WhatsApp, política mobile e desinformação: a hidra nas eleições presidenciais de 2018

WhatsApp, mobile politics and misinformation: the Hydra of Brazil's 2018 presidential election

WhatsApp, política móvil y desinformación: la hidra en las elecciones presidenciales de 2018

João Guilherme Bastos dos Santos
Instituto Nacional de Ciência e Tecnologia em Democracia Digital (INCT-DD).
E-mail: santos.jgb@gmail.com.

Miguel Freitas
Centro de Estudos de Telecomunicações da PUC-Rio. E-mail: miguel@cpti.cetuc.puc-rio.br.

Alessandra Aldé
Universidade Estadual do Rio de Janeiro (UERJ). E-mail: ale3alde@gmail.com.

Karina Santos
Instituto Nacional de Ciência e Tecnologia em Democracia Digital (INTC-DD)
E-mail: karinasantos93@hotmail.com.

Vanessa Cristine Cardozo Cunha
Instituto Nacional de Ciência e Tecnologia em Democracia Digital (INTC-DD). E-mail: vanessa_cardozo07@hotmal.com.

RESUMO
Iniciamos esta pesquisa a dez meses das eleições executivas de 2018, período no qual investigamos o comportamento coletivo de 90 grupos de WhatsApp interconectados e de apoio aos seis principais presidenciáveis, bem como os mais de 500 mil textos e imagens enviados pelos usuários, por meio dessa ferramenta, durante os cinco meses de campanha eleitoral. Com este estudo, identificamos que o alcance ampliado do aplicativo se dá através da viralização de mensagens como consequência direta da interconexão estrutural entre esses grupos. Neste cenário, confirmamos nossas hipóteses sobre a importância das características e das topologias da rede para uma compreensão aprofundada sobre a desinformação em larga escala via WhatsApp. Como resultado, reconhecemos quais métricas da rede são bem sucedidas na previsão e caminhos preferenciais para a circulação da desinformação segmentada e possibilidades de rastreio de fontes originais das notícias.
Palavras-chave: WhatsApp. Viralização. FakeNews. Eleições presidenciais 2018; *Mobile instant messaging services*.

ABSTRACT
Starting ten months before the 2018 presidential election, this paper analyses the collective behaviour of 90 interconnected WhatsApp groups supporting six different candidates, and more than 500 thousand texts and images sent during five campaign months. Large scale reach in the application relies on its counterintuitive viral spread, a direct consequence structural interconnection among groups. It confirms the importance of network topologies and characteristics in any serious understanding of WhatsApp large scale misinformation. We identified which network metrics are successful in predicting preferential pathways for segmented misinformation.
Keywords: WhatsApp. Viral spread. FakeNews, 2018 presidential election campaign. *Mobile instant messaging services*.

RESUMEN
Iniciamos esta investigación a diez meses de las elecciones ejecutivas de 2018, período en el cual investigamos el comportamiento colectivo de 90 grupos de WhatsApp interconectados y de apoyo a los seis principales presidenciables, así como los más de 500 mil textos e imágenes enviados por los usuarios, a través de esa herramienta durante los cinco meses de campaña electoral. Con este estudio, identificamos que el alcance ampliado de la aplicación se da a través de su viralización contra-intuitiva, consecuencia directa de la interconexión estructural de esos grupos. Por lo tanto, confirmamos nuestra hipótesis sobre la importancia de las características y las topologías de la red para una comprensión en profundidad sobre la desinformación a gran escala a través de WhatsApp. Como resultado, reconocemos qué métricas de la red tienen éxito en la previsión y caminos preferenciales para la circulación de la desinformación segmentada
Palabras clave: WhatsApp. Viralización. FakeNews. Elecciones presidenciales 2018. *Mobile instant messaging services*.

Submissão: 25-2-2019
Decisão editorial: 24-5-2019

1. Introdução

O WhatsApp mantém um design de rede privada com criptografia ponta-a-ponta, mas as estruturas de rede decorrentes de suas apropriações sociais no Brasil o transformaram em uma poderosa ferramenta de difusão de informações para grandes públicos. Como alternativa aos serviços de SMS pagos, a possibilidade de serviços de mensagens *mobile* sem custos de internet atraiu um contingente grande de pessoas que não têm acesso à rede de outro modo, ajudando o aplicativo a alcançar a marca de 120 milhões de usuários ativos em 2018. Se os *sites* de *fact-checking* exigem acesso à internet paga e muitas organizações jornalísticas restringem o acesso aos seus produtos, campanhas políticas souberam utilizar esta conjuntura a seu favor espalhando desinformação com viés eleitoral em momentos chave.

Um caso emblemático ocorreu às vésperas das eleições presidenciais de 2014, quando a notícia falsa de que Alberto Youssef foi envenenado durante sua prisão na Superintendência da Polícia Federal em Curitiba viralizou por WhatsApp atingindo *smartphones* do país inteiro menos de 24 horas antes do pleito. Uma montagem colocava a manchete no portal de notícias G1, acompanhada pelo rumor de que haveria envolvimento do Partido dos Trabalhadores no

crime supostamente para impedir uma delação de Youssef. O rumor foi desmentido publicamente pela Polícia Federal brasileira e pelo G1, e sua circulação foi condenada pelo Ministro da Justiça, mostrando a preocupação de diversos atores com as consequências desta mentira em um cenário eleitoral acirrado e polarizado – mas não havia como dar uma resposta proporcional na mesma velocidade ou retirar a montagem que circulava por mensagens privadas.

O impacto deste tipo de estratégia teria potencial muito maior em 2018. De acordo com *Digital News Report*[1] entre 2014 e 2018 o uso de *smartphones* para consumo de notícias cresceu de 35% para 65% enquanto a utilização de computadores passou de 64% para 62%. O uso do Facebook para notícias caiu 17 pontos entre 2016 e 2018 enquanto o WhatsApp cresceu alcançando a marca de 46%. Autoridades e comentaristas, no entanto, subestimaram consideravelmente o potencial estrago da apropriação do aplicativo repetindo e multiplicando, de modo muito mais sistemático e eficaz, a dinâmica do 'caso Youssef'. Esta possibilidade faz com que a campanha no WhatsApp – ainda cara devido a cobranças 'por envio' – comece a ser economicamente mais interessante às campanhas eleitorais.

