Automatic detection of influential actors in disinformation networks

Steven T. Smith[a,1,2], **Edward K. Kao**[a,1], **Erika D. Mackin**[a,1], **Danelle C. Shah**[a], **Olga Simek**[a], **and Donald B. Rubin**[b,c,d,1,2]

[a]MIT Lincoln Laboratory, Lexington MA 02421 USA; [b]Fox School of Business, Temple University, Philadelphia, PA 19122; [c]Yau Mathematical Sciences Center, Tsinghua University, Beijing 100084, China; [d]Department of Statistics, Harvard University, Cambridge, MA 02138

Contributed by Donald B. Rubin, November 10, 2020 (sent for review July 21, 2020; reviewed by Michael Sobel, Kate Starbird, and Stefan Wager)

The weaponization of digital communications and social media to conduct disinformation campaigns at immense scale, speed, and reach presents new challenges to identify and counter hostile influence operations (IOs). This paper presents an end-to-end framework to automate detection of disinformation narratives, networks, and influential actors. The framework integrates natural language processing, machine learning, graph analytics, and a network causal inference approach to quantify the impact of individual actors in spreading IO narratives. We demonstrate its capability on real-world hostile IO campaigns with Twitter datasets collected during the 2017 French presidential elections, and known IO accounts disclosed by Twitter over a broad range of IO campaigns (May 2007 to February 2020), over 50,000 accounts, 17 countries, and different account types including both trolls and bots. Our system detects IO accounts with 96% precision, 79% recall, and 96% area-under-the-PR-curve, maps out salient network communities, and discovers high-impact accounts that escape the lens of traditional impact statistics based on activity counts and network centrality. Results are corroborated with independent sources of known IO accounts from US Congressional reports, investigative journalism, and IO datasets provided by Twitter.

causal inference | networks | machine learning | social media | influence operations

Although propaganda is an ancient mode of statecraft, the weaponization of digital communications and social media to conduct disinformation campaigns at previously unobtainable scales, speeds, and reach presents new challenges to identify and counter hostile influence operations (1–6). Before the internet, the tools used to conduct such campaigns adopted longstanding—but effective—technologies. For example, Mao's guerrilla strategy emphasizes—"[p]ropaganda materials are very important. Every large guerrilla unit should have a printing press and a mimeograph stone" (ref. 7, p. 85). Today, many powers have exploited the internet to spread propaganda and disinformation to weaken their competitors. For example, Russia's official military doctrine calls to "[e]xert simultaneous pressure on the enemy throughout the enemy's territory in the global information space" (ref. 8, section II).

Online influence operations (IOs) are enabled by the low cost, scalability, automation, and speed provided by social media platforms on which a variety of automated and semiautomated innovations are used to spread disinformation (1, 2, 4). Situational awareness of semiautomated IOs at speed and scale requires a semiautomated response capable of detecting and characterizing IO narratives and networks, and estimating their impact either directly within the communications medium, or more broadly in the actions and attitudes of the target audience. This arena presents a challenging, fluid problem whose measured data is comprised of large volumes of human- and machine-generated multimedia content (9), many hybrid interactions within a social media network (10), and actions or consequences resulting from the IO campaign (11). These characteristics of modern IOs can be addressed by recent advances in machine learning in several relevant fields: natural language processing (NLP), semisupervised learning, and network causal inference.

This paper presents a framework to automate detection and characterization of IO campaigns. The contributions of the paper are: 1) an end-to-end system to perform narrative detection, IO account classification, network discovery, and estimation of IO causal impact; 2) a robust semisupervised approach to IO account classification; 3) a method for detection and quantification of causal influence on a network (10); and 4) application of this approach to genuine hostile IO campaigns and datasets, with classifier and impact estimation performance curves evaluated on confirmed IO networks. Our system discovers salient network communities and high-impact accounts in spreading propaganda. The framework integrates natural language processing, machine learning, graph analytics, and network causal inference to quantify the impact of individual actors in spreading IO narratives. Our general dataset was collected over numerous IO scenarios during 2017 and contains nearly 800 million tweets and 13 million accounts. IO account classification is performed using a semisupervised ensemble-tree classifier that uses both semantic and behavioral features, and is trained and tested on accounts

Significance

Hostile influence operations (IOs) that weaponize digital communications and social media pose a rising threat to open democracies. This paper presents a system framework to automate detection of disinformation narratives, networks, and influential actors. The framework integrates natural language processing, machine learning, graph analytics, and network causal inference to quantify the impact of individual actors in spreading the IO narrative. We present a classifier that detects reported IO accounts with 96% precision, 79% recall, and 96% AUPRC, demonstrated on real social media data collected for the 2017 French presidential election and known IO accounts disclosed by Twitter. Our system also discovers salient network communities and high-impact accounts that are independently corroborated by US Congressional reports and investigative journalism.

S.T.S., E.K.K., E.D.M., D.C.S., O.S., and D.B.R. designed research; S.T.S., E.K.K., and E.D.M. performed research; and S.T.S., E.K.K., E.D.M., and D.B.R. wrote the paper.

Reviewers: M.S., Columbia University; K.S., University of Washington; and S.W., Stanford University.

The authors declare no competing interest.

[1]S.T.S., E.K.K., E.D.M., and D.B.R. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: stsmith@ll.mit.edu or rubin@stat.harvard.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2011216118/-/DCSupplemental.

Fig. 1. Framework block diagram of end-to-end IO detection and characterization.

from our general dataset labeled using Twitter's election integrity dataset that contains over 50,000 known IO accounts posting between May 2007 to February 2020 from 17 countries, including both trolls and bots (9). To the extent possible, classifier performance is compared to other online account classifiers. The impact of each account is inferred by its causal contribution to the overall narrative propagation over the entire network, which is not accurately captured by traditional activity- and topology-based impact statistics. The identity of several high-impact accounts are corroborated to be agents of a foreign influence operations or influential participants in known IO campaigns using Twitter's election integrity dataset and reports from the US Congress and investigative journalists (9, 11–15).

Framework

The end-to-end system framework collects contextually relevant data, identifies potential IO narratives, classifies accounts based on their behavior and content, constructs a narrative network, and estimates the impact of accounts or networks in spreading specific narratives (Fig. 1). First, potentially relevant social media content is collected using the Twitter public application programming interface (API) based on keywords, accounts, languages, and spatiotemporal ranges. Second, distinct narratives are identified using topic modeling, from which narratives of interest are identified by analysts. In general, more sophisticated NLP techniques that exploit semantic similarity, e.g., transformer models (16), can be used to identify salient narratives. Third, accounts participating in the selected narrative receive an IO classifier score based on their behavioral, linguistic, and content features. The second and third steps may be repeated to provide a more focused corpus for IO narrative detection. Fourth, the social network of accounts participating in the IO narrative is constructed using their pattern of interactions. Fifth, the unique impact of each account—measured using its contribution to the narrative spread over the network—is quantified using a network causal inference methodology. The end product of this framework is a mapping of the IO narrative network where IO accounts of high impact are identified.

Methodology

Targeted Collection. Contextually relevant Twitter data are collected using the Twitter API based on keywords, accounts, languages, and spatiotemporal filters specified by the authors. For example, during the 2017 French presidential election, keywords

Fig. 2. Word clouds associated with the 'offshore accounts' topic in English (above) and French (below). Topics are selected from those generated from the English corpus ($N = 152{,}203$) and the French corpus ($N = 1{,}070{,}158$) as described in *SI Appendix*.

