# Leaders or Followers? A Temporal Analysis of Tweets from IRA Trolls

**Siva K. Balasubramanian[1], Mustafa Bilgic[2], Aron Culotta[3], Libby Hemphill[4], Anita Nikolich[5], Matthew A. Shapiro[6]**

[1]Stuart School of Business, Illinois Institute of Technology, Chicago, IL
[2]Department of Computer Science, Illinois Institute of Technology, Chicago, IL
[3]Department of Computer Science, Tulane University, New Orleans, LA
[4]School of Information, University of Michigan, Ann Arbor, MI
[5]School of Information Sciences, University of Illinois, Urbana-Champaign, IL
[6]Department of Social Sciences, Illinois Institute of Technology, Chicago, IL

**Abstract**

The Internet Research Agency (IRA) influences online political conversations in the United States, exacerbating existing partisan divides and sowing discord. In this paper we investigate the IRA's communication strategies by analyzing trending terms on Twitter to identify cases in which the IRA leads or follows other users. Our analysis focuses on over 38M tweets posted between 2016 and 2017 from IRA users (n=3,613), journalists (n=976), members of Congress (n=526), and politically engaged users from the general public (n=71,128). We find that the IRA tends to lead on topics related to the 2016 election, race, and entertainment, suggesting that these are areas both of strategic importance as well having the highest potential impact. Furthermore, we identify topics where the IRA has been relatively ineffective, such as tweets on military, political scandals, and violent attacks. Despite many tweets on these topics, the IRA rarely leads the conversation and thus has little opportunity to influence it. We offer our proposed methodology as a way to track the strategic choices of future influence operations in real-time.

## 1 Introduction

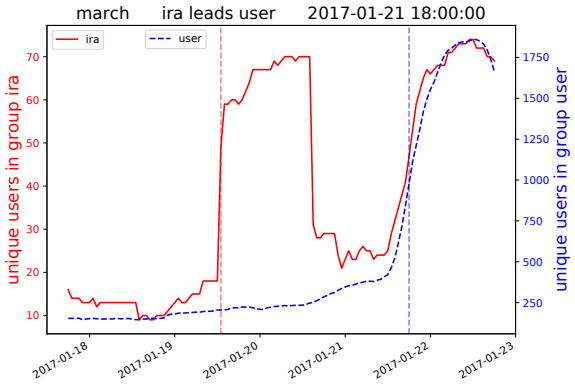
The efforts of the Internet Research Agency (IRA) to influence online political discourse in the United States in the 2016 presidential election and beyond are by now well-documented (Aral and Eckles 2019; McKay and Tenove 2020; Lukito et al. 2020). Especially since Twitter released lists of IRA-associated accounts and tweets, numerous studies have characterized the content of IRA tweets and their retweet networks, even identifying instances where mainstream news sources refer to IRA accounts directly (Lukito et al. 2020). While compelling, these direct measures of influence are rare, limited in scope, and do not address the potential widespread influence of the IRA's campaigns over political discourse through more subtle means, such as by exacerbating existing partisan divides and sowing discord. While these indirect paths of influence are inherently more difficult to quantify, identifying them could help us better understand the strategies and breadth of such campaigns.

To investigate these issues, in this paper we focus on the *temporal precedence* of salient words on Twitter to distinguish between instances where the IRA is a *leader* or *follower* in a trending conversation. For example, Figure 1 shows two instances, one where the IRA leads other Twitter users in discussing the Women's March of January 2017, and one where the IRA follows other users in discussing health care in October of 2016. We argue that patterns such as these can provide insight into both the priorities and potential impact of influence campaigns. IRA leadership indicates a greater level of effort to be "ahead of the curve" and suggests a greater potential to frame and influence the conversation than messages posted after a trend has already been established.

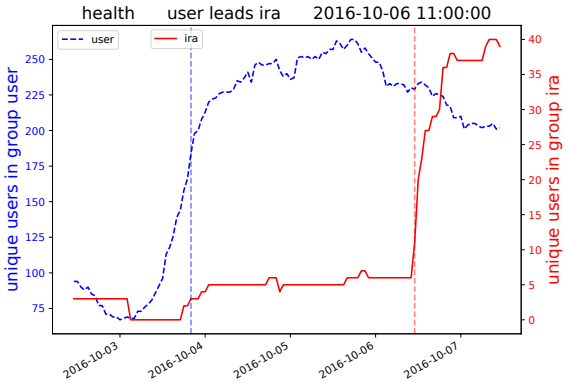
The primary contributions of this paper are (1) to establish a methodology for identifying the *temporal precedence* of salient word trends in online media, and (2) to apply it to understand the strategic choices and potential impact of IRA campaigns. We analyze over 38M tweets from the IRA, members of Congress (MOCs), journalists, and "ordinary" users posted between January 2016 and December 2017 and investigate the following research questions:

- **RQ1: For which topics is the IRA more likely to lead than follow the trend?** We conduct topic clustering analysis to group terms into topics and compute statistics over leading and following frequencies. We find that the IRA is most likely to lead on topics related to the election, race, and entertainment; it is more likely to follow on topics related to the military, health policy, and violent attacks.
- **RQ2: How does this temporal precedence vary by user group?** We conduct additional analyses for Twitter users who are MOCs and journalists to assess how the leading and following relationships vary. We find, for example, that journalists tend to lead on topics related to scandals and technology, while MOCs tend to lead on topics related to gun policy and the military.
- **RQ3: What is the potential return on investment to the IRA from their efforts to influence conversations on each topic?** By comparing the total number of tweets the IRA posts on a topic with the number of users who tweet after the IRA on that same topic, we can identify the IRA's differential potential impact across topics. For example, even though the IRA tweets with greater frequency about violent attacks and military relative to other

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.




(a) IRA leading Users on the term "march," referring to the Women's March of January 2017.




(b) Users leading the IRA on the term "health," referring to the healthcare debate of October 2016.

Figure 1: Two example word spikes, one where the IRA leads and one where it follows the User group.

topics, they rarely lead the conversations, and so have little opportunity to influence them.

In the remainder of this article, we first review some of the extensive background on influence operations in general and the IRA specifically (§2); we then describe our data collection and analysis methods (§3). Next, we describe the core results (§4) and discuss their implications for protecting online discourse in the future (§5), and end with a concluding summary (§6). Additional sensitivity analyses and sample data are included in the Appendix[1].