Crescimentos súbitos na adesão a mensagens específicas são tema de investigações em torno de ações coletivas que antecedem a internet. Propostas explicativas giram em torno da ideia de cooperação condicional, em que diferentes fatores – visibilidade do comportamento, pressão por pares, percepção sobre viabilidade da ação, quantidade de pessoas

[1] Disponível em: < http://www.digitalnewsreport.org/survey/2017/brazil-2017/ > Acesso em: junho de 2018.

envolvidas, traços de personalidade – aumentariam subitamente a quantidade de adesões a partir de um limiar, massa crítica ou ponto de virada relacionados a um destes fatores. A premissa comum é, portanto, que as escolhas das pessoas muda de acordo com informações sobre quantas outras pessoas participam de determinada ação coletiva (SCHELLING, 1978; GRANOVETTER, 1978).

Apropriações de dispositivos com acesso à internet conferem nova potência a esta dinâmica. De ações em larga escala coordenadas sem a existência de organizações centrais (BIMBER et al, 2012) e o súbito crescimento em escala relativamente independente de lideranças centralizadas (MARGETTS et al, 2015), colocam em evidência pautas políticas que fogem à tutela de seus produtores iniciais – trazendo ao debate termos como *organização sem organizações* ou *liderança sem líderes*, respectivamente. Há ainda a ação conectiva (plataformas digitais como fundamento da organização política na forma de redes de compartilhamento de narrativas pessoais) (BENNETT E SEGERBERG, 2012) e a relevância de elementos narcisistas na adesão a ações com alta visibilidade em redes *online* (PAPACHARISSI, 2010). Revisões sobre internet e mudanças abruptas na escala de apoiadores em ações coletivas (MARGETTS et al, 2016) apontam a relevância da interação entre visibilidade e informação social (e seus desdobramentos em termos de constrangimento, aprovação, pressão etc.) juntamente com variações em traços de personalidade na guinada abrupta no número adesões.

A viralização em larga escala e recorrente de mensagens políticas em campanhas no WhatsApp contraria todas estas chaves explicativas. O aplicativo

não possui perfis públicos localizáveis por busca, algoritmos de impulsionamento de visibilidade, agregação automática de informação social ou entrega direcionada de conteúdo. Pelo contrário, limita o número de encaminhamentos diretos e o número de pessoas que podem pertencer a cada grupo, descartando elementos considerados peças chave na viabilidade de viralizações rápidas e recorrentes em plataformas como Facebook. A viralização de uma notícia falsa exige um aumento exponencial de visibilidade a cada encaminhamento, incompatível com índices normais de compartilhamento individual em redes de contatos privados. É neste ponto que os grupos de WhatsApp dedicados à política, em geral segmentados, com mais de duzentas e cinquenta pessoas cada e canais de comunicação entre si, entram em cena. Propomos que a utilização de métodos de análise de redes complexas pode elucidar a possibilidade de viralização neste cenário. Este artigo se diferencia, portanto, das análises de conversações que caracteriza a produção aindaincipiente sobre novas especificidades relacionadas aos aplicativos de mensagens instantâneas por celular (Mobile Instant Messaging Services) (VALERIANI E VACCARI, 2018).

Associado a criptografia, a possibilidade de viralização torna-se uma poderosa arma para estratégias criminosas. Por isso sua apropriação é previsível, mantendo a fonte relativamente segura e tornar difícil seu rastreamento, escondida do escrutínio público geral, mas em contato com o segmento alvo em particular. A segmentação é uma possibilidade mesmo que o aplicativo não ofereça estes dados: em um modelo mais sofisticado, pode ser feita por algoritmos que cruzam dados de diferentes redes *online*, superan-

do déficits em dados de qualquer uma destas redes tomada individualmente; em modelos mais simples, cruzando números de celular e CEPs presentes *online* – possibilitando segmentação geográfica por rua e inferências demográficas –, através de páginas militantes que divulgam links para grupos de WhatsApp em redes visíveis como o Facebook (possibilitando registros sobre perfil das páginas e grupos associados, além de fazer com que estes links possam ser achados por ferramentas de busca), entre outros. Ao replicar conteúdo de modo dificilmente rastreável e anonimizando a fonte, o WhatsApp da um passo adiante em relação a ferramentas como *dark posts* e *microtargeting*, utilizados para difundir informações para nichos específicos enquanto as mantém ocultas do escrutínio público no Facebook.

Como apontado por estudos sobre crescimentos abruptos em escala de campanhas políticas (MARGETTS *et al*., 2015), a identificação de traços específicos de personalidade pode indicar maior ou menor propensão a compartilhar conteúdos, possibilitando seleção de pontos com maior probabilidade de gerar o que chamamos de viralização. O caso extremo do escândalo com a Cambridge Analytica mostra que dados sobre traços de personalidade podem ser inferidos a partir de informações de comportamento *online* registradas por *sites* de redes sociais. Os chamados robôs podem atuar, portanto, identificando nichos e enviando mensagens regularmente, utilizando a propensão destes nichos a compartilhar um tipo específico de informação ao mesmo tempo em que cria uma aparência de campanha orgânica.

É importante frisar que a apropriação visando difusão de desinformação criminosa e cripto-

grafada é uma distorção específica e reprimível, e que os aplicativos de comunicação segura utilizando criptografia fim-a-fim também podem ser uma ferramenta importante na defesa do direito à privacidade. A inovação da criptografia fim-a-fim está em permitir que qualquer usuário possa criptografar suas mensagens em seu próprio celular com uma chave de criptografia segura, protegendo-o até mesmo de agências de espionagem governamentais, que dificilmente seriam capazes de decodificar o seu conteúdo. A mensagem criptografada é decodificada apenas no dispositivo do destinatário, ficando assim imune a interceptações que possam ocorrer durante o seu trânsito pela rede. Esta mesma dinâmica, no entanto, exige medidas específicas para evitar ações criminosas como a viralização sistemática de notícias falsas com finalidade de manipulação eleitoral.