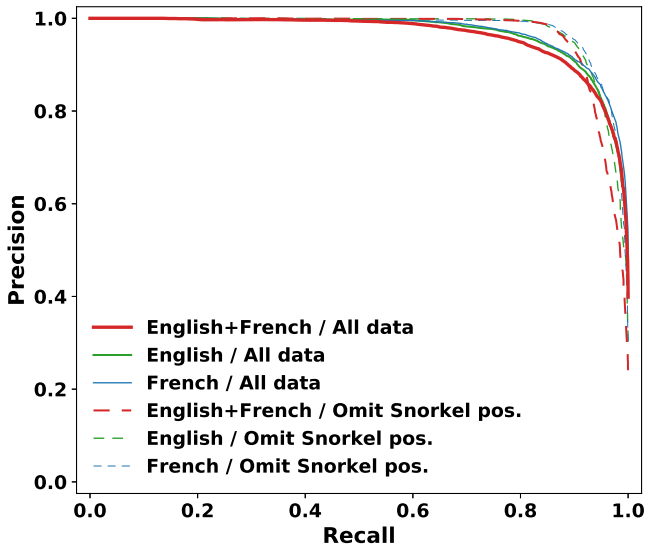


Fig. 3. P-R classifier performance with 3,151 known IO accounts (9) and language-specific training data: combined English and French (red curves, $N = 17{,}999$); English only (green curves, $N = 13{,}155$); French only (blue curves, $N = 13{,}159$). Cross-validated using a 90 : 10 split and all data (solid curves), and Snorkel positives omitted (dashed curves).

include the leading candidates, #Macron and #LePen, French election-related issues, including hostile narratives expected to harm specific candidates, e.g., voter abstention (17) and unsubstantiated allegations (6, 18). Because specific narratives and influential IO accounts are discovered subsequently, they offer additional cues to either broaden or refocus the collection. In the analysis described in the preceding subsection, 28 million Twitter posts and nearly 1 million accounts potentially relevant to the 2017 French presidential election were collected over a 30-d period preceding the election, and a total of nearly 800 million tweets and information on 13 million distinct accounts were collected.

Narrative Detection. Narratives are automatically generated from the targeted Twitter data using a topic modeling algorithm (19). First, accounts whose tweets contain keywords relevant to the subject, or exhibit pre-defined, heuristic behavioral patterns within a relevant time period are identified. Second, content from these accounts are passed to a topic modeling algorithm, and all topics are represented by a collection, or bag of words. Third, interesting topics are identified manually. Fourth, tweets that match these topics above a pre-defined threshold are selected. Fifth, a narrative network is constructed with vertices defined by accounts whose content matches the selected narrative, and edges defined by retweets between these accounts. In the case of the 2017 French elections, the relevant keywords are: "Macron," "leaks," "election," and "France"; the languages used in topic modeling are English and French.

IO Account Classification. Developing an automated IO classifier is challenging because the number of actual examples is necessarily small, and the behavior and content of these accounts can change over both time and scenario. Semisupervised classifiers trained using heuristic rules can address these challenges by augmenting limited truth data with additional accounts that match IO heuristics within a target narrative that does not necessarily appear in the training data. This approach has the additional intrinsic benefits of preventing the classifier from being overfit to a specific training set or narrative, and provides a path to adapt the classifier to future narratives. IO account classifier design is implemented using this semisupervised machine-learning approach built with the open source libraries scikit-learn and Snorkel (20, 21), and soft labeling functions based on heuristics of IO account metadata, content, and behavior. The feature space comprises behavioral characteristics, profile characteristics, languages used, and the 1- and 2-grams used in tweets; full details are provided in *SI Appendix*, section **??**, and feature importances are illustrated in *Results* (see Fig. 4). Our design approach to semisupervised learning is to develop labeling functions only for behavioral and profile characteristics, not narrative- or content-specific features, with the expectation that such labeling functions are transferable across multiple scenarios.

The classifier is trained and tested using sampled accounts representing both IO-related and general narratives. The approach is designed to prevent overfitting by labeling these sampled accounts using semisupervised Snorkel heuristics. These accounts are collected as described in *Targeted Collection*. There are four categories of training and testing data: known IO accounts (known positives from the publicly available Twitter elections

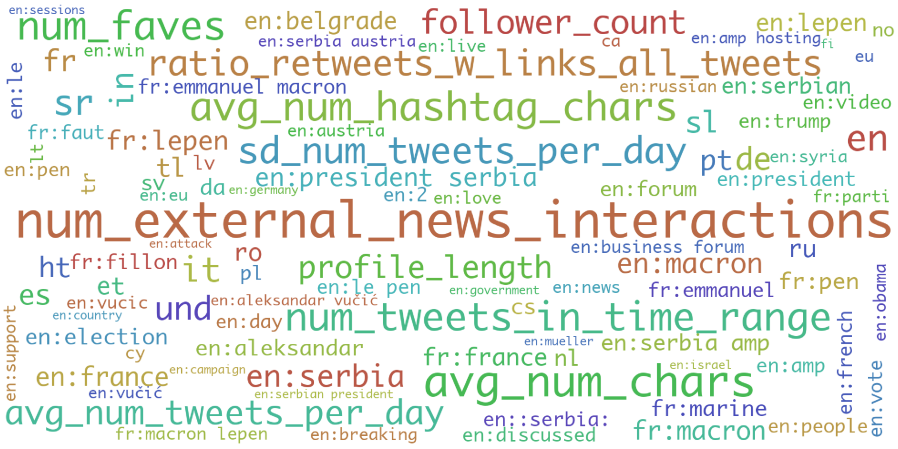


Fig. 4. The 100 most important features in the French and English combined classifier, represented by relative size in a word cloud.

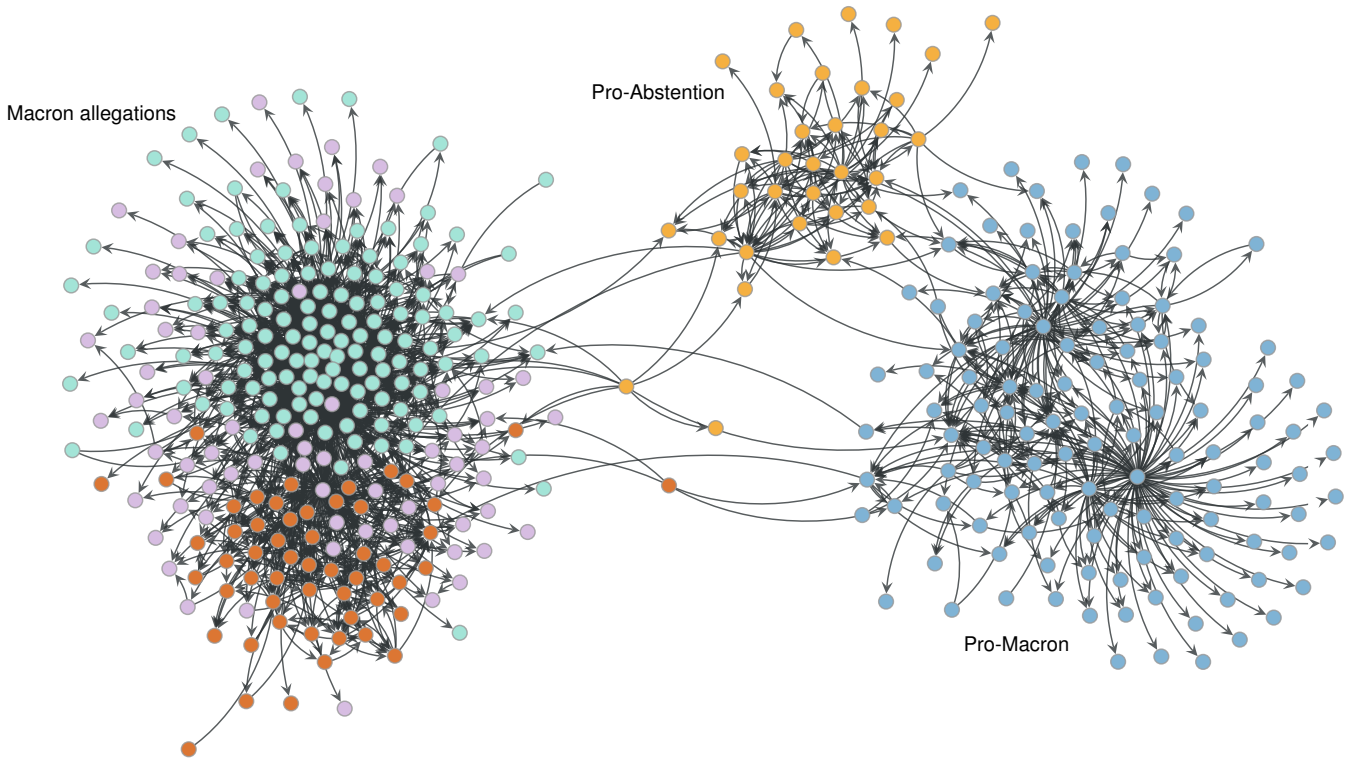


Fig. 5. Community membership in the French narrative network (Fig. 2). Colors show inferred membership from a blockmodel (22).