## 2 Russian Influence Operations and the IRA

In light of the role the IRA played during the 2016 U.S. presidential election, we contextualize our analysis of the IRA's strategies and potential influence with a brief overview of Russia's Influence Operations – sometimes called Information Operations (IO), other times called Disinformation (Jowett and O'Donnell 2018). Russia utilized these techniques during the Soviet era for both adversaries and its own citizens, ingraining messages of patriotism and loyalty (Bittman 1985),[2] and it has honed its IOs since then to the point where they are a crucial tool of statecraft.[3] Information Warfare (IW), a subset of IO used during the Cold War, further incorporates electronic means of active manipulation (Harknett 1996), protecting Russian ideology from Western influence and weaponizing information to influence opinion and foster unique narratives.

The IRA was formed shortly after the 2012 election of Vladimir Putin, which spurred a slew of domestic Internet censorship measures. Specifically, the IRA focused on bolstering positive sentiment among the populace by hiring workers to write positive content on Russian blogs and sites. While the IRA was officially registered in 2013 to Russian billionaire Yevgeny Prigozhin, who was indicted by the US in 2018 for interfering in the 2016 elections,[4] all evidence indicates that the IRA operates under the direction and authority of the Kremlin. Yet, little was known about the structure of IRA's operations, tactics, and political goals until Russian undercover reporter Alexandra Garmazhapova published an exposé in 2013.[5] For the American audience, the first of two *The New York Times* exposés were published in 2015, drawing attention to the IRA's existence and "troll factory" tactics and highlighting the IRA's geopolitical goals of sowing discord in countries targeted by the Kremlin as enemies of Russian ideology.[6]

IRA trolls engage in tasks that are specific but, when coordinated, function much like an industrial effort (Linvill and Warren 2020), playing ideologies off of each other and working both sides of an issue (Golovchenko et al. 2020; Linvill and Warren 2020; Zhang et al. 2021). Their efforts to organize protests on opposite sides of an issue have had

[1] Online at: https://arxiv.org/abs/2204.01790

[2] This is not to say that foreign governments besides Russia do not also engage in these behaviors or that countries besides the U.S. are not attacked. See Martin, Shapiro, and Nedashkovskaya (2019) for a broad assessment of foreign influence efforts from 2013 to 2018.

[3] Rooted in psychology, communications, public relations and operations research, Russia's weaponized IO originated under Joseph Stalin in 1923, when the KGB's precursor, the GPU, created a special disinformation office to conduct active intelligence operations. These Soviet operations were under more scrutiny during the Cold War under President Reagan, and were monitored by the U.S. State Department, which put out annual reports such as "Soviet Active Measures: Forgery, Disinformation, Political Operations" until the end of the 1980's (Manning et al. 2004).

[4] Source: https://www.nytimes.com/2018/02/16/world/europe/prigozhin-russia-indictment-mueller.html

[5] Source: https://novayagazeta.ru/articles/2013/09/09/56265-gde-zhivut-trolli-kak-rabotayut-internet-provokatory-v-sankt-peterburge-i-kto-imi-zapravlyaet

[6] Sources: https://www.nytimes.com/2015/06/07/magazine/the-agency.html and https://www.nytimes.com/2018/02/18/world/europe/russia-troll-factory.html

significant consequences, such as motivating African Americans to boycott elections, increasing distrust of political institutions among Latinos, prompting right-wing voters to be confrontational (Im et al. 2020), and spreading fake news (Howard et al. 2019). Linguistic innovations are often employed by the IRA: e.g., in response to violence involving immigrants, one Russian account tweeted, "Between the #rapefugees and the #refujihadis I think we've all had quite enough of this 'refugee' farce." The introduction of new linguistic terms like these (i.e. #rapefugees and #refujihadis) allows the IRA to frame the debate over immigration as one of national security and violence in order to influence citizens' reactions and political views.

The IRA also amplifies conversations and messages about particular policy issue areas that may be subsequently read by others, namely MOCs, journalists, and the public. Yet, the IRA approaches "disinformation in ongoing topics differently based on the political affiliation of their target audience: US conservative audiences are... targeted... about general topics, [while] African American audiences are... targeted with tweets about...Black Lives Matter[, the purpose of which is] to manipulate and radicalize, with some gaining meaningful influence in online communities after months of behavior designed to blend their activities with those of authentic and highly engaged US users" (Howard et al. 2019, 27).

The Senate Select Committee on Intelligence analyzed 10.4M tweets provided by social media platforms in 2017 and found that the IRA's behavior on Twitter was arguably "organic." That is, topics are chosen in a reactive manner along the lines of a "digital marketing agency," focusing on current events rather than the more careful cultivation of themes on Facebook or Instagram (DiResta et al. 2018). We invoke these claims of "organic" Twitter use given the fluid and exploratory nature of the IRA's activities. For example, trolls employ multiple personas to assess which ones have the greatest impact.[7] As well, tests for the efficacy of IRA's statements across different social media platforms suggest that IRA activities on Twitter are often preceded by related Reddit-based activity one week prior (Lukito 2020), implying that the IRA was testing its strategies on one platform before using them in a more widespread fashion on other platforms. Golovchenko et al. (2020) classify this sort of behavior as a "pre-propaganda strategy."

The release of historical IRA Twitter datasets in the past few years has led to some new insights into how the IRA's messaging propagated in social media. For example, Zannettou et al. (2020) study how images flow from Twitter to Reddit and other platforms; Stewart, Arif, and Starbird (2018) examine IRA retweet networks around the #BlackLivesMatter movement; and Badawy, Ferrara, and Lerman (2018) study how retweeting IRA accounts varies by political stance.

Despite the IRA's continued efforts to affect Americans' exposure to specific narratives (Linvill et al. 2019), including recent misinformation posted by the IRA about the efficacy of COVID-19 vaccines (Walter, Ophir, and Jamieson 2020), we acknowledge that the IRA may effect little change for Twitter users with extreme political beliefs and attitudes (Bail et al. 2020; Lazer 2020). Yet, the combination of strategies employed by the IRA can serve to bolster messages of dubious accuracy. With sufficient exposure to these types of messages, people begin to treat such information as being more reliable (Lewandowsky et al. 2012; Berinsky 2017), particularly if the information or rumor is discussed at length within a particular social cluster (DiFonzo et al. 2013).