Apesar de esta tecnologia estar disponível de forma pública e gratuita desde 1991 (LAUZON, 1998), diversos motivos explicam porque sua adoção ficou restrita a nichos tecnológicos ou de ativistas. O Caso Snowden, revelado pelo jornalista Glenn Greenwald em 2013, mostrou como a internet estava sendo utilizada para burlar mecanismos legais de proteção pessoal e monitorar indiscriminadamente cidadãos americanos e estrangeiros (GREENWALD, 2014; MIGUEL, 2016). Esse evento chamou a atenção para a necessidade de prover mecanismos para que os cidadãos possam se defender do abuso de Estados que não respeitem seus direitos individuais. Na sequência dos mesmos eventos, o governo brasileiro transforma em prioridade a aprovação da lei conhecida como Marco Civil da Internet, trazendo como pedra fundamental o princípio da privacidade.

Diferentes aplicativos de mensagens instantâneas ganharam popularidade com a disseminação dos "*smart phones*" nos anos recentes. Para destacar três dos mais relevantes atualmente, WhatsApp, Telegram e Signal foram lançados, respectivamente, em 2009, 2013 e 2014. Ao contrário dos dois últimos, o WhatsApp não tinha como apelo inicial a característica de fornecer criptografia fim-a-fim. Foi provavelmente uma decisão comercial, movida por pressão dos concorrentes pós-Snowden e de uma demanda de mercado, que leva o WhatsApp a finalmente adotar este recurso para todos usuários em 2016.

Ainda em 2017, uma série de entrevistas com profissionais de internet e campanha política (SANTOS E NEHRER, 2017), apontavam expectativas de profissionais familiarizados com a pauta das notícias falsas com o uso do WhatsApp em 2018. Juntamente com pesquisas anteriores sobre utilização de redes sociais *online* por apoiadores de Bolsonaro (ALDÉ E SANTOS, 2012; SANTOS E CUNHA, 2014), estas entrevistas foram utilizadas como base para o desenvolvimento de metodologias específicas capazes de analisar o uso do WhatsApp para disseminar informações falsas e definir quais dinâmicas tornam possível a viralização neste ambiente opaco.

A apropriação bem-sucedida do WhatsApp por apoiadores de Jair Bolsonaro resulta da cooperação de diversos grupos – incluindo eleitores – e um conhecimento específico voltado para viralização sistemática de conteúdo. A compreensão deste fenômeno exige uma alteração no modo como este aplicativo é analisado, superando obstáculos colocados por seu design voltado para troca de mensagens privadas,

através do cruzamento de dados, composição de redes e análises de fluxo de conteúdo.

Tendo em vista esses aspectos fundamentados em estudos prévios sobre o campo, testamos quatro hipóteses envolvendo a desinformação em larga escala via WhatsApp, durante as eleições presidenciais, são estas:

> (H1) Pessoas presentes em mais de um grupo podem viabilizar uma estrutura de grupos interconectados no WhatsApp, fazendo com que o aplicativo esteja sujeito a lógicas de rede bipartite e tornando possível que a desinformação se viralize rapidamente. O que propomos não é que esta seja uma característica a priori no aplicativo, mas um resultado da estrutura de rede contingente resultante de opções pessoais de usuários ao distribuírem-se em grupos;
> (H2) Usando métricas de algoritmos de rede, poderíamos encontrar correspondências entre a centralidade das redes e a relevância dos grupos nesse processo assimétrico, avançando, assim, em nossa análise sobre a ampliação viral da desinformação;
> (H3) Trata-se de um processo variante no tempo, que pode ser separado em estágios/encaminhamentos, no qual a desinformação vai dos nós centrais para os periféricos, ampliando exponencialmente por meio do encaminhamento de grupos.
> (H4) Acreditamos que haja indícios do difusor inicial nas mensagens viralizadas, hipótese cujo teste se beneficia da identificação dos grupos em que esta informação falsa foi circulada primeiro em decorrência da confirmação das hipóteses anteriores – possibilitando parcerias com esferas dedicadas ao aperfeiçoamento da legislação sobre campanhas políticas e a responsabilização dos possíveis produtores profissionais envolvidos.

2. H1: WhatsApp como uma rede bipartite

Com base na recomposição de redes e análise da estrutura de grupos pudemos testar a hipótese norteadora da presente análise: mais do que uma rede de pessoas conectadas através de grupos, o WhatsApp está sujeito à dinâmicas de uma rede *bipartite* de grupos interconectados por participantes em comum[2] que regulam o intercâmbio de informações, permitem o aumento exponencial de visibilidade e lógicas de difusão viral de notícias falsas mesmo dentro de uma rede fechada (H1). Essa rede de grupos mantém um fluxo de informações criptografado, opaco ao escrutínio público e intenso em períodos eleitorais. Transitando rapidamente em diferentes grupos, essas notícias podem fomentar ondas de compartilhamento que invadem novas levas de grupos a cada etapa, fluxo passível de mapeamento a partir da aplicação de métodos de constituição estrutural e análise de redes.