integrity dataset) (9), known non-IO accounts (known negatives comprised of mainstream media accounts), Snorkel-labeled positive accounts (heuristic positives), and Snorkel-labeled negative accounts (heuristic negatives). All known positives are eligible to be in the training set, whether they are manually operated troll IO accounts or automated bots. Because we include samples that represent both accounts engaging in the IO campaign and general accounts, our training approach is weakly dependent upon the IO campaign in two ways. First, only known IO accounts whose content includes languages relevant to the IO campaign are used. Second, a significant fraction (e.g., 67%) of Snorkel-labeled accounts (either positive or negative) must be participants in the narrative. Finally, we establish confidence that we are not overfitting by using cross-validation to compute classifier performance. Additionally, overfitting is observed without using Snorkel heuristics (see *SI Appendix*, section **??**), supporting the claim that semisupervised learning is a necessary component of our design. Dimensionality reduction and classifier algorithm selection is performed by optimizing precision–recall performance over a broad set of dimensionality reduction approaches, classifiers, and parameters (see *SI Appendix*, section **??**). In the results section, dimensionality reduction is performed with Extra-Trees (ET) (20) and the classifier is the Random Forest (RF) algorithm (20). Ensemble tree classifiers learn the complex concepts of IO account behaviors and characteristics without over-fitting to the training data through a collection of decision trees, each representing a simple concept using only a few features.

Network Discovery. The narrative network—a social network of participants involved in discussing and propagating a specific narrative—is constructed from their observed pattern of interactions. In *Results*, narrative networks are constructed using retweets. Narrative networks and their pattern of influence are represented as graphs whose edges represent strength of interactions. The (directed) influence from (account) vertex v_i to vertex v_j is denoted by the weighted edge a_{ij}. For simplicity in the sequel, network vertex v_i is also referred to as i. The influence network is represented by the adjacency matrix $\boldsymbol{A} = (a_{ij})$. Because actual influence is not directly observable, the influence network is modeled as a random variable with Poisson distribution parameterized by the observed evidence of influence. Specifically, influence a_{ij} is modeled with prior distribution $a_{ij} \sim$ Poisson(frequency of interactions from i to j), as counts of interactive influence in real-world networks. Observations of past interactions or influence on a subset of edges can be used to estimate the rates on the missing edges through inference on a network model that captures realistic characteristics such as sparsity, varying vertex degrees, and community structure (23).

Impact Estimation. Impact estimation is based on a method that quantifies each account's unique causal contribution to the overall narrative propagation over the entire network. It accounts for social confounders (e.g., community membership, popularity) and disentangles their effects from the causal estimation. This approach is based on the network potential outcome framework (24), itself based upon Rubin's causal framework (25). Mathematical details are provided in *SI Appendix*, section **??**.

The fundamental quantity is the network potential outcome of each vertex, denoted $Y_i(\boldsymbol{Z}, \boldsymbol{A})$, under exposure to the narrative from the source vector $\boldsymbol{Z}$ via the influence network $\boldsymbol{A}$. Precisely, $\boldsymbol{Z}$ is a binary N-vector of narrative sources (a.k.a. treatment vector). In this study, vertices are user accounts, edges represent influence as described in *Network Discovery*, and the potential outcomes are the number of tweets in the narrative. The influence

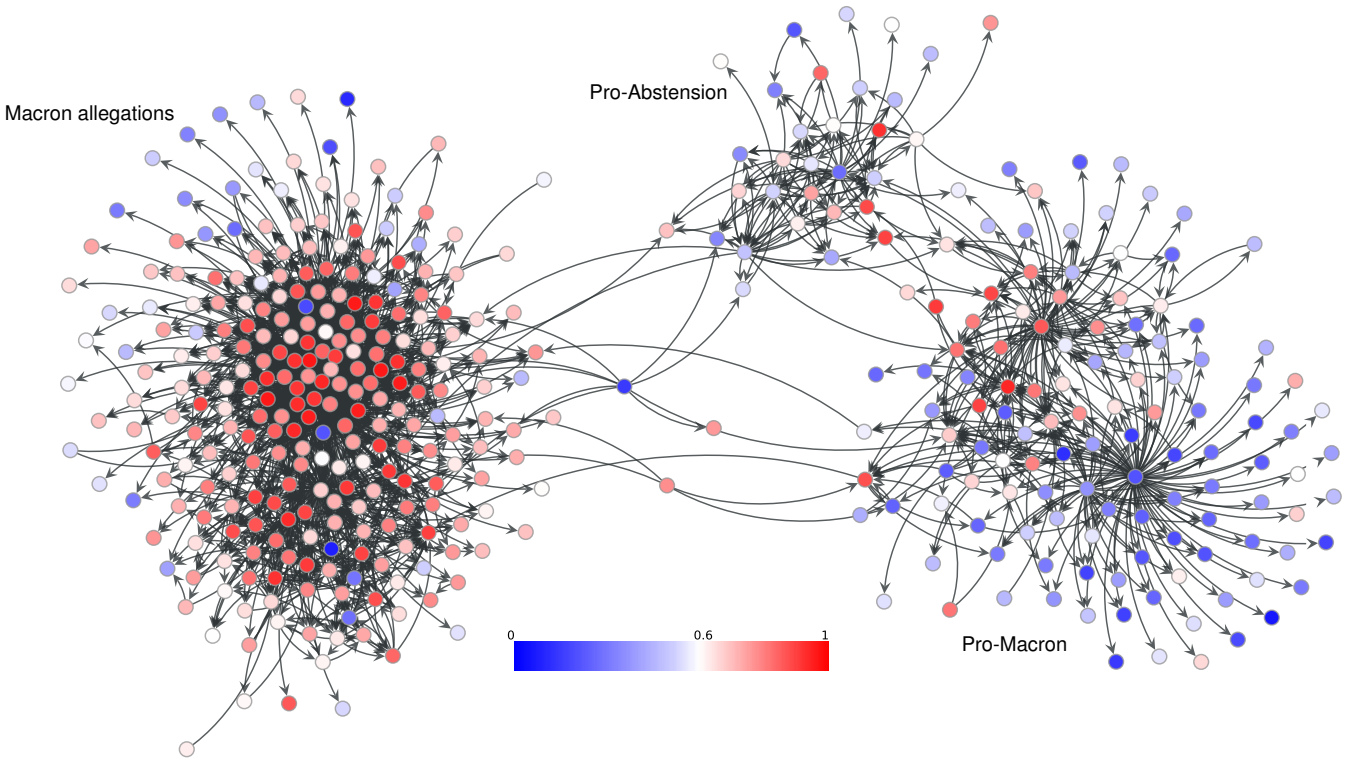


Fig. 6. Classifier scores over the French narrative network (Fig. 2). The 0–1 score range indicates increasing similarity to known IO accounts.

network is an important part of the treatment exposure mechanism. An account's exposure to the narrative is determined by both the sources as well as exposures to them delivered through the influence network. The impact ζ_j of each vertex j on the overall narrative propagation is defined using network potential outcome differentials averaged over the entire network:

$$\zeta_j(z) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N}(Y_i(\boldsymbol{Z}=\boldsymbol{z}_{j+},\boldsymbol{A}) - Y_i(\boldsymbol{Z}=\boldsymbol{z}_{j-},\boldsymbol{A})). \quad [1]$$

This causal estimand is the average difference between the individual outcomes with v_j as a source such that $\boldsymbol{z}_{j+} := (z_1,\ldots,z_j := 1,\ldots,z_N)^{\mathrm{T}}$, versus v_j not a source, $\boldsymbol{z}_{j-} := (z_1,\ldots,z_j := 0,\ldots,z_N)^{\mathrm{T}}$. This impact is the average (per vertex) number of additional tweets generated by an user's participation in the narrative. The source is said to be uniquely impactful if it is the only source.