Given that the effects of general misinformation (i.e., not solely IRA-related misinformation) linger long after one's exposure to it and even after one has been exposed to corrections (Wittenberg and Berinsky 2020; Nyhan and Reifler 2015, 2010), attempts to understand how the IRA may have influenced the American public must be rooted in an attempt like ours: to understand how the IRA initially chooses how and what to communicate.[8]

# 3 Methodology

In this section, we describe our computational approach to data collection, term extraction, time series processing, and topic discovery.[9]

## 3.1 Data

- **Russian troll accounts (IRA):** We downloaded the October 2018 Twitter release of ∼2.9M tweets from 1,635 accounts found to be affiliated with Russia's Internet Research Agency.
- **Members of Congress accounts (MOCs):** ∼2.5M tweets from 526 members of the 115th U.S. Congress (2017-2019) (Hemphill and Schöpke-Gonzalez 2020).
- **Journalist accounts:** ∼900K tweets collected from 976 journalist accounts. These accounts were identified in a semi-automated way. First, we used a query[10] to identify 20 Twitter Lists that contain the term "journalist," and then retained accounts that appear on at least two lists to reduce noise. Additionally, we manually searched for the phrases "liberal journalist" and "conservative journalist" on Twitter's search page and identified 24 additional accounts. The final list contains a mix of very popular TV personalities (e.g., @AndersonCooper, @WolfBlitzer) as well as many journalists from smaller media outlets across the political spectrum.
- **Users:** ∼31M tweets from 71,128 "regular" users. To identify a sample of politically interested, ordinary Twitter users, we sampled users who follow at least one MOC or journalist collected above. To identify these users, we first collected up to 50K followers of all MOC and journalist accounts (15M unique accounts), and then we sampled 100 followers of each account. We collected up to

[7] See Xia et al. (2019) for a case study of a single IRA persona.

[8] The steps to achieve this understanding are outlined by Aral and Eckles (2019) as follows: identify impressions of manipulative content, match one's impressions to one's voting patterns, establish causality between the two, and then identify the impacts on election outcomes.

[9] Replication materials are at: https://github.com/tapilab/icwsm-2022-leader

[10] The Google query was: "journalist site:twitter.com inurl:lists"

5,000 historical tweets from each user, retaining those users who tweeted between 2016-01-01 and 2017-12-31.

In total, the dataset contains ∼37M tweets from 74,265 users, posted between 2016-01-01 and 2017-12-31.

### 3.2 Text Processing

To identify candidate words of interest, we first processed all tokens from IRA accounts by converting them to lowercase; removing punctuation; retaining hashtags, mentions, and emojis; removing URLs; and normalizing numbers. We removed words that occurred in fewer than 50 IRA tweets or more than 40% of IRA tweets. This resulted in a vocabulary of 47,408 unique terms appearing in ∼416M tokens from users from all four groups. To further focus on words that are shared among these four user groups while exhibiting short bursts of popularity, we retained words that are used a minimum of 200 times by at least one of the journalist, MOC, or regular user group while appearing in no more than 1% of any user group's tweets. This reduced the vocabulary to 8,535 words. We retained tweets that contained at least one of these words, resulting in ∼2.7M IRA tweets, ∼2.5M MOC tweets, 880k journalist tweets, and 29.6M user tweets.

### 3.3 Time Series Processing

We define a **word spike** to be a sharp increase in the usage of a word by a group in a certain time interval, and our overall approach is to identify instances where a word spike for one group immediately precedes a spike in the same word for another group. While such temporal precedence is insufficient to infer a causal relationship between spikes, it provides suggestive evidence regarding the types of words for which IRA is likely to be a *leader* or a *follower*.

To operationalize this concept, we first construct a time series for each word/user group combination. Let the value $n_{wgt}$ represent the number of unique users from user group $g$ that use the word $w$ in the prior 24 hours from time $t$. We use the number of unique users, rather than tweets, to limit the impacts of a single user tweeting the same term many times. Thus, this value is a measure of the group adoption of a term in the given time period. These values are computed for each hour from 2016-01-01 to 2017-12-31.

We next identify candidate word spike events as follows. For each word time series, we compute the difference vector $\Delta_{wgt} = n_{wgt} - n_{wg(t-1)}$ and retain the top three values, indicating the biggest three "spikes" in the usage of term $w$ by group $g$. To reduce noise, we omit any spikes where $n_{wgt} < 5$. This resulted in 35,826 total word spike events from the four user groups.[11]

To identify potential leader-follower relationships between user groups, we select pairs of word spike events $\{\langle w, g', t' \rangle, \langle w, g, t \rangle\}$ where a sudden increase in the use of word $w$ by group $g$ is immediately preceded by a sudden increase in the use of the same word by another group $g'$. We restrict these pairs to those in which the spike for group $g'$ occurs no more than 4 days prior to the spike for group $g$. This resulted in 10,599 word spike pairs involving 3,415 unique words.

For each pair of word spike events, we categorize the first group as the *leader* and the second group as the *follower*. As seen previously, Figure 1 shows two word spike pair events, one in which the IRA leads, and one in which it follows. While the data do not allow us to make causal claims about such events (e.g., we cannot conclusively determine that the leader *caused* the follower to use this term), it is suggestive of a greater level of interest and effort on the part of the leading group with respect to this topic. Furthermore, being a leader on a topic increases the potential for having an influence on that topic. Thus, this methodology allows us to focus on topics for which the IRA has a potential impact, as opposed to examining all topics the IRA discusses while ignoring the evidence of temporal precedence.

### 3.4 Term Clustering

To better understand the topics for which the IRA tends to lead other groups, we used a semi-automated approach to cluster terms into meaningful topics. For each of the 3,415 unique words identified in the previous step, we collected all of the corresponding word spike pairs. We then collected the tweets containing each word posted by each user group involved in the leader-follower relationships, restricted to four days prior and one day after the word spike for that group. We then represent each word by a feature vector indicating the count of all other words mentioned in the same tweet as the target word (the context vectors of each word). These context vectors are converted into term frequency-inverse document frequency representations, normalized to unit length. We then cluster each context vector into one of 500 topics using K-Means clustering.[12] For example, one cluster contains the words "#womensmarch", "parade," "inauguration," in reference to the Women's March that occurred the day after President Trump's inauguration in January 2017.