Metodologicamente, esta constatação tem desdobramentos relevantes. Primeiro, para além de propriedades *topológicas*, a análise de redes reconhece propriedades *dinâmicas* de viralização e contágio. As limitações de visibilidade e tamanho de grupos afastam o WhatsApp do modelo de rede que caracteriza o Facebook – *preferential attachment/scale free*, em que atores bem conectados possuem uma vantagem cumulativa, atraindo mais conexões e concentrando centralidade –, aproximando-o de modelos descentralizados. Menos centralizado, este modelo de rede apresenta maior resistência a ataques ou desativação de *nós/grupos* (ALBERT E BARABÁSI, 2000) em compa-

[2] Uma rede bipartite, ou seja, uma rede em que grupos não estão formalmente associados entre si, mas em que participantes em comum podem constituir conexões e fomentar dinâmicas de rede.

ração ao Facebook, o que significa que a retirada de grupos do ar por decisão judicial afeta pouco a dinâmica da rede. Isso não impede que, via de regra, uma pessoa esteja intencionalmente presente em um número grande de grupos.

Em ambientes políticos polarizados, esse modelo baseado em vários grupos separados pode levar a composição, intencional ou não, das chamadas *redes policêntricas segmentadas e integradas* (GERLACH, 2001). Entre as inovações desta estrutura destacam-se as funções adaptativas ou comportamento de Hidra: o alcance a diversos núcleos sociais devido a diferenças entre os perfis sociais, culturais, táticos de cada grupo da coalizão, bem como às funções assumidas; impossibilidade de repressão centralizada, uma vez que a independência e autonomia entre os grupos fazem com que a destruição de um destes não interrompa o comportamento global da rede e que para cada grupo neutralizado surja um novo com funções semelhantes; a possibilidade de reorganização e compensação, uma vez que a falha de um grupo não necessariamente implica falha dos demais, o que permite um aprendizado coletivo por tentativa e erro sem que o coletivo de grupos precise incorrer em erro. A integração dialoga com a composição de identidade por antagonismo, dedicada à utilização de *gatilhos emocionais* capazes fomentar coesão interna ao grupo, muitas vezes através do incentivo à hostilidade contra atores externos (MCDERMOTT, 2011).

A viralização nesse cenário acontece porque o aumento no estabelecimento de conexões torna mais provável que pessoas de grupos diferentes estabeleçam pontes, levando a um aumento abrupto na quantidade de pessoas indiretamente conectadas

cada vez que um novo grupo ou conjunto de grupos se conectam. Limites para a quantidade de pessoas em cada grupo potencializam a segmentação que caracteriza esta dinâmica. Esse comportamento não linear conduz a um ponto de guinada em que os vértices passam abruptamente a estar conectados ao chamado *componente gigante*, identificado em três sistemas naturais: difusão de doenças epidêmicas em redes de contato físico, redes neurais e problemas de rede de matriz genética (RAPOPORT E HORVATH, 1961).

O modelo serve tanto para entender a composição da estrutura da rede, quanto para analisar a dinâmica de circulação de informação viral nesta estrutura, embora estas duas dinâmicas obedeçam a critérios de expansão diferentes. Utilizamos esta abordagem como alternativa à ideia de viralização associada a algoritmos de visibilidade e limiares de ação política (MARGETTS *et al*., 2015) entendendo que uma lógica similar de crescimento acelerado após determinado limiar está associada a um limiar da *estrutura* da rede, e o crescimento na *quantidade* é um efeito colateral – e não a causa – do aumento abrupto no número de elementos atingidos pela viralização.

Essa questão traz a possibilidade de (H2) desenvolver métodos voltados para as redes que rapidamente identificam quais grupos estão mais propensos a estar no *início* do processo de viralização, bem como os (H3) caminhos preferenciais para desinformação *segmentada*, pontos cruciais que impedem a propagação desses conteúdos e que retardam o processo viral, podendo torná-lo inviável.

Todos os testes empíricos feitos pelo grupo de pesquisa em Tecnologias da Comunicação e Política (TCP) da UERJ confirmaram esta possibilidade (H1). Ini-

ciando suas conexões a partir da entrada em grupos de WhatsApp em apoio a candidatos e segmentos diferentes (incluindo Jair Bolsonaro, Fernando Haddad, Ciro Gomes, Marina Silva, Geraldo Alckmin e Henrique Meirelles) em maio, os sete pesquisadores do TCP envolvidos no experimento terminaram conectados ao mesmo *componente gigante*. Na rede formada por 9.812 perfis distribuídos em 90 grupos de WhatsApp especializados em campanha e discussão política, 99,11% dos perfis – independentemente de viés político – formam um *componente gigante*, comprovando a viabilidade de dinâmicas virais.

A descentralização também foi confirmada, afastando-se de dinâmicas de redes como Facebook (a distribuição de grau indica que 8.354 perfis estão em apenas um grupo, 1.107 em dois, 225 em três, 81 em quatro, 23 em cinco, 10 em seis, 9 em sete dois em 8 grupos e um em 16). A possibilidade de circulação de links, no entanto, conecta os dois modelos de rede, permitindo a usuários do Facebook a distribuição de links para grupos de WhatsApp (com entrada condicionada ao limite de membros) e perfis do WhatsApp a distribuição de links para postagens do Facebook, canalizando a participação dos membros, promovendo ondas de comentários, curtidas ou ataques súbitos sem que a atuação dos grupos que promoveram esta onda seja visível. Os links entre grupos também são utilizados para ataques, em que links de convite são jogados em grupos adversários para entradas em massa seguidas por ondas de xingamento e conteúdo impróprio ou ainda em que adversários se passam por apoiadores que precisam se tornar administradores para poder adicionar amigos ao grupo e assim que

conseguem esta função mudam o nome do grupo invertendo apoios, alteram foto e excluem membros.

3. H2 e H3: Viralização (sem algorítimo de visibilidade ou informações sociais) e o papel das métricas de rede

Reconhecer o caráter viral destas dinâmicas exige a aceitação de duas premissas, cujos desdobramentos metodológicos são relevantes para compreensão do uso político do WhatsApp. Primeiro, viralização implica *direcionalidade* (relações assimétricas entre uma fonte e seus destinatários) e um processo *variável no tempo* cuja progressão pode ser avaliada em etapas (e em que destinatários em uma etapa podem se tornar fontes na etapa seguinte). Isso coloca a questão: como identificar rapidamente que grupos têm mais chances de estar no início deste processo? Este é o ponto vulnerável da rede, uma vez que impedir a propagação a partir destes pontos desacelera e pode inviabilizar viralização.

