Andree Thieltges / Simon Hegelich

Manipulation in sozialen Netzwerken

Risikopotenziale und Risikoeinschätzungen

Einleitung

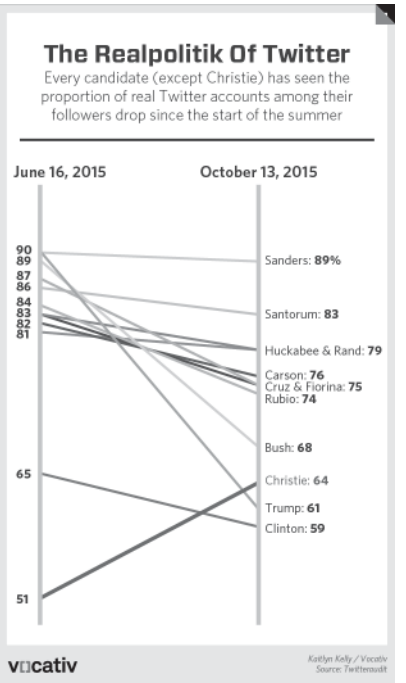
Digitale soziale Netzwerke und die in ihnen stattfindenden Interaktionen sowie die darin auffindbaren Informationen sind inzwischen ein wichtiger Teil der weltweiten Kommunikation. Dies lässt sich nicht nur an der stetig steigenden Nutzerzahl von Netzwerken wie Facebook oder Twitter ablesen, sondern auch an der steigenden Anzahl versuchter Manipulationen die dort stattfinden: Dabei wird häufig und mit unterschiedlichsten Mitteln versucht, Falschinformationen zu verbreiten oder die Popularität bestimmter Nutzerinnen und Nutzer mit massenhafter Zustimmung zu fördern. Dies bezieht sich zumeist auf das Tagesgeschehen, d.h. politische und ökonomische Aussagen und Ereignisse. Zudem werden Personen die in der Öffentlichkeit stehen sowie deren Ziele und Zwecke gezielt unterstützt oder marginalisiert. Ein aktuelles Beispiel dafür, liefert der amerikanische Präsidentschaftswahlkampf: Das amerikanische Onlinemagazin »vocativ« analysierte auf der Grundlage eines Twitter-Tools (Twitter-Audit) die Entwicklung der »follower« aller Kandidatinnen und Kandidaten des Präsidentschaftswahlkampfes. Innerhalb von 4 Monaten (von Juli bis Oktober 2015) war der Anteil der menschlichen »follower« bei fast allen Kandidatinnen und Kandidaten rückläufig (s. Abbildung 1).

Im Hinblick auf die Gesamtzahl bedeutet dieses Absinken der menschlichen »follower« den Anstieg sogenannter »fake-follower«, d.h. maschinell verwaltete Nutzerprofile, welche vorgeben menschliche Nutzer zu sein und im Netzwerk Twitter die politischen Ansichten der Kandidatinnen und Kandidaten massenhaft weiterleiten (retweet) oder diese selbstständig kommentieren und darüber massenhaft verbreiten (tweet). Die mutmaßliche Zielsetzung dieser Manipulation: Der Öffentlichkeit/Wählerschaft über den Zuwachs an Unterstützern und an Popularität einen erfolgreichen Wahlkampf zu suggerieren und darüber eine potenzielle Wahlbereitschaft zu beeinflussen.

Die »fake-follower« sind allerdings nur eine Form der automatisierten Manipulation die sich aktuell in digitalen sozialen Netzwerken finden lässt: Der Obergriff »social bots« oder deren Zusammenschluss zu einem »social botnet« beschreibt auf Algorithmen basierende maschinelle Nutzer, die vollautomatisiert konkrete Aufgaben in digitalen sozialen Netzwerken abarbeiten und dabei weitgehend autark agieren. So können social bots inzwischen mit menschlichen Nutzerinnen und Nutzern interagieren und kommunizieren, Gesprächen folgen und dem Kontext entsprechende Kommentare er-

 DOI: 10.5771/0044-3360-2017-4-493

Abbildung 1: Abnahme der menschlichen Twitter-Follower im US-amerikanischen Präsidentschafts-vorwahlkampf. Quelle: Twitteraudit



widern, sich selbstständig in digitalen sozialen Netzwerken anmelden und Benutzerprofile anlegen oder sich mit anderen (menschlichen) Usern verbinden (über die dafür vorgesehene Infrastruktur des Netzwerks, beispielsweise »like«, »friend«, »follow«). Dabei versuchen sie zumeist ihre maschinelle »Natur« zu verschleiern, was grundsätzlich darauf hindeutet, dass die Personen die hinter den social bots stehen, ihre Identität und die Zielsetzung des social bot – Einsatzes nicht preisgeben möchten. Auch wenn es sich bei dieser Feststellung über die »Hintermänner« sowie die gewünschte Wirkung, die mit den social bots erzielt werden soll um eine durchaus begründete Annahme handelt, kann sie bisher nicht bewiesen werden. Nachweisbar ist allerdings, dass inzwischen social bots massenhaft in bestimmten Kontexten eingesetzt werden, um Einfluss zu nehmen.

Diese potentiellen und tatsächlichen Manipulationen schaffen Risiken, die sowohl Nutzerinnen und Nutzer von sozialen Netzwerken als auch netzwerkexterne Personen betreffen. Zudem ist es möglich, dass gezielt verbreitete Falschmeldungen und deren Weiterleitung über das soziale Netzwerk hinaus, zu gesellschaftlichen Problemlagen führen können. Um zu verstehen, wie sich aus (teil-) algorithmisierten Manipulationen in sozialen Netzwerken Gefährdungen für Individuen und Gesellschaften entwickeln können, muss zunächst erklärt werden, wie diese funktionieren. Deshalb wird

im ersten Teil dieses Artikels anhand von Beispielen aus der Politik und der Ökonomie erläutert, wie social bots und Trolle massenhaft dazu eingesetzt werden, um die interne »Infrastruktur« der digitalen sozialen Netzwerke zu unterwandern. Ein vielfach unterschätztes Risiko ist dabei, dass social bots inzwischen in der Lage sind, menschliches Verhalten fast perfekt zu adaptieren. Dies tun sie zumeist, um nicht als Maschinen enttarnt und unglaubwürdig zu werden. Aus den Netzwerk- und Nutzermanipulationen die hier besprochen werden, entstehen potenzielle Risiken und damit verbunden Gefährdungen auf unterschiedlichen Ebenen und für unterschiedliche Nutzer und Nutzergruppen. Diese werden am Ende des jeweiligen Absatzes zusammengefasst.

Die durch social bots und Trolle[1] »gelenkten« Manipulationsversuche dienen oftmals dazu, einzelnen Nutzern oder Nutzergruppen sowie Netzwerktrends oder abgeschlossene Kontexte zu beeinflussen. Sie unterscheiden sich allerdings sowohl in ihrer Wirkweise als auch in ihrer Zielsetzung: Es gibt Beeinflussungsversuche, die beispielsweise »nur« auf eine kontextbezogene Diskussion innerhalb des sozialen Netzwerks fokussieren und solche, die (weit) darüber hinausgehen. Daraus entwickeln sich unterschiedliche Risiken, die netzwerkinterne und/oder -externe Gefährdungen für Individuen und »Gesellschaft« nach sich ziehen. Im zweiten Teil des Artikels wird versucht, eine Risikomatrix zu entwickeln, die diese unterschiedlichen Gefährdungen einordnet. Zwei wesentliche Punkte die für die Verbreitung von manipulativen Inhalten eine Rolle spielen, sind die interne »Infrastruktur« sozialer Netzwerke und das Verhalten der Nutzerinnen und Nutzer. Anhand von verschiedenen theoretischen Modellen (aus der Epidemiologie und der Spieltheorie) wird verdeutlicht, wie die »soziale« Dynamik und die Beziehung der Nutzerinnen und Nutzer aufeinander zur Verbreitung von Falschmeldungen und manipulierten Inhalten im Netzwerk führen können. Daran anschließend werden die Gelingensbedingungen für eine netzwerkexterne »Verlängerung« von Manipulationen beschrieben und abschließend eine Einschätzung über die verschiedenen Eintrittswahrscheinlichkeiten versucht.