We next manually coded each cluster into topics using an open-ontology approach. Four co-authors independently reviewed each cluster, inspecting the words as well as the contexts in which they appeared, and assigned a label to each cluster. The labels were not pre-specified, but rather were chosen by the annotators separately. These labels were then discussed jointly and merged into a unified schema, resulting in the following 22 topic labels: climate change, disasters, economy, election, entertainment, foreign policy, gun policy, health policy, holidays, immigration, in-memory, military, other, other policy, politics, protests, race, scandals, stopwords, technology, violent attacks, women/LGBTQ. As the "other" and "stopwords" clusters did not contain any semantically meaningful content, we removed them from further analysis, leaving a total of 20 clusters. Table 1 in the Appendix lists the 20 clusters, the number of unique word spike terms in each, and example terms.

[11] The number of spikes per group is user: 19,940, IRA: 10,066, MOC: 4,282, journalist: 998.

[12] The number of clusters was not optimized – we chose a large number, knowing that human annotators would reduce to a more manageable size in the next step.

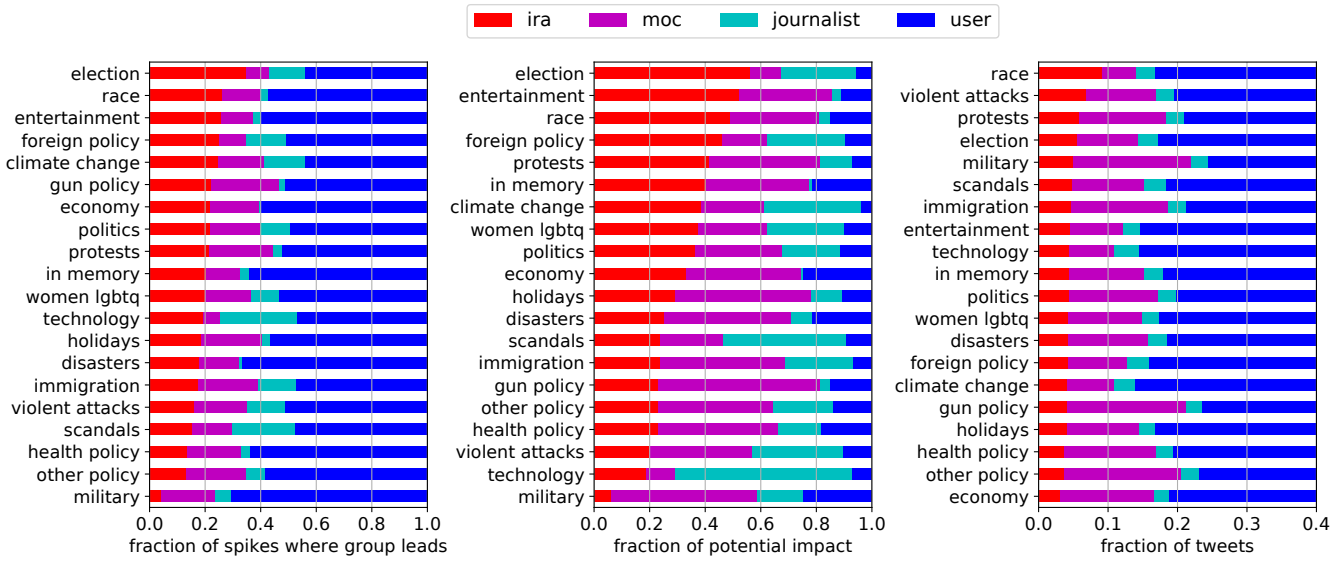



Figure 2: Left panel: Fraction of word spikes where each group leads; center panel: fraction of total potential impact for each group per cluster; right panel: fraction of all tweets from each group containing at least one term from a cluster

## 4 Results

To summarize our research questions from §1, we are interested in identifying which topics the IRA leads versus follows (**RQ1**), how these topics vary by user group (**RQ2**), and the trade-off between the effort allocated to each topic and its potential for influencing other conversations (**RQ3**).

To address **RQ1**, the left panel of Figure 2 shows the main results by topic, indicating the fraction of word spikes in each topic for which each group leads. For example, we find that of all the term spikes related to the election, IRA leads roughly 35% of the time. In contrast, the IRA leads less than 5% of the time for the military topic. Based on this ordering, the IRA appears to lead most often on topics of election, race, and entertainment. The relatively high ranking of term spikes related to race is consistent with research showing that IRA members are effective when trolling as Black activists (Freelon et al. 2020). Yet, given the impact of the IRA on the 2016 U.S. presidential election (Aral and Eckles 2019; McKay and Tenove 2020; Lukito et al. 2020), we would have been surprised if the election topic had not been ranked at or at least near the top.

This initial analysis, however, ignores the overall volume of each discussion. It could be the case that the IRA leads on many terms in a topic but does not lead on the terms that are involved in the high volume conversations. To address this distinction, we introduce an additional measure called *potential impact*. For a spike pair $\{\langle w, g', t' \rangle, \langle w, g, t \rangle\}$, where group $g'$ leads group $g$, the potential impact of group $g'$ is the total number of users in group $g$ who use term $w$ in the four days following $t'$. In other words, it is the total number of users who have the potential to have been influenced by group $g'$, based on temporal precedence. For example, in the "march" example from Figure 1(a), there are 1,860 users who use the term "march" in the four days following the IRA spike on January 19th. We aggregate these values across all spike pairs and group by topic.

The center panel of Figure 2 shows the fraction of potential impact accounted for by each user group per topic. While the ranking of topics is similar to the left panel, there are re-orderings that reflect distinctions between the topics. For example, while the "protests" topic is only at rank 9 in the left panel, it rises to rank 5 in the center panel, indicating that, while the IRA does not lead on many conversations about protests, when it does, it leads popular conversations. The converse is true for gun policy, suggesting that, while the IRA often leads such conversations, those conversations are less popular conversations.