It is impossible to observe the outcomes at each vertex with both exposure conditions under source vectors $\boldsymbol{z}_{j+}$ and $\boldsymbol{z}_{j-}$; therefore, the missing potential outcomes must be estimated, which can be accomplished using a model. After estimating the model parameters from the observed outcomes and vertex covariates, missing potential outcomes in the causal estimand ζ_j can be imputed using the fitted model. Potential outcomes are modeled using a Poisson generalized linear mixed model (GLMM) with the canonical log link function and linear predictor coefficients (τ,γ,β,μ) corresponding to the source indicator Z_i, n-hop exposure vector $s_i^{(n)}$, the covariate vector $\boldsymbol{x}_i$, and the baseline outcome. The covariate vector $\boldsymbol{x}_i$ includes the potential social confounders such as popularity and community membership. These confounders are accounted for through covariate adjustment in order to disentangle actual causal impact from effects of homophily (birds of a feather flock together) and vertex degree on outcomes, by meeting the key unconfounded influence network assumption **??** in *SI Appendix*, *Assumption* **??**. For correctness and rigor, imputation of the missing potential outcomes are designed to meet the unconfoundedness assumptions that lead to ignorable treatment exposure mechanism under network interference, detailed in *SI Appendix*, section **??**. The GLMM model for the potential outcomes is

$$Y_i(\boldsymbol{Z},\boldsymbol{A}) \sim \text{Poisson}(\lambda_i),$$

$$\log\lambda_i = \tau Z_i + \left(\sum_{n=1}^{N_{\text{hop}}} s_i^{(n)}\tau\prod_{k=1}^{n}\gamma_k\right) + \beta^{\mathrm{T}}\boldsymbol{x}_i + \mu + \epsilon_i, \quad [2]$$

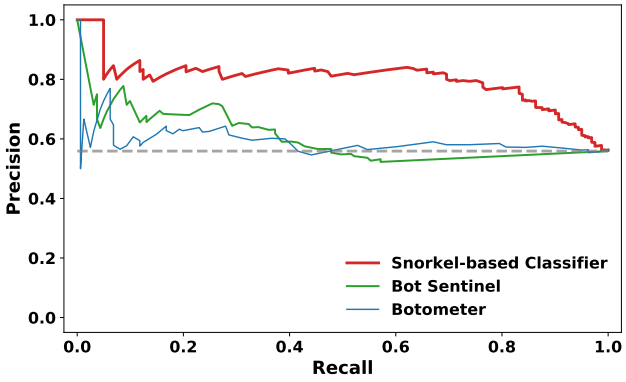


Fig. 7. Classifier performance comparison: Snorkel-based classifier (red curve, $N = 415$) vs. Botometer (green curve, $N = 289$) vs. Bot Sentinel (blue curve, $N = 288$), given proxy, community-based truth. The dashed gray horizontal line at 63% is the fraction of presumptively true examples in the community-based proxy for known IO accounts, and therefore represents random chance precision performance.

where τZ_i represents the primary effect from the source, $\sum_{n=1}^{N_{\text{hop}}} s_i^{(n)} \tau \prod_{k=1}^{n} \gamma_k$ represents the accumulative social influence effect from n-hop exposures $s_i^{(n)}$ to the source, γ_k (between 0 and 1) represents how quickly the effect decays over each additional kth hop, $\beta^{\mathrm{T}} \boldsymbol{x}_i$ is the effect of the unit covariates $\boldsymbol{x}_i$ including potential social confounders such as popularity and community membership, μ is the baseline effect on the entire population, and $\epsilon_i \sim \text{Normal}(0, \sigma_\epsilon^2)$ provides independent and identically distributed variation for heterogeneity between the units. The amounts of social exposure at the nth hop are determined by $(\boldsymbol{A}^{\mathrm{T}})^n \boldsymbol{Z}$. This captures narrative propagation via all exposure to sources within the narrative network. Diminishing return of additional exposures is modeled using (elementwise) log-exposure, $\boldsymbol{s}^{(n)} = \log((\boldsymbol{A}^{\mathrm{T}})^n \boldsymbol{Z} + 1)$. The influence matrix $\boldsymbol{A}$, with prior distribution specified in *Network Discovery*, is jointly estimated with the model parameters τ, γ, β, μ, through Markov Chain Monte Carlo (MCMC) and Bayesian regression.

Results

Targeted Collection. The targeted collection for the 2017 French presidential election includes 28,896,185 potentially relevant tweets and 999,883 distinct accounts, all collected over a 30-d period preceding the election on 7 May 2017. Targeted collections for several other IO scenarios were also collected during 2017, resulting in a dataset with 782,678,201 tweets and 12,723,995 distinct accounts. Of these, there are 3,151 known IO accounts that posted in English or French (see *SI Appendix*, Fig. **??**). The great majority of accounts are unrelated to both these known IO accounts and the French election, but will provide negative examples for IO classifier training.

Narrative Detection. Narratives immediately preceding the election are generated automatically by dividing this broad content into language groups, restricting the content and time period to election-related posts within 1 wk preceding the election's media blackout date of 5 May 2017, and filtering accounts based

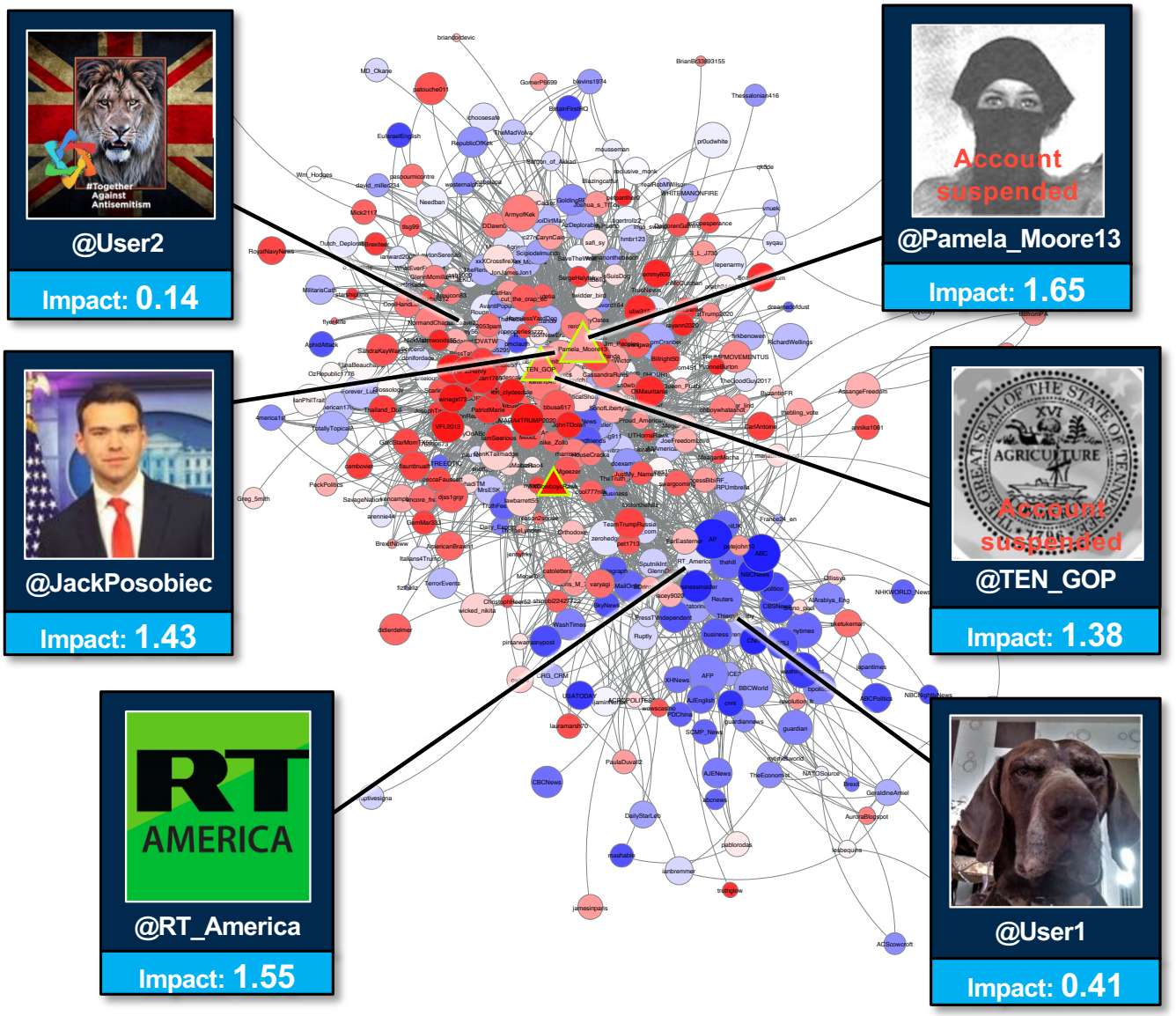


Fig. 8. Impact network (accounts sized by impact) colored by IO classifier score on the English narrative network Fig. 2). Known IO accounts are highlighted in triangles. Image credits: Twitter/JackPosobiec, Twitter/RT_America, Twitter/Pamela_Moore13, Twitter/TEN_GOP.