»Korrigierte« Diskussionsverläufe und die »influence bots«

Von allen algorithmisierten und teilalgorithmisierten Manipulationsformen in digitalen sozialen Netzwerken geht ein generelles Risiko aus. Sie zielen auf eine Beeinflussung der Nutzerinnen und Nutzer: Bewusste Täuschung, Betrug und Intransparenz sind die Mittel, mit denen diese Manipulationen versucht werden. Passend zur internen Struktur der meisten digitalen sozialen Netzwerke werden häufig große Massen von Nachrichten, Diskussionsbeiträgen oder Befürworter respektive Gegner maschinell erzeugt. Das oben genannte Beispiel der »fake follower« der US-amerikanischen Präsident-

1 Unter dem Begriff »Troll« versteht man eine real existierende »Servicekraft«, die für Meinungsmache in Internetforen und Kommentarseiten beauftragt und bezahlt wird. Dabei besteht die Manipulation darin, dass die Trolle in den jeweiligen Foren eine große Anzahl Kommentare »posten« und somit die authentische Diskussion über eine Person oder einen Sachverhalt verwässern, sowie gezielt dahin führen, dass die Diskussion abbricht oder versuchen, diese zu Gunsten ihres Auftraggebers zu gestalten.

schaftskandidatinnen und -kandidaten kann beispielsweise als Versuch interpretiert werden, die Öffentlichkeit bewusst über die Popularität der einzelnen Kandidatinnen und Kandidaten zu täuschen, um so einen Einfluss auf die Entscheidung der Wählerinnen und Wähler zu nehmen. Intransparenz wird beispielsweise über die »Marginalisierung« von authentischen Diskussionen geschaffen. Dazu wird mit Hilfe von social bots oder »Trollen« eine kontextbezogene Diskussion »infiltriert« um darin massenhaft eine bestimmte Meinung zu posten. Darüber entsteht der Eindruck, dass die von den social bots oder den Trollen verbreitete Meinung, tatsächlich von der Mehrheit der Diskussionsteilnehmerinnen und -teilnehmer vertreten wird.

Social Bots und Trolle, die versuchen, über das massenhafte Verfassen von Beiträgen eine Diskussion zu manipulieren, kommt dabei die chronologische Struktur der meisten Diskussionsseiten und –foren entgegen: Je mehr Beiträge von ihnen gepostet werden, desto mehr werden die Diskussionsbeiträge und Kommentare von »echten« Usern verdrängt, respektive in der chronologischen Abfolge nach unten verschoben. So entsteht der Eindruck, dass eine bestimmte Meinung die Diskussion dominiert. Ein aktuelles Beispiel für diese Strategie in Deutschland ist die Diskussion um Carsten Maschmeyer und dessen vermeintliche Trolle: Der Tagesspiegel[2] berichtete über »verdächtig positive Kommentare«, die, wann immer Pressemeldungen über Maschmeyer, dessen Frau oder einen befreundeten Fußballtrainer in Online-Foren diskutiert werden, die dort geäußerte Kritik überlagern und/oder marginalisieren sollen.

Die Strategie, reale Diskussionen in digitalen sozialen Netzwerken auf diese Weise zu »unterwandern«, zu unterbrechen oder als Plattform für eine bestimmte Meinung zu benutzen, ist inzwischen eine weit verbreitete Art der Einflussnahme auf den öffentlichen Meinungsaustausch in sozialen Netzwerken. Im US-amerikanischen Wahlkampf wurde kürzlich bekannt, dass die demokratische Partei ihre Kandidatin Hillary Clinton mit einem »political action commitee« (PAC)[3] unterstützt, das zum Ziel hat kritische Äußerungen über Clinton und ihre politische Position mit gezielten Kommentaren zu relativieren[4]. Auch hier kommen bezahlte und beauftragte Trolle zum Einsatz, die bestimmte Hashtags[5] in Twitter und bestimmte Gruppen im Netzwerk Facebook[6] gezielt überwachen und dann reagieren, wenn ein anderer User dort kritische Kommentare über Clinton hinterlässt.

2 Sonja Álvarez, Maschmeyer und Freunde – verdächtig positive Kommentare. PR-Angriff oder wahre Verehrung in: *Der Tagespiegel,* http://www.tagesspiegel.de/wirtschaft/pr-angriff-oder-wahre-verehrung-maschmeyer-und-freunde-verdaechtig-positive-kommentare/13492040.html, (Zugriff am 15.6.2016.).

3 Dieses PAC hat sich zweckmäßig »Correct the Record« genannt.

4 Clare Foran, »A $1 Million Fight Against Hillary Clinton's Online Trolls. A supper PAC has a plan to defend the Democratic presidential front-runner and her supporters on social media. Will it work?« in: *The Atlantic* (2016), http://www.theatlantic.com/politics/archive/2016/05/correct-the-record-online-trolls/484847/ (Zugriff am 15.6.2016).

5 Bspw. #NeverClinton, #NeverHillary, #CrookedHillary.

6 Bspw. »Stop Hillary Clinton in 2016«, »Americans Against Hillary Clinton«.

Die »Defense Advanced Research Projects Agency (DARPA)« veranstaltete im letzten Jahr einen Wettbewerb[7], um sogenannte »influence bots« im Netzwerk Twitter aufzuspüren und unschädlich zu machen. Vorangegangen war eine Kontroverse über die tatsächliche Anzahl von social bots im digitalen Netzwerk Twitter[8]. Dabei definierten die teilnehmenden Wissenschaftlerinnen und Wissenschaftler »influence bots« als algorithmisch gesteuerte Twitter-Accounts die versuchen, Twitter-Diskussionen innerhalb eines bestimmten Kontextes oder Themas gezielt zu beeinflussen [9].

Die von den »influence bots« betriebene, massenhafte Verbreitung/Weiterleitung einer bestimmten Meinung oder eines bestimmten Inhalts, bedient sich der vorhandenen Kommunikationsinfrastruktur des jeweiligen Netzwerks: In Netzwerken wie beispielsweise Facebook bilden alle Nutzerinnen und Nutzer den sogenannten »social graph«. Darin werden die Aktivitäten des Users, ihre/seine Verbindungen mit anderen Nutzerinnen und Nutzern sowie die Häufigkeit ihrer/seiner eingestellten Kommentare und Inhalte gemessen. Zudem wird registriert, wie oft ihre/seine Beiträge von anderen Benutzerinnen und Benutzern kommentiert, »geliked« und/oder auf andere Art und Weise weiterverbreitet werden. Aus der Gesamtheit dieser »Benutzer-Vermessung« und dem Vergleich mit allen anderen Nutzerinnen und Nutzern, wird eine Positionierung des einzelnen Users im sozialen Netzwerk errechnet. Dementsprechend ergibt sich ein individuelles Ranking für verschiedene Nutzerkategorien, was dazu führt, dass beispielsweise Nutzerinnen und Nutzer, deren Inhalte und Kommentare massenhaft von »influence bots« weiterverbreitet werden, »automatisch« in das Zentrum des Netzwerks vorrücken und somit netzwerkintern einen größeren Einfluss bekommen. Ihre Nachrichten und Kommentare werden dementsprechend schneller und weiter verbreitet. Inzwischen ist die Möglichkeit über »influence bots«, fake follower oder ähnliche (teil-) algorithmisierte und automatisierte Profile, Einfluss auf die Informationen und Kommunikationen in sozialen Netzwerken zu nehmen, für politische Akteure immer öfter ein probates Mittel, um ihre Ziele und Zwecke voranzubringen.

7 Den sogenannten DARPA Twitter Bot Challenge.

8 2014 veröffentlichte das Unternehmen Twitter gemeinsam mit der US-amerikanischen Security and Exchange Commission einen Bericht, in dem festgestellt wird, dass «up to approximately 8.5% of all active users used third party applications that may have automatically contacted our servers for regular updates without any discernible additional user-initiated action." In der öffentlichen Berichterstattung existiert seitdem die Lesart, dass die hier genannten 8,5% der Twitter-User, die sich automatisch über »Anwendungen von Fremdanbietern« einwählen (social) bots sind (bei aktuell ca. 320 Millionen Twitter-Usern wären das 27,2 Millionen social bots alleine in diesem Netzwerk). Dieser Darstellung wurde von Twitter allerdings widersprochen (vgl. Jack Linshi, *Twitter Refutes Report That 23 Million Active Users Are Bots. Clarification follows misinterpretations of its SEC filing, which disclosed how many users are on third party apps,* http://time.com/3103867/twitter-bots/ (Zugriff am 6.7.2016).