Finally, in the right panel of Figure 2, we report the fraction of tweets from each group that contain at least one word from each topic. Note that the $x$-axis in this figure is truncated since the User group accounts for at least 70% of all tweets on all topics due to their much greater size. This panel begins to address **RQ3** – how does the number of tweets the IRA allocates to a topic relate to the potential impact? For example, while the IRA allocates a significant number of tweets to the "violent attacks" topic, both the fraction of spikes where the IRA leads on this topic, as well as the fraction of potential impact, are near the bottom of the list.

This relationship between "effort" as measured by the number of tweets on a topic and potential impact and propensity to lead is shown more clearly in Figure 3. On the $x$-axis is the raw number of tweets from the IRA on each topic ($\log$ scale); on the $y$-axis are the fraction of spikes where the IRA leads and the fraction of potential impact. As implied above, we find the "violent attacks" topic in the lower right quadrant of both panels, showing the limited returns on investment in this topic. In contrast, topics like "entertainment" and "protests" have high potential impact de-

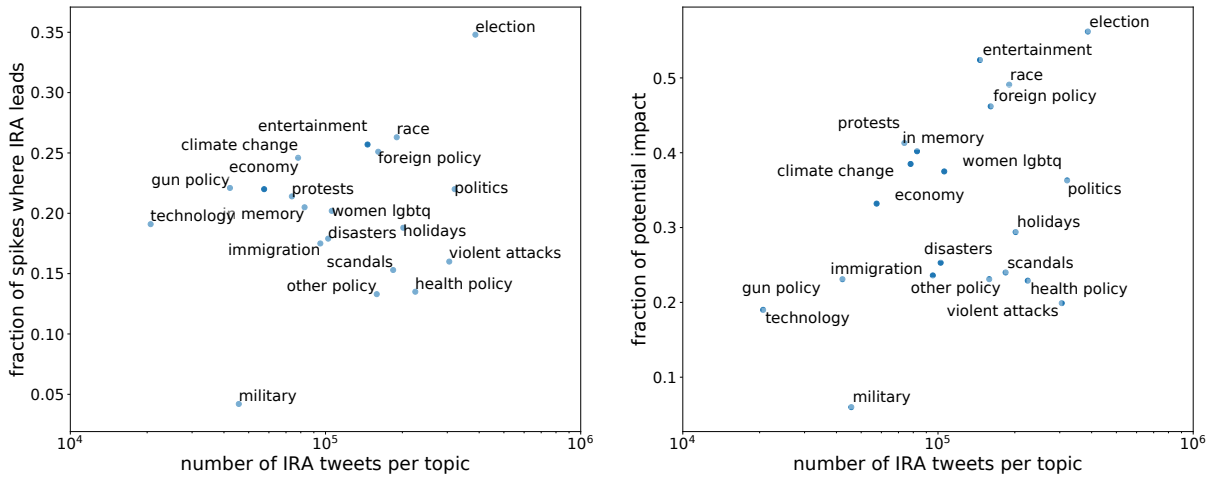



Figure 3: Scatter plots showing the relationship between the number of IRA tweets on a topic and the fraction of spikes where the IRA leads (left panel) and the fraction of potential impact (right panel). Topics in the top left quadrant suggest higher potential return on investment, while topics in the lower right quadrant suggest lower potential returns.

spite the relatively smaller number of tweets on the topic.

The fraction of word spikes for most IRA-led content ranges from 15% to 25%. The connection between content and efficiency shown in Figure 3 illustrates how the IRA prioritized the election topic and how, with differentiated efficiency, the IRA utilizes its resources to tweet about other non-election-related topics while focusing on the election. In light of the IRA's limited resources, this multifaceted approach may reflect not just narrative switching, identified in Dawson and Innes (2019), but also how the IRA engages in an "organic" and reactive process, consistent with DiResta et al. (2018).

Returning to **RQ2**, we next look more closely at how the temporal precedence of word spikes varies by user group. To do so, for the MOC, journalist, and user groups, we identify every spike pair involving the IRA group. We then plot by topic the fraction of spikes where the IRA leads or follows the other group. Figure 4 shows these results for each of the three groups. For example, of the 22 immigration spike pairs involving the IRA and MOC groups, IRA led on 16 of them (73%), indicated by the immigration row in the first panel. We also observe for the IRA-MOC pair that the IRA leads on most topics, which is consistent with findings related to the 113th Congress showing that MOCs were not likely to lead online discussions about particular issues, tending rather to follow their supporters (Barbera et al. 2019). That said, MOCs do lead the IRA on the following topics: military, protests, and violent attacks.

In terms of the IRA-journalist pair, the center panel of Figure 4 shows that spikes are predominantly IRA-based with regard to the topics of climate change, health policy, the election, and entertainment. However, the IRA has apparently little interest or is unable to lead the discussion with journalists on the following topics: technology, gun policy, economy, military, and in-memory. The fact that IRA content leads media content across such a large number of topics is potentially problematic given that the media may directly quote IRA tweets (Zhang et al. 2021; Lukito et al. 2020). Any potential influence by the IRA on journalists would likely reinforce polarization given that information from left and right-leaning media sources is typically consumed by, respectively, people on the left and right (Tyler, Grimmer, and Iyengar 2020).

Turning now to the IRA-user pair, for no single topic does the IRA lead more than users do. This may reflect the fact that the number of tweets posted by the user group is much greater than those posted by the IRA group. That said, the IRA does lead users on the election topic more than 40% of the time. The implication of the IRA's temporal precedence on the election topic in particular may have serious effects for conservatives, as they are exposed significantly more to IRA-based information relative to liberals (Hjorth and Adler-Nissen 2019), complementing research about 2016 showing that "top influencers" who shared news on the left affected Clinton supporters, while Trump supporters affected the behavior of "top fake news spreaders" (Bovet and Makse 2019).

In addition to the primary empirical results in this section, please refer to the Appendix for a number of sensitivity analyses to assess the robustness of the results to changes in parameter choices, as well as to view additional sample time series to provide further insights into the nature of the leading/following relationships.