Essa abordagem é especialmente útil em casos de redes criptografadas como o WhatsApp, cujas características podem ser válidas para compreender os modelos de rede (considerando algumas inferências sobre os *padrões* que ocorrem fora de nossa amostragem). Isso é particularmente interessante num cenário no qual a ausência de um número total de grupos que integram globalmente essa rede proíbe métodos que dependam de proporções quantitativas ou representativas. A primeira consequência importante do reconhecimento dessa assimetria é que os grupos não são equivalentes e, portanto, a simples coleta e quantificação de seus conteúdos ignora as funções estruturais da rede. Esse erro, combinado com a fal-

ta de mensuração da representatividade nessa rede privada, pode inviabilizar qualquer generalização de conclusões quantitativas.

Assim, nossos critérios para entrar em grupos segmentados, de apoio aos seis candidatos anteriormente mencionados e já ativos no início de 2018, ocorreram através do acesso a links em páginas favoráveis a esses presidenciáveis no Facebook. Por meio desses links, fomos automaticamente adicionados a outros grupos especializados durante o período pré-eleitoral, o que nos permitiu obter análises que variam conforme o tempo. Sobre esse aspecto, vale ressaltar que considerar o tempo também é relevante na análise das redes, dado que os perfis não podem visualizar informações postadas no grupo antes de suas entradas.

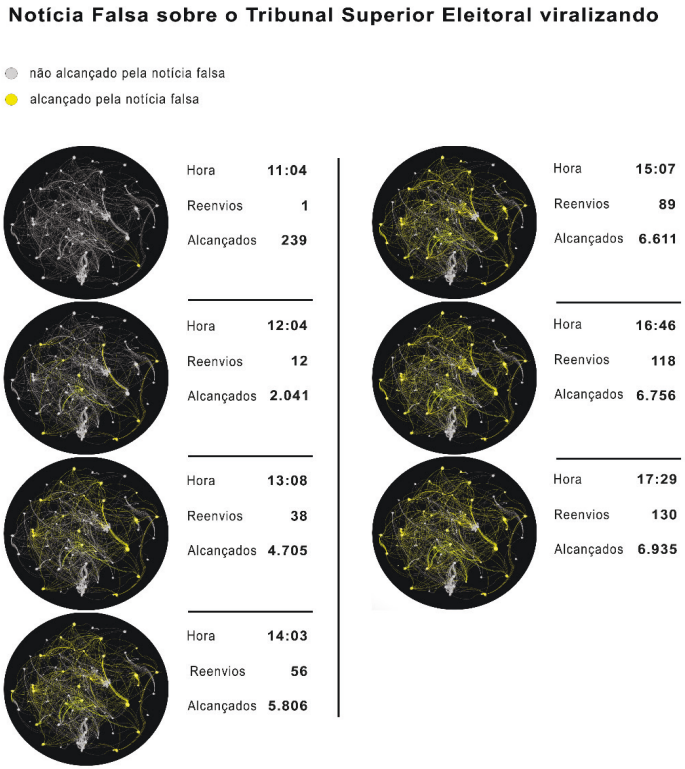
A posição estrutural de cada grupo define sua relevância neste processo: se para circular por diversos pontos da rede uma informação necessariamente passa por um grupo, este grupo tem um nível de *centralidade* nesta rede, aumentando quando outros grupos centrais passam a estar conectados graças a este grupo em particular (*eigenvector*, medida que varia entre 0 para grupos sem centralidade e 1 para grupos com centralidade máxima). Isso também faz com que as informações cheguem com mais probabilidade a este grupo, e aumente suas chances de viralização a partir do momento em que o alcança. A criptografia da fonte e impedimentos de acesso a conteúdos anteriores à entrada do perfil em um grupo deixa as possibilidades de análise deste fenômeno restritas a pesquisadores que já estivessem acompanhando quando o fenômeno ocorreu, capazes de cruzar estes dados com informações sobre estrutura da rede.

A imagem a seguir mostra a viralização de uma notícia falsa sobre o Tribunal Superior Eleitoral na rede descrita acima, atingindo 6.935 dos 9.812 perfis em poucas horas – aglomerados de pontos cinza indicam grupos sem contato com a notícia, amarelos aqueles que tiveram contato com ela e as linhas apontam conexões entre eles. A notícia afirma que o Tribunal Superior Eleitoral teria informado a anulação de 7,2 milhões de votos e que 2 milhões teriam sido necessários para que Bolsonaro vencesse no primeiro turno. A viralização por compartilhamento traz a possibilidade de analisar as etapas de difusão e encontrar a fonte inicial e o papel de cada grupo neste processo. Entre os dez primeiros a receber a notícia falsa viralizada, seis possuíam centralidade maior do que 0,90; dois acima de 0,85; e os restantes com centralidade 0,69 e 0,64. Entre os dez últimos a receber a notícia, três têm centralidade inferior a 0,18; quatro entre 0,52 e 0,42; e os três restantes possuem centralidades de 0,60, 0,73 e 0,85. Acusações contra o TSE foram uma constante entre apoiadores de Jair Bolsonaro e são sugeridas pelo próprio candidato em sua primeira declaração a jornais após o resultado do primeiro turno. Das 438,4 mil mensagens textuais coletadas nesta rede entre junho e outubro de 2018, 3.348 envolviam 'urnas' e o TSE, a maior parte delas falsa.

Ao identificar *padrões* estruturais de conexão e fluxo, modelos de rede permitem utilizar uma escala reduzida da rede para projetar e compreender a dinâmica da rede em seu todo. Esta abordagem é *impositiva* uma vez que não é possível visualizar a rede 'completa' de grupos de WhatsApp e estabelecer qualquer parâmetro de representatividade ou fidedignidade para métodos de *quantificação* simples.