9 V. S. Subrahmanian / Amos Azaria / Skylar Durst / Vadim Kagan, *The DARPA Twitter Bot Challenge,* https://www.google.de/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwiS-cT-6qnNAhWJoRQKHShCApYQFggcMAA&url=https%3A%2F%2Farxiv.org%2Fpdf%2F1601.05140&usg=AFQjCNEZX8p_XZbYMDImWUQHu_6C-aWdJw (Zugriff am 15.6.2016).

Wie real diese Gefährdung und damit die Eintrittswahrscheinlichkeit einer gezielten Manipulation durch »influence bots« bereits geworden ist, zeigt die herbeigeführte Unterbrechung der Twitterdiskussionen im Nachgang der russischen Parlamentswahlen. Dabei gelang es Unbekannten mit Hilfe von über 25.000 »fake accounts« die Diskussionen auf Twitter mit massenhaft unverständlichen und widersprüchlichen tweets zum Stillstand zu bringen[10]. Den potenziellen Versuch die öffentliche Meinung zu beeinflussen, kann man aktuell der in den USA laufenden Berichterstattung über die Spitzenkandidaten des diesjährigen US-Präsidentschaftswahlkampfs entnehmen: Unter den Twitter-Followern des republikanischen Kandidaten Donald Trump befinden sich beispielsweise fake follower, die vorgeben, der mexikanischen Wählerschaft anzugehören und die Trump's Wahlkampf mit ihren Kommentaren aktiv unterstützen[11]. Dabei liegt der Verdacht nahe, dass mit diesen fake followern versucht wird die lateinamerikanische Wählerschaft zu Gunsten von Trump zu beeinflussen. Zudem tauchten in den Vorwahlkämpfen immer wieder fake follower im Netzwerk Twitter auf, die auf die eingelegten Beschwerden der Federal Communication Commission (FCC) gegen Ted Cruz hinwiesen, anscheinend mit dem Ziel diesen Konkurrenten für die Wählerschaft zu diskreditieren[12].

All die hier genannten Manipulationen weisen zunächst auf ein bei der Benutzung von digitalen sozialen Netzwerken immanentes Risiko: Politische Propaganda, »fake news«, manipulative Inhalte sowie authentische Informationen und »reale« Nachrichten stehen im Netzwerk nebeneinander und müssen vom User unterschieden werden. Allerdings wird durch die oben genannten Beispiele auch klar, dass die Gefährdung, manipulativen Inhalten aufzusitzen, Meinungsmache oder gezielte Falschinformation nicht als solche zu erkennen, zunimmt: Das objektive Unterscheidungsmerkmal der Quantität, dass beispielsweise Meinungstrends anzeigt, oder die Verbreitungsgeschwindigkeit von Informationen und Nachrichten innerhalb eines Netzwerks regelt, ist praktisch »ausgehebelt«. Dementsprechend steigt die oben genannte Gefahr für alle Benutzerinnen und Benutzer von digitalen sozialen Netzwerken, die sich dort über tagesaktuelle Nachrichten informieren oder versuchen Meinungstrends oder kontextbezogene Diskussionen auszuwerten.

Wer einmal lügt, dem glaubt man nicht…

Aus der Manipulationsform der »fake follower leitet sich allerdings auch noch ein weiteres Risiko ab: Da es oftmals schwierig ist, den Urheber und die Zielsetzung der fake follower einwandfrei nachzuweisen, können diese auch als »Waffe« benutzt werden,

10 Kurt Thomas / Chris Grier / Vern Paxson, »Adapting Social Spam Infrastructure for Political Censorship«, in: Engin Kirda (Hg.), *5th USENIX Workshop on Large-Scale Exploits and Emergent Threats, LEET '12*, San Jose, CA, USA 2012.

11 Sam Wooley / Phil Howard, »Bots Unite to Automate the Presidential Election« in: *Wired Magazin* (2016), http://www.wired.com/2016/05/twitterbots-2/ (Zugriff am 15.6.2016).

12 Aaron Sankin, *Who's behind this army of Donald Trump-loving Twitter bots?*, http://www.dailydot.com/politics/trump-twitter-bots-ruffini/ (Zugriff am 15.6.2016).

um beispielsweise Konkurrentinnen und Konkurrenten öffentlich zu diskreditieren. Dazu ein Beispiel aus der deutschen Parteienlandschaft, in der diese Art der Manipulation in digitalen sozialen Netzwerken ebenfalls seit längerem bekannt ist: 2012 berichtete der ZDF-Blog »Hyperland« über den sprunghaften Anstieg der Twitter Follower der CDU [13].Ein Jahr später verfünffachte sich die Zahl der FDP-Twitter Follower innerhalb weniger Tage. In einer Analyse der Internetplattform »Pluragraph«, die die Aktivitäten von Parteien und Stiftungen in digitalen sozialen Netzwerken untersuchte, wurde damals festgestellt, dass 81% der neuen Anhänger fake follower sind. Wie oben bereits angemerkt, ist es aufgrund der häufig betriebenen Verschleierungsbemühungen, die den social bots immanent sind[14], oftmals »nur« möglich mehr oder weniger begründete Annahmen über deren Zielsetzungen zu formulieren: Wie bei dem rasanten Anstieg der fake follower der Präsidentschaftskandidatinnen und -kandidaten im US-amerikanischen Wahlkampf, legt der Fall der FDP die Interpretation nahe, dass sich die Partei der stetig wachsenden Bedeutung von digitalen sozialen Netzwerken bewusst ist und sich deshalb in diesen Netzwerken um eine massenhafte Unterstützung und eine wachsende Popularität bemüht. Hierbei kommen demnach auch manipulative Methoden zum Einsatz[15].

13 Mirjam Hauck, »Die wundersame Follower-Vermehrung der CDU« in: *Süddeutsche Zeitung* 2012 (2012), http://www.sueddeutsche.de/digital/twitter-die-wundersame-follower-vermehrung-der-cdu-1.1411578 (Zugriff am 15.6.2016).

14 Das »Soziale« an social bots ist die Bestrebung der Programmierer, mittels Quantifizierung und anschließender Überführung in einen Algorithmus, das menschliche Verhalten in digitalen sozialen Netzwerken zu adaptieren, um als »authentischer User« unerkannt an Diskussionen und Meinungsbildungsprozessen mitzuwirken (vgl. dazu Tim Hwang / Ian Pearce / Max Nanis, »Socialbots: Voices from the Fronts« in: *Interactions* (April 2012), S. 38–45; Scott A. Golder / Michael W. Macy, »Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures.« in: *Science*, 333, 6051 (2011), S. 1878–1881; Carlos A. Freitas / Fabrício Benevenuto / Saptarshi Ghosh / Adriano Veloso, »Reverse Engineering Socialbot Infiltration Strategies in Twitter« (2014), http://arxiv.org/abs/1405.4927v 1 (Zugriff am 15.6.2016). Simon Hegelich / Dietmar Janetzko, »Are Social Bots on Twitter Political Actors? Empirical Evidence from a Ukrainian Social Botnet«, in: Association for the Advancement of Artificial Intelligence (Hg.), *Tenth International AAAI Conference on Web and Social Media*, Palo Alto 2016, S. 579–582 stellen in ihrer Analyse eines social botnet das zum Ukrainekonflikt auf Twitter pro Tag 60.000 tweets postet, zudem fest, dass dort nicht nur politische Meinungsmache betrieben wird, sondern von den social bots auch tweets über Sport, Witze und Kalendersprüche gepostet werden. Dies, so die Schlussfolgerung, geschieht zum Zweck der Verschleierung des maschinellen Ursprungs der geposteten Inhalte.

15 Die aktuelle Berichterstattung zu einer anderen Facette der stattfindenden »manipulativen Wahlkampfpraktiken« unterstützt diese These: Anhand von großen Mengen gesammelter Daten aus dem Internet und verschiedenen digitalen sozialen Netzwerken ist es inzwischen möglich geworden, potenzielle Wähler eines politischen Lagers ausfindig zu machen und aufgrund bestimmter (Verhaltens-) Merkmalen dieser Personen sogar vorauszusagen, wie man diese am besten anspricht. Mit den Informationen aus diesem sogenannten »psychographic targeting« werden dann Wahlkämpfer zu Hausbesuchen geschickt (Tom Hamburger, *Cruz campaign credits psychological data and analytics for its rising success,* https://www.washingtonpost.com/politics/cruz-campaign-credits-psychological-data-and-analytics-for-its-rising-success/2015/12/13/4cb0baf8-9dc5-11e5-bce4-708fe33e3288_story.html (Zugriff am 25.7.2016).