## 5 Discussion

In terms of election-related content, the data suggest that IRA is clearly focused on being "ahead of the curve" – there is evidence that the IRA's volume of election-related conversations often lead MOCs, journalists, and the general public. The IRA's ability to lead conversations about elections in social media suggests, but does not conclusively show, their potential to influence them. Future work is needed to more precisely estimate possible causal effects of IRA's efforts on

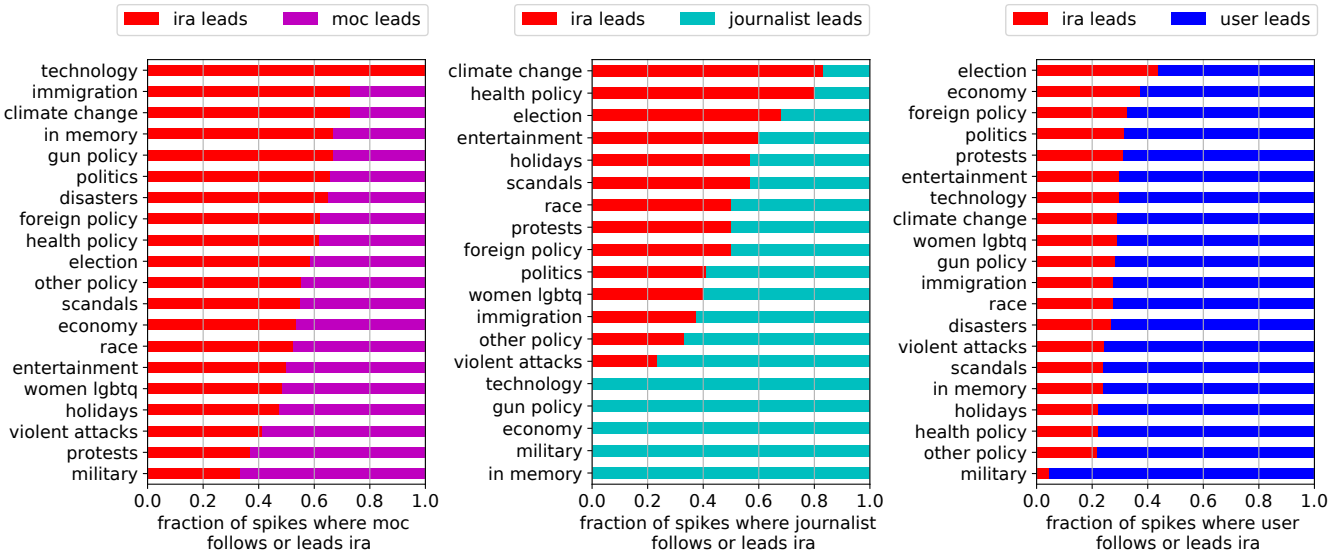



Figure 4: Left panel: Of all spike pairs involving IRA and MOC groups, the fraction of spikes for which IRA leads versus follows; center panel: for journalists; right panel: for users.

human behavior, especially since unopposed efforts to deliberately mislead can harm American democratic institutions (Rodriguez 2019; McKay and Tenove 2020).

Our results provide insights into how the IRA's temporal precedence in other topics has occurred in the context of content focusing on the election. We acknowledge that the IRA is likely aware that simply targeting political candidates in the context of elections would be an ineffective influence campaign. Thus, the IRA would lead (or follow) non-election topics in an attempt to hide its identity and diversify its image, as discussed in Bastos and Farkas (2019).

The identification of a rough division of topics between high and low potential impacts and high and low propensities for the IRA to lead is a novel technique. Such impacts and propensities have likely affected the IRA's decisions regarding the design and prioritization of its content. With evidence that posts about certain topics result in disproportionately greater attention, the IRA might redirect its efforts and thus refocus its audiences' attention in order to increase the polarization and radicalization of the American public. The modification of its communication strategy in the wake of receiving this information would further reflect the IRA's application of an "organic" and reactive approach (DiResta et al. 2018).

To assess the IRA's lead-or-follow tendencies, we have distinguished between groups of Twitter accounts that initially present content on certain topics from those that echo and thus reinforce those topics. In terms of what we have accomplished, we characterized the communication strategies among IRA, sometimes leading and sometimes following; we developed quantitative measures to identify words/phrases that are most indicative of each communication strategy; and we identified the temporal priority of a given set of words and topics by either the IRA or one of the other three groups.

Equipped with the techniques and insights offered here, we call for further IRA-focused research in a similar vein. We suggest that research examine the role of exogenous predictors of IRA-based Twitter activity, such as whether fewer word spikes led by the IRA co-occur with Russian holidays or cold weather in St. Petersburg (Almond, Du, and Vogel 2020). However, we also encourage future research to consider the presence of second-order (or third-order, etc.) spikes, i.e. instances where a topic initially led by one group is followed up by a second group, but then picked up again by the first group without receiving attribution for having started the discussion in the first place. Through a feedback effect, a second-order event that presents a "new" tweet could reinforce an existing narrative or rekindle an old one.

We still do not know definitively whether the benefits of temporal precedence are necessarily greater than those where the IRA may, in times where it does not lead, serve as a conduit to messages posted by others that effectively foster misinformation. The use of multiple dissemination paths would be consistent with the notion of "rumor cascades" (Friggeri et al. 2014). To address this, in line with Weeks and de Zúñiga (2021), one could examine information flows in the context of network analysis.

## 6 Conclusion

The IRA has influenced political conversations in the United States, helping the IRA to foster misinformation, slow down the sharing of accurate information (Vosoughi, Roy, and Aral 2018), increase political polarization, all quite likely to undermine deliberative democracy (McKay and Tenove 2020). While IRA Twitter operations have been described as "opportunistic real time chatter" (DiResta et al. 2018), the IRA invokes the practice of "cyber voter interference" (Hansen and Lim 2019) and continues to modernize and re-

fine the tactics of "active measures". These measures are designed to polarize communities and sow doubt about government, a strategy carried over from the Communist era (Bittman 1985) and adapted for the new media era.

Our broad goal here has been to explain the connections between the content and the temporal precedence of IRA-based Twitter information dissemination to other groups of information receivers, namely MOCs, journalists, and the public. Paths of influence among these four groups are neither uniform, linear, nor simple to predict (Zhang et al. 2021). To be explicit, our work is distinct from a growing body of research that address the distinction between rumors, misinformation, and disinformation (Guess and Lyons 2020), the susceptibility of people to misinformation (Pennycook and Rand 2019), and the means of countering misinformation's effects through an information campaign of one form or another (Wittenberg and Berinsky 2020; Kuklinski et al. 2000).[13] We identified conversations where IRA effectively led, those where it strategically followed, and highlighted potential paths of influence on U.S. politics.