Uma vez que delimitamos nossa rede, o aumento na quantidade de pessoas alcançadas desacelera nas etapas finais a despeito da continuidade de compartilhamentos. No cenário real, em que a difusão segue para outros grupos periféricos fora da rede analisada, esta desaceleração levaria muito mais tempo.

Figura 1



Fonte: autores.

Portanto, a notícia progride preferencialmente *de* grupos com maior centralidade *para* grupos pe-

riféricos – numa lógica policêntrica, quando outros grupos centrais são atingidos a dinâmica se repete, propagando a viralização. A cada etapa, a multiplicação faz com que a quantidade de informações replicadas para o próximo conjunto de grupos seja exponencialmente maior do que a anterior. Para fora da rede de grupos dedicados/especializados em política, grupos mais difundidos socialmente como de família e outras afinidades, tendem a ser atingidos. Isso faz com que a simples *quantificação* de tipos de grupos em que a notícia falsa pode ser encontrada, como os 'de família', sem levar em consideração sua *centralidade* na rede que promoveu a viralização, conduza a erros graves na atribuição de relevância, *invertendo* completamente a lógica da rede. Embora sejam mais numerosos e conjuntamente possam ter um número maior de *eleitores*, a presença de notícias falsas em grupos periféricos é a *consequ*ência e não *causa* da difusão sistemática de uma notícia falsa específica. É justamente a ignorância em relação a este processo que faz alguns apoiadores assumirem que "se eu não recebi para compartilhar este conteúdo, ninguém recebeu e sua difusão é orgânica", sem questionar quem as produziu/difundiu *antes* de seu contato com a notícia.

Vale frisar que, entre os grupos que receberam a desinformação repetidamente, também pudemos identificar padrões confirmando nossa proposta: entre os dez grupos com maior número de repetições essa informação aparece em média 8,6 vezes, a centralidade menor é 0,84 e a centralidade média é 0,92 – 6 de dez grupos de centralidades são superiores a 0,90. Número médio de membros do grupo é 234. Por outro lado, entre os grupos de dez grupos que receberam

apenas uma vez, a centralidade média é de 0,38 e os membros médios do grupo é 97,25.

Confirmando H2, o uso de algoritmos para identificar comunidades estruturadas considerando apenas a *topologia* de rede para identificar comunidades estruturadas (algoritmo de modularidade) foi bem sucedido em agrupar automaticamente os apoiadores de diferentes partes em diversas categorias, identificando subgrupos entre os defensores do mesmo candidato e também categorizando redes de discussão política com nenhum candidato específico – com erros de categorização raros envolvendo grupos de discussão sem candidato específico. Isso mostra que a heterogeneidade entre os atores envolvidos, e que essas diferenças podem ser vistas na estrutura da rede, incluindo diferenças entre aqueles que apoiam o mesmo candidato. Apesar de identificar subgrupos no apoio dos candidatos, separando-os, o algoritmo não agrupou os apoiadores de diferentes candidatos – com a exceção mencionada acima.

Nossa análise do caso envolvendo TSE também indica que houve viés de preferência partidária na circulação desta notícia. Ela foi compartilhada 202 vezes (retornando a circulação depois do último registro nos grafos), mas a despeito das interconexões entre os 90 grupos, só 41 são atingidos: 37 são grupos de apoio a Bolsonaro entre conservadores, 'de direita' ou pró-militares, e quatro são de discussão política sem candidato definido. A preferência por um candidato pode indicar um maior ou menor filtro ao compartilhamento de notícias específicas pelos perfis que conectam grupos favoráveis a este candidato a outros grupos da rede.

Reconhecidos os caminhos preferenciais e a dinâmica favorecida por grupos segmentados mais dispostos a compartilhar, entendemos que a viralização no WhatsApp envolve ao menos três etapas: primeiro a etapa de produção e difusão inicial; em seguida sua circulação em grupos segmentados dedicados a política, interconectados por membros mais dispostos a compartilhá-la e inseri-la em uma dinâmica de viralização; e por fim grupos periféricos não dedicados a política, quantitativamente mais numerosos, embora proporcionalmente irrelevantes na etapa mais intensa da viralização. Ao se aproximar de grupos com maior centralidade (*eigenvector*) tendemos a nos aproximar da fonte primária da notícia falsa.

Esse tipo de percepção ajuda a elucidar as dificuldades encontradas pelos jornalistas e comentaristas na tentativa de entender o papel do WhatsApp em *Outubro*, como uma reação à participação eleitoral: *(i)* eles não podiam entrar em grupos políticos que já alcançavam o limite dos participantes, *(ii)* quando os grupos não estavam lotados, os jornalistas dependiam de links de convite presentes fora do WhatsApp, em um cenário de ataques recíprocos que faziam com que muitos grupos políticos restringissem seus convites de links para listas de encaminhamento internas; *(iii)* quando finalmente conseguiram fazer parte de um grupo, nenhum comentário anterior a sua entrada podia ser visualizado. Esses profissionais da mídia, então, se depararam com uma série de filtros sobre qualquer conteúdo disponível, tendo apenas acesso a mensagens de novos grupos com links postados fora do WhatsApp – em geral, esses grupos não estavam preocupados com "estrangeiros" e não possuíam muitos membros.

4. H4 - A arquitetura de segurança do WhatsApp, rastreamento e o caso brasileiro em 2018

O potencial das pesquisas efetivadas no auxílio da compreensão destes casos paralelamente aos desencontros entre funcionamento da tecnologia e as iniciativas legais motivaram, em novembro de 2018, a denúncia encaminhada à Procuradoria Geral da República (PGR) com indícios que poderiam ser usados para identificar os criadores de diversos conteúdos falsos ("*fake news*") que circularam nos grupos políticos do WhatsApp no período eleitoral. Para tal, foram analisadas mensagens utilizadas no teste das hipóteses anteriores, já divulgadas em diferentes congressos acadêmicos (SANTOS et al, 2018a; SANTOS et al, 2018b), identificando as postagens que possuíam as melhores características de rastreabilidade.