Die FDP bestritt damals allerdings vehement, für die fake follower gezahlt zu haben und forderte den Netzwerkbetreiber Twitter umgehend auf, diese zu löschen. Da der tatsächliche Auftraggeber nicht ermittelt werden konnte, wäre es also ebenfalls denkbar, dass ein offensichtlicher Manipulationsversuch von der politischen Gegenseite oder anderen Akteuren gesteuert ist, um die Partei in den digitalen sozialen Netzwerken und darüber hinaus in Misskredit zu bringen, ihre Vertrauenswürdigkeit zu schmälern und ihren authentischen Zuspruch innerhalb und außerhalb der digitalen Netzwerke zu torpedieren. Auf Basis dieser möglichen Manipulationsstrategie, ergibt sich ein generelles Risiko für Personen und Institutionen, die – beispielsweise aufgrund einer exponierten gesellschaftlichen Stellung – ihren Netzwerkauftritt zu Präsentations- und Repräsentationszwecken nutzen. Das »Unterschieben« von maschinell generierten Freunden und Folgern ist durchaus eine denkbare und überaus wirksame Manipulation: Einmal als »Nutznießer« von massenhaften fake followern enttarnt, wird es schwierig zu beweisen, dass man Opfer und nicht Täter ist.

»Spread the news…«

Ein zunehmendes Risiko bezogen auf die Manipulationsversuche in sozialen Netzwerken besteht darin, dass sich die Falschmeldungen und Desinformationen außerhalb des Netzwerks weiterverbreiten und in der »realen« Welt Schaden anrichten. Hier spielt oftmals die Adaption menschlichen Verhaltens als Täuschungsstrategie eine entscheidende Rolle. Als Basis für die Diffusion von Falschmeldungen in unterschiedlichen Nachrichtenkanälen, dient zunächst die bereits oben genannte Strategie der massiven Infiltration von social bots oder Trollen: Je *höher die Quantität* gleichlautender Kommentare und Falschmeldungen zu einem bestimmten Kontext, desto wahrscheinlicher, dass diese als »glaubwürdig« erscheint und deshalb als authentische Information behandelt werden. Allerdings sind social bots zunehmend in der Lage, menschliches Verhalten täuschend echt zu simulieren. Eine sehr ausgereifte Möglichkeit der maschinellen Adaption menschlichen Verhaltens und der damit verbundenen unerkannten Manipulation wurde auf der diesjährigen »Black Hat«-Konferenz in Las Vegas vorgestellt: Das Konzept des »automated spear phishings« ermöglicht so etwas wie den perfekten »Netzwerk-Freund« für einen beliebigen User zu erschaffen. Der Algorithmus analysiert die Nachrichten der Benutzerin/des Benutzers und versucht auf Grundlage der Auswertung dieser Daten selbständig Nachrichten zu generieren, die für diesen bestimmten User von großem Interesse sind. Die Nachrichten werden mit einem Link verbunden, der auf eine Website mit Schadsoftware verweist. Diese Manipulationsform ist eine zugespitzte Weiterentwicklung des sogenannten »phishing«: Dabei werden massenhaft Nachrichten verbreitet, von denen die Manipulantin/der Manipulant hofft, dass sie für viele Nutzerinnen und Nutzer interessant sind. Im Vergleich zu dieser »herkömmlichen« Art es phishings, bei der ca. 5% der User die Website mit der Schadsoftware tatsächlich öffnen, erreicht das »spear phishing« über die Anpassung der Nachrichten an die Interessen der Zielperson eine Erfolgsrate von ca. 50%. Das Ge-

fährdungspotenzial, das von einer solchen Manipulationsform ausgeht, ist immens: Einmal automatisiert und mit einem social bot Profil verbunden, kann praktisch jede Nutzerin/jeder Nutzer von solchen »automated spear phishing bots« angegriffen und ausgespäht werden [16].

Inzwischen ist auch das vollautomatisierte »clonen« kompletter Benutzerkonten eine Strategie, die darauf zielt bei einer bestimmten Zielgruppe den Eindruck zu erwecken, eine vertrauenswürdige Person oder Institution zu sein. Wie realistisch solche »clone-bots« Inhalte selbstständig adaptieren können, deutet das auf Twitter sehr populäre Beispiel des DeepDrumpf-Bot an: Der am MIT entwickelte Algorithmus erkennt selbstständig Muster in den Texten und Reden von Donald Trump und kompiliert diese zu neuen, »authentischen« Aussagen, die dann auf dem gleichnamigen Twitter-Account gepostet werden [17].

Welche Risiken der manipulative Einsatz von »clone-bots« in ökonomischen Kontexten entfalten kann, zeigt das Beispiel des Chipherstellers »Audience Inc.« : 2013 sank dessen Aktienkurs innerhalb weniger Sekunden um 25%. Grund dafür war eine gezielte Falschmeldung über die finanzielle Schieflage des Unternehmens. Ein »geklontes« Benutzerkonto einer US-amerikanischen Firma, die sich auf »Leerverkäufe« spezialisiert hat, war Ausgangspunkt dieser »erfolgreichen Manipulation« [18]. Das wohl bekannteste Beispiel einer Beeinflussung mit weitreichenden ökonomischen Folgen ist das des »New Yorker Börsencrashs« aus dem Jahr 2013, mit einem zwischenzeitlichen Verfall des Aktienkurses um mehr als 143 Punkte. Dieser wurde technisch allerdings nicht durch »clone bots«, sondern durch einen Hackerangriff einer Gruppe realisiert, die sich die »syrische Internet Armee« nannte. Dennoch verdeutlichen der Verlauf und die Folgen dieser Manipulation die potenzielle Gefährdung, die von Falschmeldungen aus digitalen sozialen Netzwerken inzwischen ausgeht. Der Twitter-Kanal des Nachrichtendienstes »Associated Press«(AP) wurde infiltriert und eine gezielte Falschinformation lanciert. Diese berichtete über einen Angriff auf das Weiße Haus, bei dem Präsident Obama verletzt worden sei. Über die bestehenden Follower-Abonnements des Associated Press-Kanals wurde die Nachricht binnen Sekunden an mehreren tausend Nutzer im Netzwerk Twitter weitergeleitet, die ihrerseits diese Nachricht retweeteten oder als »favorit« markierten. Getäuscht von der »authentischen Quelle« zogen etliche andere Nachrichtenkanäle nach[19]. Das Resultat dieser gezielten Manipulation: Innerhalb von 4 Minuten verlor der Standard & Poor 500 Index durch Panikverkäufe 143

16 Simon Hegelich, »Invasion der Meinungs-Roboter« in: *Analysen und Argumente der Konrad Adenauer Stiftung*, Nr. 221 (2016), http://www.kas.de/wf/doc/kas_46486-544-1-30.pdf?160927132147 (Zugriff am 29.9.2016), S. 1–9.

17 Stuart Dredge, »Deep Drumpf: the Twitter bot trying to out-Trump the Donald. MIT project uses artificial-intelligence algorithm to learn Republican frontrunner's speech patterns before publishing 'remarkably Trump-like statements'« in: *The Guardian* (2016), https://www.theguardian.com/technology/2016/mar/04/donald-trump-deep-drumpf-twitter-bot (Zugriff am 16.6.2016).

18 John McCrank / David Gaffen, *Hoax tweets send Audience shares atwitter,* New York 2013.

19 Tero Karppi / Kate Crawford, »Social Media, Financial Algorithms and the Hack Crash« in: *Theory, Culture & Society* 33, Nr. 1 (2015), S. 73–92.

Punkte und 136,5 Milliarden US-Dollar[20]. Viele Broker und Spekulanten hatten den tweet oder den retweet gelesen, diesen als gesicherte Information eingestuft und daraufhin große Aktienpakete von US-amerikanischen Firmen abgestoßen. Dieses Beispiel unterstreicht einerseits den »Einfluss«, den digitale soziale Netzwerke und die dort verbreiteten Informationen inzwischen für den finanzkapitalistischen Teil der Ökonomie haben: Digitale soziale Netzwerke bieten große Mengen von Informationen die in sehr kurzer Zeit bereitstehen und eine Bewertung der ökonomischen Lage von Unternehmen sowie soziökonomischen Entwicklungsprognose von ganzen Weltregionen erlauben. Andererseits zeigt diese Beispiel das Gefährdungspotential gezielter Falschmeldungen die durch »vermeintlich verlässliche« Informationsquellen in digitalen sozialen Netzwerken entstehen können: Je authentischer der Verfasser der Nachricht erscheint, oder je größer die Quantität der tweets und retweets auf einer vertrauten Profilseite, umso größer das Risiko, dass die manipulative Information als »wahr« eingestuft wird, Verbreitung findet und Folgen für die sozioökonomische Realität zeitigt.