## Ethical Statement

The data in this paper is derived from publicly-accessible user-generated content online. While our focus is on aggregate trending keywords and not individual user characteristics, such data carry risks for issues of privacy and "right-to-be-forgotten." To mitigate these issues and comply with terms of service, we will release only tweet IDs for the data used in this study.

## References


Almond, D.; Du, X.; and Vogel, A. 2020. Russian Holidays Predict Troll Activity 2015-2017. Working Paper 28035, National Bureau of Economic Research. doi:10.3386/w28035. URL http://www.nber.org/papers/w28035.

Aral, S.; and Eckles, D. 2019. Protecting elections from social media manipulation. *Science* 365(6456): 858–861. ISSN 0036-8075. doi:10.1126/science.aaw8243. URL https://science.sciencemag.org/content/365/6456/858.

Badawy, A.; Ferrara, E.; and Lerman, K. 2018. Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. In *2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*, 258–265. IEEE.

Bail, C. A.; Guay, B.; Maloney, E.; Combs, A.; Hillygus, D. S.; Merhout, F.; Freelon, D.; and Volfovsky, A. 2020. Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017. *Proceedings of the National Academy of Sciences* 117(1): 243–250. ISSN 0027-8424. doi:10.1073/pnas.1906420116. URL https://www.pnas.org/content/117/1/243.


[13] Warning labels may have only modest effects on specific beliefs and virtually none on political parties (Guess 2020), but the media itself can reduce public misperceptions with the journalistic fact-checking mechanism (Nyhan et al. 2020).


Barbera, P.; Casas, A.; Nagler, J.; Egan, P. J.; Bonneau, R.; Jost, J. T.; and Tucker, J. A. 2019. Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data. *American Political Science Review* 113(4): 883–901. doi:10.1017/S0003055419000352.

Bastos, M.; and Farkas, J. 2019. "Donald Trump Is My President!": The Internet Research Agency Propaganda Machine. *Social Media + Society* 5(3).

Berinsky, A. J. 2017. Rumors and Health Care Reform: Experiments in Political Misinformation. *British Journal of Political Science* 47(2): 241–262. doi:10.1017/S0007123415000186.

Bittman, L. 1985. *The KGB and Soviet disinformation : an insider's view*. Washington: Pergamon-Brassey's. ISBN 0080315720.

Bovet, A.; and Makse, H. A. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications* 10. URL https://doi.org/10.1038/s41467-018-07761-2.

Dawson, A.; and Innes, M. 2019. How Russia's Internet Research Agency Built its Disinformation Campaign. *The Political Quarterly* 90(2): 245–256. doi:https://doi.org/10.1111/1467-923X.12690. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-923X.12690.

DiFonzo, N.; Bourgeois, M. J.; Suls, J.; Homan, C.; Stupak, N.; Brooks, B. P.; Ross, D. S.; and Bordia, P. 2013. Rumor clustering, consensus, and polarization: Dynamic social impact and self-organization of hearsay. *Journal of Experimental Social Psychology* 49(3): 378–399. ISSN 0022-1031. doi:https://doi.org/10.1016/j.jesp.2012.12.010. URL https://www.sciencedirect.com/science/article/pii/S0022103112002570.

DiResta, R.; Shaffer, K.; Ruppel, B.; Sullivan, D.; Matney, R.; Fox, R. B.; Albright, J.; and Johnson, B. 2018. The tactics & tropes of the Internet Research Agency. *New Knowledge Organization* URL https://apo.org.au/node/211296.

Freelon, D.; Bossetta, M.; Wells, C.; Lukito, J.; Xia, Y.; and Adams, K. 2020. Black Trolls Matter: Racial and Ideological Asymmetries in Social Media Disinformation. *Social Science Computer Review* 0(0): 0894439320914853. doi:10.1177/0894439320914853.

Friggeri, A.; Adamic, L.; Eckles, D.; and Cheng, J. 2014. Rumor Cascades. In *International AAAI Conference on Web and Social Media*. URL https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8122.

Golovchenko, Y.; Buntain, C.; Eady, G.; Brown, M. A.; and Tucker, J. A. 2020. Cross-Platform State Propaganda: Russian Trolls on Twitter and YouTube during the 2016 U.S. Presidential Election. *The International Journal of Press/Politics* 25(3): 357–389. doi:10.1177/1940161220912682. URL https://doi.org/10.1177/1940161220912682.

Guess, A. M. 2020. Misinformation research, four years later. In *News and Information Disorder in the 2020 Presidential Election*.

Guess, A. M.; and Lyons, B. A. 2020. Innovation and Intellectual Property Rights. In Persily, N.; and Tucker, J. A., eds., *Social Media and Democracy: The State of the Field and Prospects for Reform*, chapter 2, 10–33. Cambridge: Cambridge University Press.

Hansen, I.; and Lim, D. J. 2019. Doxing democracy: influencing elections via cyber voter interference. *Contemporary Politics* 25(2): 150–171.

Harknett, R. J. 1996. Information warfare and deterrence. *The US Army War College Quarterly: Parameters* 26(3): 9.

Hemphill, L.; and Schöpke-Gonzalez, A. M. 2020. Two Computational Models for Analyzing Political Attention in Social Media. *Proceedings of the International AAAI Conference on Web and Social Media* 14(1): 260–271. URL https://ojs.aaai.org/index.php/ICWSM/article/view/7297.

Hjorth, F.; and Adler-Nissen, R. 2019. Ideological Asymmetry in the Reach of Pro-Russian Digital Disinformation to United States Audiences. *Journal of Communication* 69(2): 168–192. ISSN 0021-9916.

Howard, P. N.; Ganesh, B.; Liotsiou, D.; Kelly, J.; and Francois, C. 2019. The IRA, Social Media and Political Polarization in the United States, 2012-2018. University of Oxford.