Primeiro destacamos, dentre os conteúdos de mídia mais populares, aqueles que seriam comprovadamente falsos e que preservassem a mesma URL criptografada original por grandes períodos de tempo, mesmo aparecendo em grupos diferentes. A identificação bem-sucedida no âmbito acadêmico aponta a possibilidade de que órgãos competentes solicitarem ao WhatsApp, mediante ordem judicial, informações sobre o usuário e/ou endereço IP responsável pelo "upload" do arquivo/notícia falsa para os servidores da companhia.

O material resultante desse teste e enviado a autoridades competentes é composto por dois documentos: (1) parecer técnico demonstrando o funcionamento da plataforma, os limites impostos pela criptografia fim-a-fim e o tipo de informação que pode ser solicitado à empresa de forma tecnicamente plausível. (2) listagem de conteúdos falsos/caluniosos

com detalhamento de sua propagação nos grupos monitorados e seus identificadores únicos (*hashes* e URLs) que deveriam ser solicitados pela justiça.

Para produzir esta listagem de conteúdos de mídia rastreáveis, o primeiro passo consiste em agrupar as imagens e vídeos por similaridade visual, de forma automatizada. É comum, principalmente no caso de vídeos, que várias versões do mesmo conteúdo circulem na rede, diferindo apenas em termos de qualidade ou resolução de imagem.

Entre os achados, destacamos: qualquer modificação no conteúdo do arquivo, por menor que seja, produz um identificador *hash*[3] – tipo de identificador único que é gerado, utilizando uma função matemática, a partir do conteúdo do arquivo – completamente diferente. Dois *hashes* são encontrados dentro da plataforma WhatsApp, o *hash* do arquivo de mídia original e o *hash* do arquivo criptografado. Por motivos técnicos e de segurança, cada vez que o arquivo é criptografado, obtém-se um resultado diferente e, portanto, um *hash* diferente. Isto conduz a segunda variabilidade frequentemente encontrada nas mensagens de WhatsApp: um mesmo arquivo original enviado (por upload e não por encaminhamento) à rede por diferentes usuários produz versões criptografadas diferentes, com diferentes *hashes* correspondentes. A observação deste padrão de propagação permite discriminar casos em que o conteúdo original foi inicialmente distribuído por outra plataforma, por exemplo, via Facebook. Quando diferentes usuários baixam o arquivo do Facebook para então retrans-

[3] No WhatsApp os *hashes* são sempre calculados por meio do algoritmo SHA256 e codificados com BASE64.

miti-lo, de forma independente, dentro do WhatsApp é produzido um padrão diferente.

É possível perceber que em alguns casos, apesar de o *hash* da mídia (arquivo jpg) ser igual em todas as mensagens mostradas, o *hash* do arquivo criptografado (arquivo enc) apresenta versões diferentes e, para cada uma delas, o WhatsApp produziu uma URL diferente. Esta URL pode ser digitada em um *browser*, permitindo que qualquer pessoa possa fazer o *download* do arquivo criptografado caso este ainda esteja na rede. O arquivo, no entanto, somente será decodificado no conteúdo original de posse das chaves de criptografia que são encaminhadas na própria mensagem e restritas aos emissores e destinatários do arquivo.

Em casos de elevada possibilidade de que esse conteúdo tenha sido distribuído originalmente em outra plataforma, casos marcados por várias URLs não são considerados casos de boa rastreabilidade. Não seria produtivo encontrar os diferentes usuários que fizeram a cópia de uma plataforma para a outra, pois – apesar de a prática ser igualmente problemática – não há nenhum indicativo de que estes teriam qualquer relacionamento com o criador original do conteúdo.

Nesse ponto, a identificação de casos de viralização interna no WhatsApp são relevantes. Eles indicam casos em que praticamente todos os encaminhamentos se referem ao mesmo *hash* de mídia original – devido à lógica viral – e também a mesma URL (e consequentemente o mesmo *hash* criptografado). Um caso de boa rastreabilidade nesse sentido é a suposta mensagem entre Gabrielli e Fernando Haddad combinando a publicação de uma "bomba"

na Folha. Foram encontradas centenas de mensagens encaminhadas com a mesma URL, o que nos leva a concluir que o conteúdo é originário da própria plataforma WhatsApp e compartilhado quase exclusivamente por meio da opção de "encaminhamento" oferecida e ativando lógicas virais de difusão. Se o WhatsApp fornecesse às autoridades o IP responsável pelo upload desta URL poderíamos chegar ao usuário criador deste conteúdo.

Em resposta oficial, porém, o WhatsApp nega armazenar os registros de *upload* de arquivos que permitiriam a identificação do criador dos conteúdos de mídia. A resposta do WhatsApp é problemática do ponto de vista legal, pois a empresa estaria possivelmente infringindo o artigo 15 do Marco Civil que exige que provedores de aplicação guardem tais registros pelo prazo de seis meses, sendo obrigados a fornecê-los somente por determinação judicial.

O mecanismo de rastreio aqui sugerido oferece uma janela segura para investigações moderadas, isto é, sem permitir abusos e o acesso em massa pelo Estado. Isto se aplicaria também a investigações de diferentes naturezas não eleitorais como, por exemplo, crimes de pedofilia. É necessário avançar com este debate na sociedade, reestabelecendo limites e deveres das empresas de tecnologia para que estas possam colaborar efetivamente com investigações legítimas sem violar os direitos individuais.

Considerações finais

A pesquisa conseguiu confirmar empiricamente quatro hipóteses: o WhatsApp está sujeito a lógicas de uma rede bipartite graças a sua estrutura de grupos segmentados interconectados e métricas podem

identificar grupos centrais neste processo variante no tempo através de diferentes etapas. De etapa em etapa, a desinformação vai de nós centrais para nós periféricos ampliando seu alcance exponencialmente e se tornando viral. O cruzamento entre centralidade (*eigenvector*) e identificação de comunidades estruturadas (*modularidade*) oferece um mecanismo automatizado de detecção de rotas preferenciais para difusão de notícias virais em cada comunidade ou coalizão de comunidades encontrada na rede. Replicar a utilização deste algoritmo em um número maior de redes pode avançar estabelecendo *modelos de rede*.