Durch die massenhafte Verbreitung von Falschmeldungen entwickeln sich zudem mittel- und langfristig »retrospektive« Risiken für die netzwerkexterne Ebene, da sie im Netzwerk verbleiben und somit darauf beruhende Auswertungen und Analysen verfälschen können: Unter dem Schlagwort »big data« wird immer öfter und in unterschiedlichen ökonomischen Bereichen versucht, das Nutzerverhalten in sozialen Netzwerken einzuschätzen und auszuwerten. Dies hat zum Ziel, gesellschaftliche Trends zu erkennen und beispielsweise für die Produktentwicklung nutzbar zu machen oder (potenzielle) Kundschaft zu analysieren und zu kategorisieren sowie Erkenntnisse über Marktanteile und die Position der eigenen Marke/des eigenen Produktes zu gewinnen[21]. Ähnliche big data - Analysen kommen inzwischen auch in der Politik zum Einsatz: In Deutschland eher noch selten[22], entwickeln sich international immer mehr Anbieter für Politikberatung, die auf Grundlage von Netzwerkanalysen agieren und Handlungsempfehlungen erstellen. Akteure wie beispielsweise »Civics Analytics« (für Hillary Clinton) oder »Cambridge Analytica« (von Donald Trump beauftragt) schätzen gesellschaftliche Meinungstrends ein und geben darauf beruhende »Ratschläge« für eine kurzfristige politische Strategie (beispielsweise im Wahlkampf) oder eine zukünftige politische Ausrichtung. Wie die oben genannten Beispiele zeigen, ist es durch den massenhaften Einsatz von social bots sowie ihrer täuschend echten Simulation menschlichen Verhaltens möglich geworden, Meinungen/Meinungstrends und somit

20 Alina Selyukh, *Hackers send fake market-moving AP tweet on White House explosions* 2013.

21 Simon Hegelich / Cornelia Fraune / David Knollmann, »Point Predictions and the Punctuated Equilibrium Theory. A Data Mining Approach -US Nuclear Policy as Proof of Concept« in: *Policy Studies Journal*, 43 (2) (2015), S. 228–256.

22 Simon Hegelich / Morteza Shahrezaye, »The Communication Behavior of German MPs on Twitter. Preaching to the Converted and Attacking Opponents« in: *European Policy Analysis* 1, Nr. 2 (2015) zeigen in ihrer Analyse zum Kommunikationsverhalten deutscher Politiker im Netzwerk Twitter potentielle Anwendungsbereiche für zukünftige big data-Auswertungen.

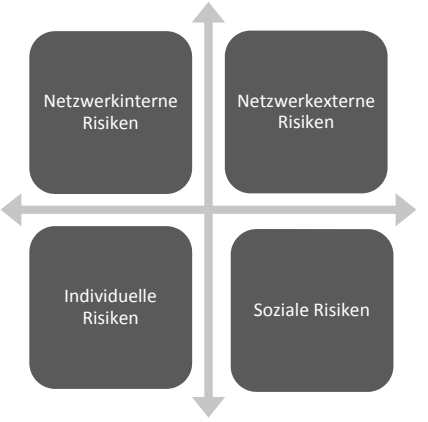
auch die darauf beruhenden Analysen zu manipulieren. Daraus resultiert ein Risikopotential für Akteurinnen und Akteure, die big data Analysen anwenden oder (politische) Entscheidungen darauf gründen. Die potenzielle Gefährdung, die von solchen manipulierten Analysen ausgeht, reicht dabei von einer einmalig fehlerhaften Beurteilung eines Trends oder einer Sachlage bis hin zu einer falschen politischen oder ökonomischen Strategie mit realen Folgen.

Eine Diffusion von Falschinformation in andere Medienformate geschieht bisher eher selten, respektive wird nicht so häufig erkannt. Wenn eine solche Weiterverbreitung allerdings »glückt«, potenziert sich auch das Manipulationsrisiko: Einmal massenhaft verbreitet, werden die manipulativen Ansichten und Inhalte auch in andere Informationskanäle getragen (beispielsweise über die (quantitative) Zusammenfassung von Twitter-Hashtags zu einem tagesaktuellen Kontext in TV, Radio, Printmedien etc.). Dies bedeutet, dass bewusst verbreitete Falschmeldungen oder gefälschte Inhalte in digitalen sozialen Netzwerken eine Legitimation durch andere Medien erfahren, wenn Journalistinnen und Journalisten diese für authentisch halten. Über eine solche »Weiterleitung« gelangen die manipulativen Inhalte zudem in Teile der Öffentlichkeit, die sich ansonsten nicht in sozialen Netzwerken informieren. Hier besteht das gravierende Risiko, dass ein manipulativer Inhalt oder eine Meinung unter Umständen eine gesellschaftliche Bedeutung bekommt, die sie durch die Diskussion im sozialen Netzwerk alleine nicht hätte erreichen können.

Risikofaktoren: Netzwerkstruktur und Nutzerverhalten

Zusammenfassend lässt sich feststellen, dass Manipulationsrisiken für unterschiedliche Ebenen des virtuellen und des realen gesellschaftlichen Lebens erkennbar sind, die sich an mehrere Gelingensbedingungen knüpfen. Dies lässt sich schematisch, beispielsweise in einer Risikomatrix darstellen (s. Abb.2):

Abbildung 2: Schematische Darstellung einer Risikomatrix in digitalen sozialen Netzwerken.



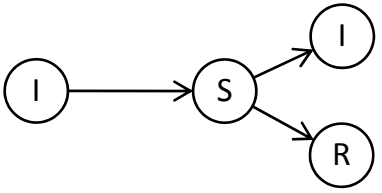
Es muss dabei zwischen netzwerkintern und -extern spezifische Risiken unterschieden werden, die sowohl den individuellen User von digitalen sozialen Netzwerken bedrohen als auch Risiken für die gesellschaftliche und/oder ökonomische Ordnung haben können. Ausgangspunkt für die auf Algorithmen beruhenden Manipulationsformen ist dabei zunächst die netzwerkinterne Struktur und Topographie. Wie oben bereits erwähnt, spielen bei der Manipulation durch social bots die quantitativen Faktoren eine entscheidende Rolle: Der Einfluss und die Popularität eines bestimmten Users verschiebt sich aufgrund von Häufigkeiten, beispielsweise über die »bloße« Anzahl der Follower (fake follower) oder Freunde, denen »netzwerkimmanent« ein Inhalt oder ein Kommentar weitergeleitet wird. Je öfter die Inhalte und Kommentare eines Users von anderen Usern geteilt, kommentiert, gefolgt oder weitergeleitet werden, umso größer sein Einfluss und eventuell auch Popularität im gesamten Netzwerk. Dabei gibt es anscheinend eine signifikanten Zusammenhang zwischen den verschiedenen netzwerkinternen Kommunikations-möglichkeiten und der Aktivität der User bei der Verbreitung von Falschinformationen: In einer Untersuchung zu dem Verhältnis von Falschmeldungen und deren Berichtigungen (dem sogenannten »fact checking«) im sozialen Netzwerk Twitter wird festgestellt, dass der Großteil der User Falschmeldungen und deren Berichtigung über die sogenannten »retweets« (Weiterleitungen der originalen Nachricht) anstatt über originäre Nachrichten (tweets) verbreiten. Bei Usern die mit manipulativer Absicht eine hohe Netzwerkaktivität aufweisen, ist dieses Verhältnis allerdings umgekehrt: »However, for top spreaders of fake news, this ratio [häufiger retweets als tweets] is much higher: These users do not retweet as much but post many original messages promoting the misinformation[23]«.