Im, J.; Chandrasekharan, E.; Sargent, J.; Lighthammer, P.; Denby, T.; Bhargava, A.; Hemphill, L.; Jurgens, D.; and Gilbert, E. 2020. Still out there: Modeling and Identifying Russian Troll Accounts on Twitter. In *WebSci 2020 - Proceedings of the 12th ACM Conference on Web Science*, 1–10. doi:10.1145/3394231.3397889.

Jowett, G.; and O'Donnell, V. 2018. *Propaganda & Persuasion*. SAGE Publications. ISBN 9781506371320. URL https://books.google.com/books?id=rB5VDwAAQBAJ.

Kuklinski, J. H.; Quirk, P. J.; Jerit, J.; Schwieder, D.; and Rich, R. F. 2000. Misinformation and the Currency of Democratic Citizenship. *The Journal of Politics* 62(3): 790–816. URL http://www.jstor.org/stable/2647960.

Lazer, D. 2020. Studying human attention on the Internet. *Proceedings of the National Academy of Sciences* 117(1): 21–22. ISSN 0027-8424. doi:10.1073/pnas.1919348117. URL https://www.pnas.org/content/117/1/21.

Lewandowsky, S.; Ecker, U. K. H.; Seifert, C. M.; Schwarz, N.; and Cook, J. 2012. Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest* 13(3): 106–131.

Linvill, D. L.; Boatwright, B. C.; Grant, W. J.; and Warren, P. L. 2019. "THE RUSSIANS ARE HACKING MY BRAIN!" investigating Russia's internet research agency twitter tactics during the 2016 United States presidential campaign. *Computers in Human Behavior* 99: 292–300. ISSN 0747-5632.

Linvill, D. L.; and Warren, P. L. 2020. Troll Factories: Manufacturing Specialized Disinformation on Twitter. *Political Communication* 37(4): 447–467.

Lukito, J. 2020. Coordinating a Multi-Platform Disinformation Campaign: Internet Research Agency Activity on Three U.S. Social Media Platforms, 2015 to 2017. *Political Communication* 37(2): 238–255. doi:10.1080/10584609.2019.1661889.

Lukito, J.; Suk, J.; Zhang, Y.; Doroshenko, L.; Kim, S. J.; Su, M.-H.; Xia, Y.; Freelon, D.; and Wells, C. 2020. The Wolves in Sheep's Clothing: How Russia's Internet Research Agency Tweets Appeared in U.S. News as Vox Populi. *The International Journal of Press/Politics* 25(2): 196–216.

Manning, M.; Manning, M.; Romerstein, H.; Olson, J.; and ProQuest. 2004. *Historical Dictionary of American Propaganda*. Greenwood Press. ISBN 9780313296055. URL https://books.google.com/books?id=1-JjwDPcOLQC.

Martin, D. A.; Shapiro, J. N.; and Nedashkovskaya, M. 2019. Recent Trends in Online Foreign Influence Efforts. *Journal of Information Warfare* 18(3): 15–48. ISSN 14453312, 14453347. URL https://www.jstor.org/stable/26894680.

McKay, S.; and Tenove, C. 2020. Disinformation as a Threat to Deliberative Democracy. *Political Research Quarterly* 0(0).

Nyhan, B.; Porter, E.; Reifler, J.; and Wood, T. J. 2020. Taking Fact-Checks Literally But Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability. *Political Behavior* URL https://www.springerprofessional.de/en/taking-fact-checks-literally-but-not-seriously-the-effects-of-jo/16409994.

Nyhan, B.; and Reifler, J. 2010. When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior* 32. URL https://doi.org/10.1007/s11109-010-9112-2.

Nyhan, B.; and Reifler, J. 2015. Displacing Misinformation about Events: An Experimental Test of Causal Corrections. *Journal of Experimental Political Science* 2(1): 81–93. doi:10.1017/XPS.2014.22.

Pennycook, G.; and Rand, D. G. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188: 39–50. ISSN 0010-0277. The Cognitive Science of Political Thought.

Rodriguez, M. 2019. Disinformation Operations Aimed at (Democratic) Elections in the Context of Public International Law: The Conduct of the Internet Research Agency During the 2016 US Presidential Election. *International Journal of Legal Information* 47(3): 149–197. doi:10.1017/jli.2019.28.

Stewart, L. G.; Arif, A.; and Starbird, K. 2018. Examining trolls and polarization with a retweet network. In *Proc. ACM WSDM, workshop on misinformation and misbehavior mining on the web*, volume 70.

Tyler, M.; Grimmer, J.; and Iyengar, S. 2020. Partisan Enclaves and Information Bazaars: Mapping Selective Exposure to Online News. *Journal of Politics* .

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science* 359(6380): 1146–1151.

ISSN 0036-8075. doi:10.1126/science.aap9559. URL https://science.sciencemag.org/content/359/6380/1146.

Walter, D.; Ophir, Y.; and Jamieson, K. H. 2020. Russian Twitter Accounts and the Partisan Polarization of Vaccine Discourse, 2015–2017. *American Journal of Public Health* 110(5): 718–724.

Weeks, B. E.; and de Zúñiga, H. G. 2021. What's Next? Six Observations for the Future of Political Misinformation Research. *American Behavioral Scientist* 65(2): 277–289.

Wittenberg, C.; and Berinsky, A. J. 2020. Misinformation and Its Correction. In Persily, N.; and Tucker, J. A., eds., *Social Media and Democracy: The State of the Field and Prospects for Reform*, chapter 8, 163–198. Cambridge: Cambridge University Press.

Xia, Y.; Lukito, J.; Zhang, Y.; Wells, C.; Kim, S. J.; and Tong, C. 2019. Disinformation, performed: self-presentation of a Russian IRA account on Twitter. *Information, Communication & Society* 22(11): 1646–1664.

Zannettou, S.; Caulfield, T.; Bradlyn, B.; De Cristofaro, E.; Stringhini, G.; and Blackburn, J. 2020. Characterizing the use of images in state-sponsored information warfare operations by russian trolls on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 774–785.

Zhang, Y.; Lukito, J.; Su, M.-H.; Suk, J.; Xia, Y.; Kim, S. J.; Doroshenko, L.; and Wells, C. 2021. Assembling the Networks and Audiences of Disinformation: How Successful Russian IRA Twitter Accounts Built Their Followings, 2015–2017. *Journal of Communication* ISSN 0021-9916.