on interaction with non-French media sites pushing narratives expected to harm specific candidates. Topic modeling (19) is applied to the separate English and French language corpora, and the resulting topics are inspected by the authors to identify relevant narratives. Two such narratives are illustrated in Fig. 2 by the most-frequent word and emoji usage appearing in tweets included within the topic. From the English corpus (N = 152,203), 15 topics are generated; from the French (N = 1,070,158), 30 topics. These automatically generated topics correspond closely to allegations claimed to be spread by WikiLeaks, candidate Marine Le Pen, and others (14, 26). The role of bots in spreading these allegations using the #MacronLeaks hashtag narrative is studied by Ferrara (27). Two topics, one in English and one in French, pertaining to unsubstantiated financial allegations (Fig. 2) will be used to identify accounts involved in spreading these narratives, independent of whether the accounts were used for narrative detection.

IO Account Classification. Twitter published IO truth data containing 50,230 known IO account identities and their multimedia content (9). IO accounts represent a tiny fraction of all Twitter accounts, e.g., Twitter's 50,230 known IO accounts are only 0.02% of its 330 million active monthly users, and bots are estimated to comprise 9–15% of all Twitter accounts (28). This dataset is used along with heuristic rules to train a semisupervised classifier (21) using the approach described in *Methodology*, and the rulesets are detailed in the *SI Appendix*, section **??**. The classifier is trained and tested using content from these sources: known IO accounts that have tweeted on any topic at least once in either English or French, 20 known non-IO mainstream media accounts, and Snorkel-labeled accounts randomly drawn from our dataset, such that 67% are topically relevant and 33% are topically neutral. There are 3,151 known IO accounts from Twitter's dataset in our dataset that have tweeted in English or French. To account for a possible upper bound of up to 15% bots (IO-related or not), we randomly select 15,000 presumptively false examples with 5,000 each of three false classes: topically relevant accounts that tweeted at least once in English, at least once in French, and topically neutral, randomly chosen from the general dataset (see *SI Appendix*, Fig. **??**).

Precision-recall performance. Precision-recall (P-R) performance of the classifier is computed via cross validation using the same dataset with a 90 : 10 split, averaged over 20 rounds (Fig. 3). Because the number of known cases is limited, weak supervision from the heuristic Snorkel labeling functions is used to identify true examples before cross validation (21). All training is performed with Snorkel-labeled data, and cross validation is performed both using both Snorkel-labeled data (solid curves in Fig. 3) and omitting Snorkel positives (dashed curves). Sensitivity to language-specific training data is also computed by restricting topically relevant, false examples to specific languages. All classifiers exhibit comparably strong performance, and small differences in relative performance is consistent with the authors' expectations. All classifiers detect IO accounts with 96% precision and 79% recall at a nominal operating point, 96% area-under-the-PR-curve (AUPRC), and 8% equal-error rate (EER). The English-only and French-only classifiers perform slightly better (nominally 0–3%) than the combined language model, consistent with the expectation that models with greater specificity outperform less specific models. This strong classifier performance will be combined with additional inferences—narrative networks, community structure, and impact estimation—to identify potentially influential IO accounts involved in spreading particular narratives.

With this dataset, the original feature space has dimension 1,896,163: 17 behavior and profile features, 61 languages, and 1,896,085 1- and 2-grams. *SI Appendix*, section **??** lists the behavior, profile features, and language features. Grid search over feature dimensionality is used to identify the best feature set for dimensionality reduction: 10 behavioral and profile features, 30 language features, and 500 1- and 2-grams. The most important features used by the classifier are illustrated by the relative sizes of feature names appearing in the word cloud of Fig. 4. Note that the most important features for IO account classification are independent of topic, and pertain instead to account behavioral characteristics and frequency of languages other than English or French. This topic independence suggests the potential applicability of the classifier to other IO narratives. Furthermore, the diversity of behavioral features suggests robustness against future changes of any single behavior.

Classifier performance comparisons. Several online bot classifiers are used to report upon and study influence campaigns (3, 27, 29), notably Botometer (formerly BotOrNot) (30) and Bot Sentinel (31). Indeed, Rauchfleisch and Kaiser (32) assert based on several influential papers that analyze online political influence that "Botometer is the de-facto standard of bot detection in academia" (p. 2). In spite of the differences between general, automated bot activity and the combination of troll and bot accounts used for IO campaigns, comparing the classifier performance between these different classifiers is important because it is widespread practice to use such bot classifiers for insight into IO campaigns. Therefore, we compare the P-R performance of our IO classifier to both Botometer and Bot Sentinel. This comparison is complicated by three factors: 1) neither Botometer nor Bot Sentinel have published classifier performance against known IO accounts; 2) neither project has posted open source code; and 3) known IO accounts are immediately suspended, which prevents post hoc analysis with these online tools.

Therefore, a proxy for known IO accounts must be used for performance comparisons. We use the observation that there exists strong correlation between likely IO accounts in our narrative network and membership in specific, distinguishable communities independently computed using an MCMC-based blockmodel (22). Community membership of accounts in the French language narrative network (Fig. 2) are illustrated in Fig. 5. Five distinct communities are detected, three of which are identified

Table 1. Comparison of impact statistics between accounts on the English network: tweets (T), retweets (RT), followers (F), first tweet time on 28 April, PageRank centrality (PR), and causal impact* (CI).

Screen name	T	RT	F	1st time	PR	CI*
@RT_America	39	8	386k	12:00	2706	**1.55**
@JackPosobiec	28	123	23k	01:54	4690	**1.43**
@User1†	8	0	1.4k	22:53	44	**0.14**
@User2†	12	15	19k	12:27	151	**0.41**
@Pamela_Moore13	10	31	56k	18:46	97	**1.65**
@TEN_GOP	12	42	112k	22:15	191	**1.38**

*Estimate of the causal estimand in Equation [1]
†Anonymized screen names of currently active accounts

to have promoted Macron allegation narratives. The other two narratives promote pro-Macron and pro-abstention narratives. Accounts in this narrative network are classified on a 0–1 scale of their similarity to known IO accounts, shown in Fig. 6. Comparing Figs. 5 and 6 shows that the great majority of accounts in the 'Macron allegation' communities are classified as highly similar to known IO accounts, and conversely, the great majority of accounts in the pro-Macron and pro-abstention communities are classified as highly dissimilar to known IO accounts. This visual comparison is quantified by the account histogram illustrated in *SI Appendix*, Fig. **??**.

