein D, Blossfeld HP (2009) Indicating mate preferences by mixing survey and process-generated data. The case of attitudes and behaviour in online mate search. Historical Social Research 34(1):77–93

Vermunt J, Magidson J (2002) Latent class cluster analysis. In: Hagenaars J, McCutcheon (eds) Applied latent class analysis. Cambridge University Press, Cambridge, UK, pp 89–106