Die technische Umsetzung des jeweiligen Netzwerks, die diesen »netzwerkimmanenten« Zusammenhang von »Netzwerkaktivität« und »Nutzer-Einfluss« festlegt, ist ein Ansatzpunkt der häufig von (teil-) algorithmisierten Manipulationsformen genutzt wird. Allerdings spielt das Nutzerverhalten bei der Verbreitung von Falschmeldungen und manipulativen Inhalten ebenfalls eine gewichtige Rolle. Wie menschliches Verhalten in digitalen sozialen Netzwerken gezielte Manipulationen befördern kann, beschreibt das aus der Epidemiologie entliehene SIR-Modell: In diesem recht einfachen Modell, dass die Wahrscheinlichkeit von Informationsdiffusion erläutert, wird davon ausgegangen, dass es drei Verhaltensweisen oder »Zustände« (states) gibt, die den Informationsfluss befördern oder verhindern: Entweder man ist offen und beeinflussbar (engl.: susceptible), oder bereits mit der neuen Information »infiziert« oder man nimmt die Information nicht (mehr) auf (engl.: recovered). Der vorgestellte Diffusionsprozess von Informationen im SIR-Modell läuft nun so ab, dass die mit einer Information »infizierte« Person (I) diese an Jemanden weiterleitet der beeinflussbar ist (S). Daraufhin setzt ein Entscheidungsprozess bei (S) ein: Entweder sie/er wird ebenfalls

23 Chengcheng Shao / Giovanni Luca Ciampaglia / Alessandro Flammini / Filippo Menczer, *Hoaxy: A Platform for Tracking Online Misinformation,* http://www2016.net/proceedings/companion/p745.pdf (Zugriff am 25.7.2016).

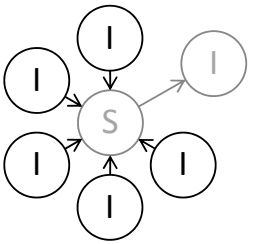
von der Information »infiziert« (und damit zu I) oder der Nachrichtenfluss endet bei ihr/ihm (R) (siehe Abb.3).

Abbildung 3: Informationsdiffusion im SIR-Model



Das Risiko von einer beeinflussbaren Person zu einer infizierten Person zu werden, erhöht sich in diesem Modell durch die Masse der direkten und indirekten Beziehungen zu anderen, bereits infizierten Personen (siehe Abb. 4).

Abbildung 4: Erhöhte Anfälligkeit für manipulative Inhalte in sozialen Netzwerken durch viele infizierte Freunde und Follower.



Bezogen auf das menschliche Verhalten in digitalen sozialen Netzwerken und deren Umgang mit Falschmeldungen verdeutlich dieses Modell, dass Benutzerinnen und Benutzer mit vielen (infizierten) Followern oder Freunden anfälliger sind für manipulative Inhalte, da die Manipulation mit höherer Wahrscheinlichkeit direkt oder indirekt zu ihnen gelangt. Das Risiko, netzwerkintern Falschmeldungen weiterzuleiten oder manipulativen Inhalten aufzusitzen, ist demnach für User die aus gesellschaftlichen, ökonomischen oder politischen Gründen viele Anhänger, Freunde oder Follower in sozialen Netzwerken haben, höher. Dies nimmt allerdings für alle andern User nichts von dem potenziellen Risiko zurück, das durch ein scheinbar weit verbreitetes Nutzerverhalten in sozialen Netzwerken befördert wird: »Few people seem to check the reliability of news before sharing them with their friends and potentially with millions of others. This is mainly due to the fact, that the Internet, in particularly social networking services, provides a complete decentralization of information on a large scale: every user is potentially a new source and often it is not trivial to establish the truth [24]«. Über Modelle aus der Spieltheorie, die ebenfalls versuchen die Diffusion von Falschinforma-

24 Marcella Tambuscio / Giancarlo Ruffo / Alessandro Flammini / Filippo Menczer, »Fact-checking Effect on Viral Hoaxes«, in: Aldo Gangemi / Stefano Leonardi / Alessandro Panconesi (Hg.), *the 24th International Conference* 2015, S. 977–982.

tionen in sozialen Netzwerken zu analysieren, lassen sich weitere Aussagen über das Nutzerverhalten und das damit verbunden Risiko einer netzwerkinternen Manipulation treffen: »The diffusion of misinformation is more related to the trust and belief in social networks[25]«. Der Informationsverbreitungsprozess wird hier als Aushandlungsprozess von unterschiedlichen Nutzerüberzeugungen (beliefs) definiert, der sich zwischen zwei Nutzerkategorien vollzieht: Den »normalen« Nutzerinnen und Nutzern und den sogenannten »zwingenden« (forceful) Nutzern[26]. Letzter kann man sich vielleicht als »Meinungsführer« in Diskussion vorstellen. In der Interaktion zweier »normaler« Nutzer kommt es zu einer Konsensüberzeugung, da sich beide gegenseitig beeinflussen. Interagiert allerdings ein »normaler« mit einem »zwingenden« Nutzer, so geschieht die Beeinflussung nur einseitig: Der normale Nutzer wird vom zwingenden Nutzer beeinflusst. Auf Grundlage dieser Annahme über das Nutzerverhalten und der daraus resultierenden Beeinflussung, lässt sich die Verbreitung von Falschinformationen in sozialen Netzwerken wie folgt zusammenfassen: »In the beginning of misinformation diffusion, all regular nodes [Netzwerkbenutzer werden in der Informatik als Knotenpunkte und die Verbindungen zwischen ihnen als sogenannte Kanten (edges) bezeichnet, Anm. d. Verf.] are assumed to have a belief probability taken from a certain probability distribution, and force individuals may hold some specific scores against the regular. Through iterations of belief exchange, the belief converges to a consensus among all individuals. The higher the consensus belief is, the more widespread misinformation is diffused over the network[27]«. Vor diesem Hintergrund werden weitere Aussagen über unterschiedliche Verbreitungsgrade von Falschinformationen sowie über das Anfälligkeitsrisiko für unterschiedliche Netzwerkstrukturen möglich: Wenn die normalen User eines Netzwerks sowohl mit einer breiten Masse weiterer normaler Nutzerinnen und Nutzern als auch mit diversen »zwingenden« Usern verbunden sind, ist für das Risiko der Verbreitung einer Falschinformation die Gesamtanzahl aller im Netzwerk befindlichen User und ihrer Verbindungen untereinander ausschlaggebend: Je kleiner die Anzahl aller im Netzwerk befindlichen Nutzerinnen und Nutzer, umso größer die Wahrscheinlichkeit einer weiten Verbreitung von Falschinformationen. Gleiches gilt für soziale Netzwerke, die aus vielen kleinen, in sich abgeschlossenen, »communities« bestehen[28]. Sind diese nicht »offen« oder unzureichend mit anderen communities vernetzt, so besteht ein erhöhtes Verbreitungsrisiko von Falschinformationen. Zum Beispiel dann, wenn eine abgeschlossene Gruppe immer aus den gleichen

25 My T. Thai / Weili Wu / Hui Xiong (Hg.), *Big Data in Complex and Social networks,* New York, London 2016.

26 Daron Acemoglu / Asuman Ozdaglar / Ali ParandehGheibi, »Spread of (mis)information in social networks« in: *Games and Economic Behavior* 70, Nr. 2 (2010), S. 194–227.

27 Liang Wu / Fred Morstatter / Xia Hu / Huan Liu, »Mining Misinformation in social Media«, in: My T. Thai / Weili Wu / Hui Xiong (Hg.), *Big Data in Complex and Social networks*, New York, London 2016.

28 In vielen sozialen Netzwerken ist es möglich, Gruppen zur Diskussion oder zum Informationsaustausch innerhalb bestimmter Kontexte einzurichten, die einer Zugangsbestätigung bedürfen. Diese sind dann in sich geschlossene und kontextbezogene communities innerhalb des Netzwerks.

Mitgliedern besteht oder sich innerhalb der Gruppe eine Kommunikationshierarchie herausgebildet hat. Ist ein »zwingender« User das »Bindeglied« zwischen zwei abgeschlossenen Nutzergruppen, so ist die Wahrscheinlichkeit, dass Falschinformationen in beide Nutzergruppen gelangen, abhängig von der »Überzeugung« dieses »zwingenden Users«[29]. Auf diese und ähnliche in sozialen Netzwerken vorfindliche Strukturen und Nutzungsverhalten, zielen die oben genannten Erscheinungsformen der social bots: Dabei wird beispielsweise versucht, social bots als zwingende Nutzer in kontextbezogene Diskussionen »einzuschleusen« oder über den massiven Einsatz einer großen Anzahl social bots die bereits bestehenden »Meinungsführer« zu marginalisieren.