Using membership in these Macron allegation communities as a proxy for known IO accounts, P-R performance is computed for our IO classifier, Botometer and Bot Sentinel (Fig. 7). Note that Botometer's performance in Fig. 7 at a nominal 50% recall is 56% precision, which is very close to the 50% Botometer precision performance shown by Rauchfleisch and Kaiser using a distinctly different dataset and truthing methodology (32, Fig. 4, "all"). Given this narrative network and truth proxy, both Botometer and Bot Sentinel perform nominally at random chance of 63% precision, the fraction of presumptive IO accounts. Our IO classifier has precision performance of 82–85% over recalls of range 20–80%, which exceeds random chance performance by 19–22%. These results are also qualitatively consistent with known issues of false positives and false negatives in bot detectors (32), though some performance differences are also likely caused by the in-

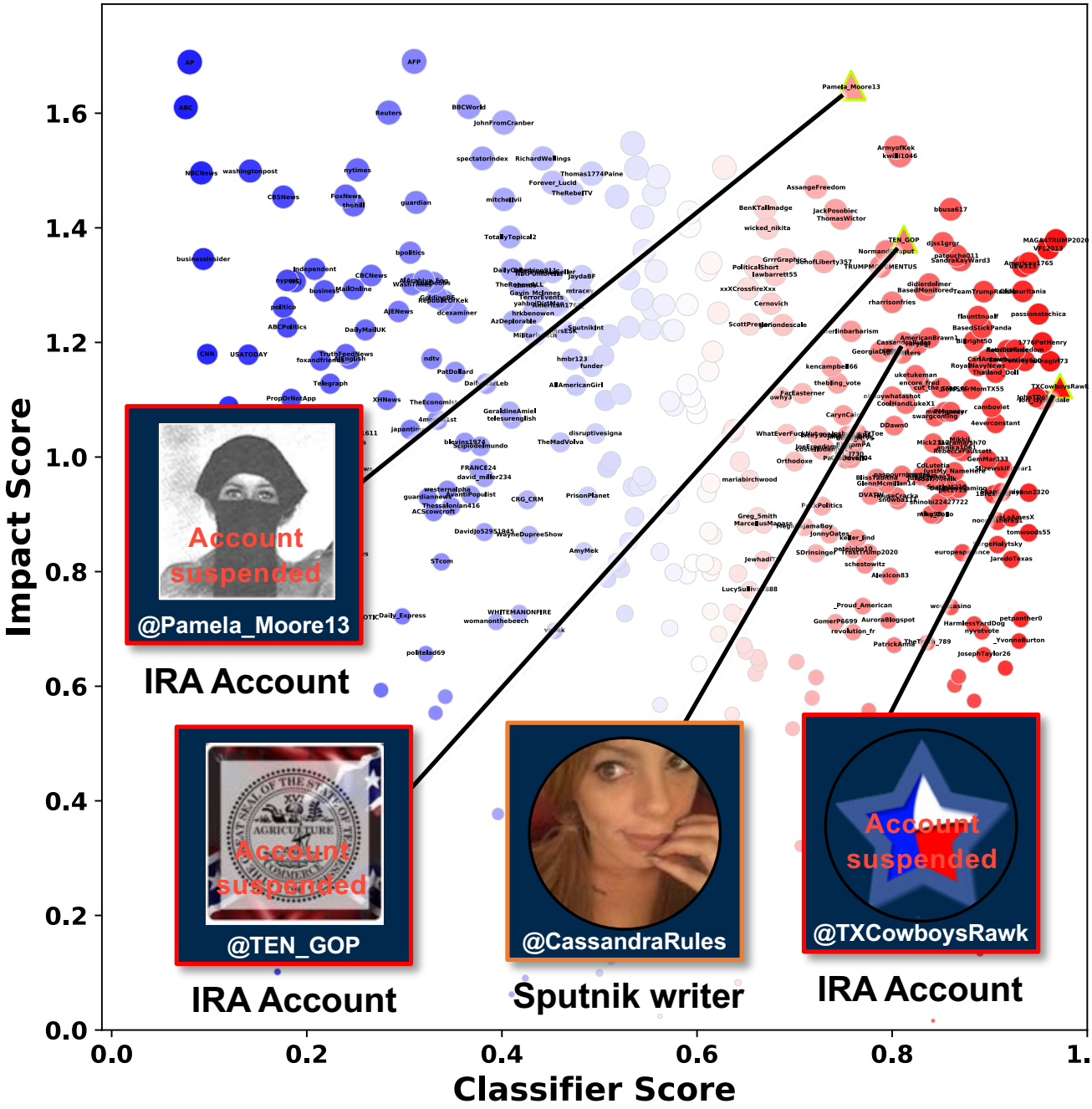


Fig. 9. Impact versus classifier score, English narrative network (Fig. 2). Known IO accounts run by the Internet Research Agency (IRA) (9, 11–13, 15) are highlighted. Image credits: Twitter/Pamela_Moore13, Twitter/TEN_GOP, Twitter/CassandraRules, Twitter/TXCowboysRawk.

tended design of Botometer or Bot Sentinel, which is to detect general bot activity, rather than the specific IO behavior on which our classifier is trained.

Network Discovery. Tweets that match topics of Fig. 2 are extracted from the collected Twitter data. French-language tweets made in the week leading up to the blackout period, 28 April through 5 May 2017, are checked for similarity to the French language topic. To ensure the inclusion of tweets on the #MacronLeaks data dump (10, 14, 27), which occurred on the eve of the French media blackout, English-language tweets from 29 April through 7 May 2017 are compared to the English topic. In total, the French topic network consists of 6,927 accounts and the English topic network consists of 1,897 accounts. For visual clarity, network figures are generated on the most active accounts, 459 in the French and 752 in the English networks.

Impact Estimation. Estimation on the causal impact of each account in propagating the narrative is performed by computing the estimand in Eq. [1], considering each account as the source. Unlike existing propagation methods on network topology (33), causal inference accounts also for the observed counts from each account to capture how each source contributes to the subsequent tweets made by other accounts. Results demonstrate this method's advantage over traditional impact statistics based on activity count and network topology alone.

Impact estimation and IO classification on the English narrative network (Fig. 2) are demonstrated in Fig. 8. Graph vertices are Twitter accounts sized by the causal impact score (i.e. posterior mean of the causal estimand) and colored by the IO classifier using the same scale as Fig. 6. Redness indicates account behavior and content like known IO accounts, whereas blueness indicates the opposite. This graph layout reveals two major communities involved in narrative propagation of unsubstantiated financial allegations during the French election. The large community at the top left comprises many accounts whose behavior and content is consistent with known IO accounts. The relatively smaller community at the lower right includes many mainstream media accounts that publish reports on this narrative. Within this mainstream journalism-focused community, the media accounts AP, ABC, RT, and Reuters are among the most impactful, consistent with expectation. The most remarkable result, however, is that known IO accounts are among the most impactful among the large community of IO-like accounts. The existence of these IO accounts was known previously (9, 12, 13), but not their impact in spreading specific IO narratives. Also note the impactful IO accounts (i.e. the large red vertices) in the upper community that appear to target many benign accounts (i.e. the white and blue vertices).

A comparison between the causal impact and traditional impact statistics is provided in Table 1 on several representative and/or noteworthy accounts highlighted in Fig. 8. The prominent @RT_America, a major Russian media outlet, and @JackPosobiec, a widely reported (14) account in spreading this narrative, corroborate our estimate of their very high causal impact scores. This is also consistent with their early participation in this narrative, high tweet counts, high number of followers, and large PageRank centralities. Conversely, @User1 and @User2 have low impact statistics, and also receive low causal impact scores as relatively non-impactful accounts. It is often possible to interpret why accounts were impactful. E.g., @JackPosobiec was one of the earliest participants and has been reported as a key source in pushing the related #MacronLeaks narrative (14, 27) (see *SI Appendix*, Fig. **??**). In that same narrative, another impactful account @UserB serves as the initial bridge from the English sub-network into the predominantly French-speaking sub-network (see *SI Appendix*, section **??**).

Known IO account (9–13) @Pamela_Moore13's involvement in this narrative illustrates the relative strength of the causal impact estimates in identifying relevant IO accounts. @Pamela_Moore13 stands out as one of the most prominent accounts spreading this narrative. Yet 'her' other impact statistics (T, RT, F, PR) are not distinctive, and comparable in value to the not-impactful account @User2. Additionally, known IO accounts @TEN_GOP and @TXCowboysRawk (9, 12, 13, 15), and Sputik writer @CassandraRules (34) all stand out for their relatively high causal impact and IO account classifier scores (Fig. 9). Causal impact estimation is shown to find high-impact accounts that do not stand out using traditional impact statistics. This estimation is accomplished by considering how the narrative propagates over the influence network, and its utility is demonstrated using data from known IO accounts on known IO narratives. Additional impact estimation results are provided in the *SI Appendix*, section **??**.