Identificando caminhos iniciais das notícias e características específicas nas *hashes* agregadas aos encaminhamentos no aplicativo, podemos explorar quais possibilidades esta tecnologia oferece para que instituições democráticas possam ser efetivas no cumprimento de algumas demandas da sociedade, particularmente visando coibir a prática de crimes e a difusão de notícias falsas por meio de recursos técnicos disponíveis.

Referências

ALBERT, Réka; BARABÁSI, Albert-László. **Topology of evolving networks: Local events and universality**. Phys. Rev. Lett. 85. 2000, p. 5234-5237.

ALDÉ, Alessandra; SANTOS, João Guilherme Bastos dos. **Petições Públicas e batalhas digitais**. XXI COMPÓS, Juiz de Fora (MG), 2012.

BENNETT, W. Lance, SEGERBERG, Alexandra. *The Logic of Connective Action*. Information, **Communication & Society**, vol. 15, no 5, p. 739-768, 2012.

BIMBER, Bruce, FLANAGIN, Andrew J, STOHL, Cynthia. **Collective Action in Organizations: Interaction and Engagement in an Era of**

Technological Change. Cambridge University Press, 2012.

FREITAS, Miguel. "*Twister: the development of a peer-to-peer micro-blogging platform.*" **International Journal of Parallel, Emergent and Distributed Systems** 31.1 (2016): 20-33.

GERLACH, Luther. The structure of social movements: environmental activism and its opponents In: ARQUILLA, John, RONFELDT, David. **Networks and netwars**: The future of terror, crime, and militancy. RAND, 2001.

GRANOVETTER, Mark. Threshold Models and Collective Behavior. American**. Journal of Sociology**, 83, pp. 1420-1443, 1978.

GREENWALD, Glenn. "No place to hide: Edward Snowden, the NSA, and the US surveillance state". Macmillan, 2014.

LAUZON, Elizabeth. "The Philip Zimmerman Investigation: The Start of the Fall of Export Restrictions on Encryption Software Under First Amendment Free Speech Issues." **Syracuse L. Rev**. 48 (1998): 1307.

MARGETTS, Helen; JOHN, Peter; HALE, Scott A.; Yasse Ri, Taha. **Political Turbulence: How Social Media Shape Collective Action**. Princeton e Oxford: Princeton University Press, 2016.

MCDERMOTT, Rose. Emotional Manipulation of Political Identity. In: Le Cheminant, Wayne; Parrish, John M (org.). **Manipulating Democracy.** Routledge, 2011.

PAPACHARISSI. Zizi A. **A Private Sphere**: Democracy in a Digital Age. Cambridge: Polity Press, 2010.

RAPOPORT A, e HORVATH, W. A study of a large sociogram. **Behavioral Science** 6. 1961, p. 279-291.

SANTOS, João Guilherme Bastos dos.; SANTOS, Karina; CARDOZO, Vanessa . **Cartografia do Whatsapp**: a rede de apoio aos presidenciáveis nas eleições de 2018. In: 3'Congresso Nacional de Estudos Comunicacionais da PUC, 2018. Minas - Poços de Caldas: Conec, 2018a.

SANTOS, João Guilherme Bastos dos; SANTOS, Karina; CARDOZO, Vanessa. La red del 'mito' 2018: Articulaciones políticas de grupos de extrema derecha en Whatsapp. Conferência Latinoamericana de Ciências Sociais, 2018. Buenos Aires, Argentina: CLASCO, 2018b.

SCHELLING, Thomas. **Micromotives and Macrobehavior**. New York: Norton, 1978.

VALERIANI, Augusto, VACCARI, Cristian. Political talk on mobile instant messaging services: a comparative analysis of Germany, Italy, and the UK. Information, **Communication and Society,** vol. 21, no 11, pp. 1715-1731, 2018.

João Guilherme Bastos dos Santos
Doutor em Comunicação Social pela Universidade do Estado do Rio de Janeiro (UERJ), pesquisador vinculado ao Instituto Nacional de Ciência e Tecnologia em Democracia Digital (INCT-DD) alocado no grupo Tecnologias da Comunicação Política. E-mail: santos.jgb@gmail.com.

Miguel Freitas
Recebeu seu B. E., M. Sc. e doutorado em engenharia elétrica pela Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), em 2002, 2004 e 2011, respectivamente. Atua como engenheiro de pesquisa no Centro de Estudos de Telecomunicações da PUC-Rio. E-mail: miguel@cpti.cetuc.puc-rio.br.

Alessandra Aldé
Doutora em Ciência Política pelo Instituto Universitário de Pesquisa do Rio de Janeiro, professora de Comunicação Política da Universidade Estadual do Rio de Janeiro (UERJ), coordenadora do grupo de pesquisa Tecnologias da Comunicação Política. E-mail: ale3alde@gmail.com.

Karina Santos
Mestranda em Comunicação Social pela Universidade Federal Fluminense (UFF), pesquisadora do Instituto de Tecnologia e Sociedade do Rio de Janeiro (ITS-RIO), membro do Instituto Nacional de Ciência e Tecnologia em Democracia Digital (INTC-DD) alocado no grupo Tecnologia da Comunicação Política. E-mail: karinasantos93@hotmail.com.

Vanessa Cristine Cardozo Cunha
Mestre em Comunicação Social pela Universidade Estadual do Rio de Janeiro (UERJ), doutoranda do Programa de Pós Graduação em Comunicação Social da UERJ, membro do Instituto Nacional de Ciência e Tecnologia em Democracia Digital (INTC-DD) alocado no grupo Tecnologia da Comunicação Política. E-mail: vanessa_cardozo07@hotmal.com.