Von der netzwerkinternen auf die -externe Ebene

Eine erfolgreiche netzwerkinterne Manipulation ist die Grundvoraussetzung um Falschinformationen, Betrügereien und manipulative Inhalte auf die externe Ebene zu heben. Das Risikopotenzial, dass dabei von manipulativen social bots ausgeht, ist in diesem Zusammenhang besonders hervorzuheben: Social bots sind inzwischen vom »normalen« menschlichen Nutzer nur noch selten als »Maschinen« zu erkennen, da es technisch möglich geworden ist, menschliches Verhalten in sozialen Netzwerken sehr »natürlich« zu adaptieren[30]: »In recent years, Twitter bots have become increasingly sophisticated, making their detection more difficult. The boundary between human-like and bot-like behavior is now fuzzier. For example, social bots can search the Web for information and media to fill their profiles, and post collected material at predetermined times, emulating the human temporal signature of content production and consumption – including circadian patterns of daily activity and temporal spikes of information generation[31]«.

Je authentischer ein social bot einen menschlichen User »simulieren« kann, umso wahrscheinlicher ist eine erfolgreiche Täuschung beispielsweise der netzwerkinternen Sicherheits- und Überwachungssysteme [32] oder der anderen User [33]. Das individuelle Risiko einem social bot aufzusitzen ist vor diesem Hintergrund recht wahrscheinlich: Auf der Plattform Twitter werden regelmäßig social bots Profile enttarnt, die vorgeben offizielle Nachrichtenkanäle zu sein und »authentische« Informationen über Promi-

29 Acemoglu / Ozdaglar / ParandehGheibi, Spread of (mis)information in social networks, aaO. (FN 1).

30 Aus der Schwierigkeit verschiedener Erkennungsmethoden, eine exakte Unterscheidung von menschlichen und maschinellen Usern vorzunehmen, entstehen weitergehende Risiken für bestimmte Benutzerinnen und Benutzer. Vgl. dazu Andree Thieltges / Florian Schmidt / Simon Hegelich, »The Devil's Triangle: Ethical Considerations on Developing Bot Detection Methods«, in: *The 2016 AAAI Spring Symposium Series*, Palo Alto, California 2016, S. 253–257.

31 Emilio Ferrara, *Manipulation and abuse on social media,* https://arxiv.org/pdf/1503.03752v 2 .pdf (Zugriff am 25.7.2016).

32 Yazan Boshmaf / Ildar Muslukhov / Konstantin Beznosov / Matei Ripeanu, »Design and analysis of a social botnet« in: *Computer Networks*, Nr. 57 (2013), S. 556–578.

33 Hegelich, Invasion der Meinungs-Roboter, aaO. (FN 15).

nente und Politiker zu verbreiten (beispielsweise »Sky Breaking News« oder »Global Associated News«). Sehr häufig sind in diesem Zusammenhang die sogenannten »death hoaxes«, also gefälschte Meldungen über das Ableben prominenter Musikerinnen und Musiker oder Schauspielerinnen und Schauspieler[34].

Eine Diffusion von Falschinformationen auf eine netzwerkexterne Ebene kann dann stattfinden, wenn es beispielsweise gelingt, dass Journalistinnen und Journalisten diese als »authentische« Nachricht oder Dokument ansehen und in anderen Medien verbreiten oder weiterleiten. Ein Beispiel dazu findet sich im Zusammenhang mit dem Amoklauf in San Bernadino im Dezember 2015: Auf der Plattform Twitter berichtet eine Userin/ein User unter dem Namen »Marie Christmas«, dass sie Augenzeugin der Schießerei gewesen sei. Die tweets wurde von mehreren Nachrichtensendern und Agenturen (u.a. CNN, Associated Press und New York Times) als »authentisch« eingestuft, für die Berichterstattung verwendet und veröffentlicht[35]. Ein weiterer Fall in einem ähnlichen Kontext[36] ereignete sich erst kürzlich: Hier wurde die Profilseite des mutmaßlichen Täters auf der Plattform Facebook als »fake account« enttarnt und aus dem sozialen Netzwerk entfernt, kurz bevor Nachrichtensender und Agenturen darüber berichteten[37]. Diese Beispiele untermauern die Annahme, dass manipulierte Inhalte und Informationen in sozialen Netzwerken häufig dann auf die netzwerkexterne Ebene gelangen, wenn sie mit Kontexten verknüpft sind, die tagesaktuelle Inhalte und Informationen aufgreifen oder in denen »kontroverse Themen« diskutiert werden. Ferrara et al.[38] vertritt in diesem Zusammenhang die These, dass sich das Risiko von Manipulationsversuchen, die aus sozialen Netzwerken auf die netzwerkexterne Ebene gelangen, immer dann erhöht, wenn Ausnahmesituationen herrschen: »However, manipulation of information (e.g., promotion of fake news) and misinformation spreading can cause panic and fear in the population, which can in turn become mass hysteria. The effects of such types of social media abuse have been observed during Hurricane Sandy at the end of 2012, after the Boston Marathon bombings in April 2013, and increasingly ever since[39]«.

34 Kirthana Ramisetti, *Memorable celebrity death hoaxes: Robert Redford, Carlos Santana, Paul McCartney, Jackie Chan among stars social media wrongfully declared dead,* http://www.nydailynews.com/entertainment/gossip/memorable-death-hoaxes-targeting-celebrities-article-1.2130301 (Zugriff am 19.7.2016).

35 Steve Buttry, *'Marie Christmas:' Some journalists fell for San Bernardino prank; others backed away,* https://stevebuttry.wordpress.com/2015/12/03/the-case-of-marie-christmas-verifying-eyewitnesses-isnt-simple-or-polite/ (Zugriff am 19.7.2016).

36 Die Ermordung von Polizisten in Baton Rogue am 18.7.2016.

37 Abby Ohlheiser, *How to spot a fake Facebook page during a breaking news situation,* https://www.washingtonpost.com/news/the-intersect/wp/2016/07/17/hundreds-were-fooled-by-this-fake-facebook-page-for-the-baton-rouge-gunman/ (Zugriff am 19.7.2016).

38 Emilio Ferrara, *Manipulation and abuse on social media,* https://arxiv.org/pdf/1503.03752v2.pdf (Zugriff am 25.7.2016).

39 Emilio Ferrara, *Manipulation and abuse on social media,* https://arxiv.org/pdf/1503.03752v2.pdf (Zugriff am 25.7.2016).

Eintrittswahrscheinlichkeit: Fazit und Ausblick

Trotz der hier aufgezeigten bekannten Möglichkeiten und Strategien, Informationen, Inhalte und ganze Kontexte in digitalen sozialen Netzwerken zu manipulieren, fällt es schwer eine konkrete Einschätzung darüber zu geben, wie hoch die Eintrittswahrscheinlichkeit eines solchen Falls ist. Um dabei zu einer zutreffenden Aussage zu gelangen, wäre es notwendig, möglichst viele verifizierte Manipulationen auf ihre Besonderheiten zu analysieren und zu vergleichen. Allerdings ist man in der Manipulationserkennung zumeist darauf verwiesen, diese erst »im Nachhinein« und nicht währenddessen sie tatsächlich passieren, analysieren zu können. Das ergibt sich zum einen aus der sehr großen interaktiven Dynamik, die in digitalen sozialen Netzwerken herrscht: Themen und ihre Kontexte können innerhalb kürzester Zeit massenhaft Nutzerinnen und Nutzer ansprechen, die dann das jeweilige Thema kommentieren, weiterverbreiten oder mitdiskutieren. Ob in einem Fall tatsächlich manipuliert wurde, ergibt sich aber erst aus der differenzierten Untersuchung des gesamten Kontextes und aller daran teilnehmenden User[40]. Zum anderen werden die Strategien der Manipulanten immer ausgefeilter: Die oben genannten Beispiele des automated spear phishings und der clone bots machen es dem »normalen User« schon fast unmöglich, solche Betrügereien und Manipulationen in sozialen Netzwerken sicher zu erkennen. Darüber hinaus ist es, wie oben bereits angedeutet, schwierig die genaue Zielsetzung einer Manipulation auszumachen, da über Urheber und Adressaten oder Zielgruppen nur spekuliert werden kann. Trotz dieser Einschränkungen kann man aus den oben genannten Beispielen und Manipulationsstrategien eine Annäherung an das Thema Eintrittswahrscheinlichkeit versuchen. Der vorgestellten Risikomatrix entsprechend, sollen nun die verschiedenen Möglichkeiten diskutiert werden:

1. Netzwerkintern-individuelle Risiken:

Die Absicht, einzelne User oder Usergruppen in sozialen Netzwerken zu manipulieren, ist relativ erfolgversprechend. Sei es die unerkannte »Meinungsmache«, das »Abgreifen« von persönlichen Daten, das Unterschieben von diskreditierenden Inhalten. Diese sind inzwischen »einfache Übungen« für (teil)- algorithmisierte Manipulationsformen die, wie gezeigt, durch die netzwerkinterne Struktur »unfreiwillig« unterstützt werden (Stichwort »social graph«). Zudem lassen die oben genannten spieltheoretischen Modelle der Diffusion von Falschmeldungen darauf schließen, dass es in unterschiedlichen Userkonstellationen möglich ist, gezielt Einzelpersonen oder ganze Gruppen zu manipulieren. Dementsprechend muss hier von einer hohen Eintrittswahrscheinlichkeit und in diesem Zusammenhang von multiplen Risiken für die User von digitalen sozialen Netzwerken ausgegangen werden.