Influential IO Account Detection. The outcome of the automated framework proposed in this manuscript is the identification of influential IO accounts in spreading IO narratives. This is accomplished by combining IO classifier scores with IO impact scores for a specific narrative (Fig. 9). Accounts whose behavior and content appear like known IO accounts and whose impact in spreading an IO narrative is relatively high are of potential interest. Such accounts appear in the upper-right side of the scatterplot illustrated in Fig. 9. Partial validation of this approach is provided by the known IO accounts discussed above. Many other accounts in the upper-right side of Fig. 9 have since been suspended by Twitter, and some at the time of writing are actively spreading conspiracy theories about the 2020 coronavirus pandemic (35). These currently-active accounts participate in IO-aligned narratives across multiple geo-political regions and topics, and no matter their authenticity, their content is used hundreds of times by known IO accounts (9) (see *SI Appendix*, section **??**). Also note that this approach identifies both managed IO accounts [e.g., @Pamela_Moore13, @TEN_GOP and @TXCowboysRawk (9, 12, 13, 15)] as well as accounts of real individuals [@JackPosobiec and @CassandraRules (14, 34)] involved in the spread of IO narratives. As an effective tool for situational awareness, the framework in this manuscript can alert social media platform providers and the public of influential IO accounts and networks, and the content they spread.

Discussion

We present a framework to automate detection of disinformation narratives, networks, and influential actors. The framework integrates NLP, machine learning, graph analytics, and network causal inference to quantify the impact of individual actors in spreading the IO narrative. Application of this framework to several authentic influence operation campaigns run during the 2017 French elections provides alerts to likely IO accounts that are influential in spreading IO narratives. Our results are corroborated by independent press reports, US Congressional reports,

and Twitter's election integrity dataset. The detection of IO narratives and high-impact accounts is demonstrated on a dataset comprising 29 million Twitter posts and 1 million accounts collected in 30 days leading up to the 2017 French elections. We also measure and compare the classification performance of a semisupervised classifier for IO accounts involved in spreading specific IO narratives. At a representative operating point, our classifier performs with 96% precision, 79% recall, 96% AUPRC, and 8% EER. Our classifier precision is shown to outperform two online Bot detectors by 20% (nominally) at this operating point, conditioned on a network-community-based truth model. A causal network inference approach is used to quantify the impact of accounts spreading specific narratives. This method accounts for the influence network topology and the observed volume from each account, and removes the effects of social confounders (e.g., community membership, popularity). We demonstrate the approach's advantage over traditional impact statistics based on activity count (e.g., tweet and retweet counts) and network topology (e.g., network centralities) alone in discovering high-impact IO accounts that are independently corroborated.

Data Availability. Comma-separated value (CSV) data of the narrative networks analyzed in this paper have been deposited in GitHub (https://github.com/Influence-Disinformation-Networks/PNAS-Narrative-Networks) and Zenodo (https://doi.org/10.5281/zenodo.4361708).

ACKNOWLEDGMENTS. This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering.

References

1. G. King, J. Pan, M. E. Roberts (2017) How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review* 111(3):484–501. doi:10.1017/S0003055417000144.
2. M. Stella, E. Ferrara, M. D. Domenico (2018) Bots increase exposure to negative and inflammatory content in online social systems. *Proc. Natl. Acad. Sci. U.S.A.* 115(49):12435–12440. doi:10.1073/pnas.1803470115.
3. S. Vosoughi, D. Roy, S. Aral (2018) The spread of true and false news online. *Science* 359(6380):1146–1151. doi:10.1126/science.aap9559.
4. K. Starbird (2019) Disinformation's spread: bots, trolls and all of us. *Nature* 571:449. doi:10.1038/d41586-019-02235-x.
5. T. Rid (2020) *Active Measures: The Secret History of Disinformation and Political Warfare*. (Farrar, Straus and Giroux, New York NY).
6. M. S. Schmidt, N. Perlroth (2020) U. S. charges Russian intelligence officers in major cyberattacks. *The New York Times*. https://www.nytimes.com/2020/10/19/us/politics/russian-intelligence-cyberattacks.html. Accessed 19 October 2020.
7. M. Tse-tung (1937) *On Guerilla Warfare*, FMFRP 12-18. (U. S. Marine Corps). https://www.marines.mil/Portals/1/Publications/FMFRP%2012-18%20%20Mao%20Tse-tung%20on%20Guerrilla%20Warfare.pdf. Accessed 1 March 2020.
8. V. Putin (2014) The military doctrine of the Russian Federation. https://rusemb.org.uk/press/2029. Accessed 1 January 2018.
9. V. Gadde, Y. Roth (2018) Enabling further research of information operations on Twitter. https://blog.twitter.com/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.html. Accessed 1 January 2020.
10. S. T. Smith, E. K. Kao, D. C. Shah, O. Simek, D. B. Rubin (2018) Influence estimation on social media networks using causal inference in *Proc. 2018 IEEE Statistical Signal Processing Workshop (SSP)*. pp. 28–32. doi:10.1109/SSP.2018.8450823.
11. E. Birnbaum (2019) Mueller identified 'dozens' of US rallies organized by Russian troll farm. *The Hill*. https://thehill.com/policy/technology/439532-mueller-identified-dozens-of-us-rallies-organized-by-russian-troll-farm. Accessed 18 May 2019.
12. US House Permanent Select Committee on Intelligence (2017) HPSCI minority exhibits during open hearing, memorandum. https://democrats-intelligence.house.gov/uploadedfiles/hpsci_minority_exhibits_memo_11.1.17.pdf. Accessed 1 January 2018.
13. US House Permanent Select Committee on Intelligence (2017) Exhibit of the user account handles that Twitter has identified as being tied to Russia's "Internet Research Agency.". https://democrats-intelligence.house.gov/uploadedfiles/exhibit_b.pdf. Accessed 1 January 2018.
14. A. Marantz (2017) The far-right American nationalist who tweeted #MacronLeaks. *The New Yorker*. https://www.newyorker.com/news/news-desk/the-far-right-american-nationalist-who-tweeted-macronleaks. Accessed 1 January 2018.
15. A. Kessler (2018) Who is @TEN_GOP from the Russia indictment? Here's what we found reading 2,000 of its tweets. *CNN*. https://www.cnn.com/2018/02/16/politics/who-is-ten-gop/index.html. Accessed 1 March 2020.
16. N. Reimers, I. Gurevych (2019) Sentence-BERT: Sentence embeddings using siamese BERT-networks in *Proc. 2019 Conf. Empirical Methods in Natural Language Processing and the 9th Intl. Joint Conf. Natural Language Processing (EMNLP-IJCNLP)*. pp. 3973–3983. doi:10.18653/v1/D19-1410.
17. RT en Français (2017) «Sans moi le 7 mai», l'abstentionnisme gagne Twitter. *RT en Français*. https://francais.rt.com/france/37496-sans-moi-7-mai-abstentionnisme-gagne-twitter. Accessed 24 April 2017.
18. J. Borger (2017) US official says France warned about Russian hacking before Macron leak. *The Guardian*. https://www.theguardian.com/technology/2017/may/09/us-russians-hacking-france-election-macron-leak. Accessed 1 January 2018.
19. A. K. McCallum (2002) Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu. Accessed 1 January 2018.
20. F. Pedregosa, et al. (2011) Scikit-learn: Machine learning in Python. *J. Machine Learning Research* 12:2825–2830. arxiv:1201.0490.
21. A. Ratner, et al. (2017) Snorkel: Rapid training data creation with weak supervision in *Proc. VLDB Endowment*. Vol. 11, pp. 269–282. doi:10.14778/3157794.3157797.
22. T. P. Peixoto (2014) Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models. *Phys. Rev. E* 89:012804. doi:10.1103/PhysRevE.89.012804.
23. E. K. Kao, S. T. Smith, E. M. Airoldi (2019) Hybrid mixed-membership blockmodel for inference on realistic network interactions. *IEEE Trans. Network Science and Engineering* 6(3):336–350. doi:10.1109/TNSE.2018.2823324.
24. E. K. Kao (2017) Causal inference under network interference: A framework for experiments on social networks. *Harvard University*. arxiv:1708.08522.
25. G. W. Imbens, D. B. Rubin (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences*. (Cambridge University Press). doi:10.1017/CBO9781139025751.
26. L. Dearden (2017) Emmanuel Macron launches legal complaint over offshore account allegations spread by Marine Le Pen. *The Independent*. https://www.independent.co.uk/news/world/europe/french-presidential-election-latest-emmanuel-macron-legal-complaint-marine-le-pen-offshore-account-a7717461.html. Accessed 1 April 2020.
27. E. Ferrara (2017) Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday* 22(8–7). https://firstmonday.org/ojs/index.php/fm/article/download/8005/6516. Accessed 1 July 2020.
28. O. Varol, E. Ferrara, C. A. Davis, F. Menczer, A. Flammini (2017) Online human-bot interactions: Detection, estimation, and characterization in *Proc. 11th Intl. AAAI Conf. Web and Social Media (ICWSM 2017)*. pp. 280–289. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/viewPaper/15587. Accessed 1 January 2018.
29. B. McEwan (2020) How social media misinformation wins—even if you don't believe it. *The Week*. https://theweek.com/articles/890910/how-social-media-misinformation-wins–even-dont-believe. Accessed 1 March 2020.
30. C. A. Davis, O. Varol, E. Ferrara, A. Flammini, F. Menczer (2016) BotOrNot: A system to evaluate social bots in *Proc. 25th Intl. Conf. Companion on World Wide Web*. pp. 273–274. doi:10.1145/2872518.2889302.
31. Bot Sentinel (2020) Platform developed to detect and track political bots, trollbots, and untrustworthy accounts. https://botsentinel.com. Accessed 1 March 2020.
32. A. Rauchfleisch, J. Kaiser (2020) The false positive problem of automatic bot detection in social science research. *PLOS ONE* 15(10):1–20. doi:10.1371/journal.pone.0241045.
33. S. T. Smith, E. K. Kao, K. D. Senne, G. Bernstein, S. Philips (2014) Bayesian discovery of threat networks. *IEEE Trans. Signal Proc.* 62(20):5324–5338. doi:10.1109/TSP.2014.2336613.
34. C. Fairbanks (2020) Cassandra Fairbanks. Sputnik News. https://sputniknews.com/authors/cassandra_fairbanks. Accessed 1 March 2020.
35. J. Donati (2020) U.S. adversaries are accelerating, coordinating coronavirus disinformation, report says. *The Wall Street Journal*. https://www.wsj.com/articles/u-s-adversaries-are-accelerating-coordinating-coronavirus-disinformation-report-says-11587514724. Accessed 21 April 2020.