2. Netzwerkintern- soziale Risiken:

Da die Usergemeinde von Facebook, Twitter und anderen digitalen sozialen Netzwerken stetig wächst, ist die Möglichkeit, dass aus manipulativer Hetze oder Meinungsma-

40 Simon Hegelich, »Decision Trees and Random Forests. Machine Learning Techniques to Classify Rare Events« in: *European Policy Analysis* 2, Nr. 1 (2016).

che in sozialen Netzwerken gesellschaftliche Probleme erwachsen, durchaus denkbar. Allerdings nicht sehr wahrscheinlich: Bisher gibt es keine Anhaltspunkte dafür, dass ein Manipulationsversuch, der gezielte Stellungen/Handlungen zu gesellschaftsrelevanten Themen fordert, dazu geführt hat massenhaft Nutzerinnen und Nutzer zu beeinflussen[41]. Die oben bereits erwähnte Analyse von Ferrara kommt bei diesem Thema zu dem Schluss, dass es gesellschaftliche Ausnahmesituationen (Terroranschläge, Naturkatastrophen etc.) braucht, um eine darauf zielende Manipulation wirksam werden zu lassen. Diese haben dann aber zumeist eine diffuse Wirkung (Panikmache und das Schüren von Angstgefühlen) und sind nur von sehr kurzer Dauer. Deshalb kann hier von einer eher geringen Eintrittswahrscheinlichkeit ausgegangen werden.

3. Netzwerkextern – individuelle Risiken:

Wie oben bereits gezeigt, ist das Risiko, dass sich Manipulationen, die bestimmte Nutzerinnen und Nutzer sozialer Netzwerke betreffen, auf der netzwerkexternen Ebene aufgegriffen werden und verbreiten, inzwischen durchaus real. Gerade Personen und (politische) Institutionen die ständig in der Öffentlichkeit stehen und digitale soziale Netzwerke zur Repräsentation, Kommunikation und Information benutzen, sind durch die oben genannten Manipulationsformen gefährdet. Hinzu tritt, dass bereits mit einfachsten Mitteln der (teil-) algorithmisierten Manipulationsformen (fake follower, chat bots etc.) Wirkungen erzielt werden können, die auf die netzwerkexternen Ebenen gelangen. Dementsprechend kann hier ebenfalls von einer hohen Eintrittswahrscheinlichkeit gesprochen werden.

4. Netzwerkextern – soziale Risiken:

Das Risiko, dass gezielte Manipulationen die netzwerkinterne Diskussion verlassen, durch andere (Massen)-Medien aufgegriffen und so zu einem gesellschaftliches Risiko werden, ist oben bereits besprochen worden. Bisher muss man hervorheben, dass dieses Szenario hinsichtlich gesellschaftlicher Manipulationen bisher noch eher unwahrscheinlich ist: An der Schnittstelle zwischen netzwerkinterner und -externer Ebene werden die Informationen und Inhalte zumeist von Journalistinnen und Journalisten auf ihre Authentizität überprüft. Allerdings wird es vor dem Hintergrund, dass die Manipulationsformen immer ausgereifter werden, auch für diese »gatekeeper« zunehmend schwierig, authentische von manipulierten Inhalten und Informationen zu unterscheiden. Zudem ist die sehr kurze Zeitspanne, in der die Nachrichtenverbreitung abläuft, inzwischen zu einem kritischen Moment geworden: Hintergrundrecherchen, die zur Verifizierung/Falsifizierung von Informationen unerlässlich sind, werden im modernen Nachrichtengeschäft immer weniger Zeit eingeräumt[42]. Daher muss hier zu-

41 Eine Aufarbeitung des sogenannten »arabischen Frühlings«, seiner (kontra-) revolutionären gesellschaftlichen Umbrüche und der Rolle, die politischer Aktivismus in digitalen sozialen Netzwerken in diesem Zusammenhang gespielt hat, würde ggf. zu einer Klärung des Zusammenhangs führen.

42 Petter Bae Brandtzaeg / Marika Lüders / Jochen Spangenberg / Linda Rath-Wiggins / Asbjørn Følstad, »Emerging Journalistic Verification Practices Concerning Social Media« in: *Journalism Practice* (2015), S. 1–20.

nächst von einer niedrigen Eintrittswahrscheinlichkeit gesprochen werden, die sich aber zukünftig erhöhen dürfte.

Der Blick auf die technischen Umsetzbarkeit und Verfügbarkeit von (teil-) algorithmisierten Manipulationsformen zeigt darüber hinaus, dass sich die Eintrittswahrscheinlichkeit von netzwerkinternen und -externen Täuschungen weiterhin erhöht. Das oben genannte Verhältnis von netzwerkinterner und netzwerkexterner Ebene dreht sich hier um: Der Ausgangspunkt von Manipulation ist immer öfter ein kommerzialisierter »Service«, der beispielsweise social bots für alle möglichen Zwecke und in allen technisch möglichen Varianten verkauft. Auf diesem Markt werden 4000 fake follower auf Twitter bereits für 5 US-Dollar angeboten und für ein wenig mehr haben diese auch ein eigenes Profilbild. Für 3700 US-Dollar kann man 1 Mio. »Freunde« im Netzwerk Instagram erwerben[43]. Dementsprechend kann davon ausgegangen werden, dass sich für jede denkbare Manipulation in sozialen Netzwerken ein auf die Anforderungen abgestimmtes »Manipulationswerkzeug« kaufen lässt, das kurzfristig und in beliebiger Quantität geliefert wird und einsatzfähig ist. Die kommerziell »freie« Verfügbarkeit von (teil-) algorithmisierten Manipulationsformen erhöht somit das individuelle und gesellschaftliche Risiko enorm, da Anwender mit manipulativer Absicht sich die technische Umsetzung nicht mehr selbstständig erarbeiten müssen, sondern auf »passgenaue Produkte« zurückgreifen können.

Zusammenfassung

Die tägliche und massenhafte Nutzung von sozialen Netzwerken als Interaktionsplattform und Informationskanal hat sich inzwischen weltweit durchgesetzt. Allerdings gibt es immer öfter Versuche, Diskussionen, Neuigkeiten und ganze Kontexte mit Hilfe von (teil-)algorithmisierten Maschinen zu manipulieren. Im Folgenden werden verschiedene Risikopotenziale erläutert und Risikoeinschätzungen vorgenommen, die sich durch diese Art der Manipulation ergeben. Dabei wird sowohl auf die netzwerkinternen und -externen Gefährdungen als auch auf Risiken für verschiedene (politische) Nutzergruppen und die gesellschaftlichen Auswirkungen von Manipulationen in sozialen Netzwerken Bezug genommen.

Summary

Today's unexceptional and copious use of online social networks (OSN) as a place to gather information or interact with each other is globally accepted. However, there are frequently attemps to manipulate these discussions and information with partial algorithmized machines. In this paper we will assess the potential risks that may arouses

43 Nick Bilton, »Friends, and Influence, for Sale Online. There are several services that allow social media users to buy bots, which can make celebrities appear more popular and even influence political agendas.« in: *New York Times* (2014), http://bits.blogs.nytimes.com/2014/04/20/friends-and-influence-for-sale-online/?_r=0 (Zugriff am 22.6.2016).

from this sort of manipulation for the political sector in OSN's and beyond. We focus on the threats and risks for different (political) user groups and the socio economic impact of these manipulations and analyse the methodes that are used to influence users and lead to viral dispersal in both, the internal network and common media.

Andree Thieltges / Simon Hegelich: Influence and Manipulation in Online Social Networks. Potential Risks and Security Ratings