PNAS
www.pnas.org

Supplementary Information for

Automatic Detection of Influential Actors in Disinformation Networks

Steven T. Smith, Edward K. Kao, Erika D. Mackin, Danelle C. Shah, Olga Simek, and Donald B. Rubin

To whom correspondence may be addressed. Email: stsmith@ll.mit.edu or rubin@stat.harvard.edu. Distribution statement A. Approved for public release: distribution unlimited. This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering.

This PDF file includes:

Supporting Information Text

A. Additional Background. Exploitation of social media and digital communications by world powers to achieve their political objectives (1–18) at unprecedented scales, speeds, and reach presents a rising threat, especially to democratic societies (1, 3, 5, 6, 19–21). Situational awareness of influence campaigns and a better understanding of the mechanism behind social influence (4, 22) provide necessary capabilities for potential responses (2, 7, 23). In other applications, social influence can even be harnessed to promote knowledge and best practices in public policy settings for social good (24).

B. Narrative Detection Using Topic Modeling. Tweets were collected that are potentially relevant to a representative scenario in which actual IO accounts were expected to be active: the 2017 French election. A large dataset containing both English and French language corpora was created from this collection. From these corpora, 15 English language topics and 30 French topics are generated automatically using a topic modeling algorithm (25). The set of generated English topics includes several relevant to the French election as well as topics on U.S. politics and other world events. A selection of the English topics is shown in Table S1. The first topic relates to unsubstantiated financial allegations; this is the topic used for network discovery in the *main paper*. The generated French topics are predominantly focused on the French election, including a French language version of unsubstantiated financial allegations. A subset of these topics is shown in Table S2. Note that in both tables and the sequel, the notation ':emoji_symbol:' specifies an emoji symbol.

C. IO Account Classifier and Feature Engineering. The data used to train and test the IO classifier includes 3,151 known IO accounts released by Twitter and 15,000 randomly selected accounts from three subsets of the targeted collection dataset: accounts that tweeted on the French election in English, accounts that tweeted on the French election in French, and topic- and language-neutral accounts. Each subset contributes 5,000 accounts to the training dataset. Multiple feature categories are used: account behavior features, language features across all content, and features derived from the content itself. The origin of these accounts from Twitter's dataset is illustrated in Fig. S1.

C.1. Heuristics for Semisupervised Learning. Because the quantity of known IO accounts is relatively small, a semisupervised training strategy based on Snorkel (26) is employed. In this approach, training data from known truth are augmented with weakly labeled training data provided by heuristic "labeling functions" (Table S3), and the learning model learns and incorporates labeling function inaccuracies. Our Snorkel labeling functions are based on reported characteristics of IO account behavior, content, and metadata (27–30). These Snorkel heuristics were then refined by applying them to a small validation set. One hundred accounts were randomly selected out of the 5,000 that tweeted on the French election in English. An attempt was made to label each account as either a real person or an IO account by examining their Twitter profile and current tweets. However, many accounts were either suspended or difficult to determine with confidence to which category they belong. Of these 100 randomly chosen accounts, 24 were labeled as IO accounts, 31 were labeled non-IO, and the remaining were discarded from the validation set.

Though 9–15% of all Twitter accounts are estimated to be bots (31), it is possible that the charged topics analyzed in the *main paper* contain a higher proportion of bots and possible IO accounts. To account for this possibility in classifier training, all accounts that have a 70% or higher likelihood of being IO-like, according to Snorkel, are labeled as IO accounts, and accounts below the 70% threshold are labeled as non-IO accounts. This results in roughly 30% of the training accounts on the French election and 15% of the topic-neutral accounts being labeled IO accounts. The proportions of each training data subset that falls above the threshold is given for three different points in Table S4.

C.2. Classifier Design Comparisons. Classifier design is conducted by comparing the relative performance of four different classifier models, Random Forest (32), Logistic Regression, xgBoost (33), and SVM (34), and two dimensionality reduction approaches, Extra-Trees (35) and SVD. The performance of the all classifier design combinations over a grid of classifier parameters is evaluated via averaged cross validation over twenty $90:10$ splits. Cross validation is performed both over 10% of all training data (Fig. S3) and after discarding all accounts that were labeled as IO-like by Snorkel heuristics (Fig. S2). Of the eight combinations and for both methods of cross validation, the best performing method is composed of dimensionality reduction using the Extra-Trees method followed by a Random Forest classifier (Figs. S2, S3). We compare the P-R and ROC curves for Random Forest, Extra-Trees across the two methods of cross validation in Figs. S4 and S5, respectively. The standard deviation from cross-validation is shown in Fig. S4, in which it is seen that the maximum standard deviation for when all data is used is 1.6%, and 3.2% when Snorkel positives are omitted. Although xgBoost with Extra-Trees performs as well as the Random Forest, Extra-Trees classifier when the validation set includes 10% of all training data, it performs significantly worse when the Snorkel-labeled IO accounts are discarded from the validation set. Further, examination of the performance of xgBoost, Extra-Trees on accounts in the narrative networks shows that the classifier scores tend to be skewed towards extremes, rather than being distributed more evenly across the interval $[0, 1]$ (Fig. S8).

To assess the necessity of using Snorkel heuristics in our semisupervised learning approach, we compute classifier performance without Snorkel labeling functions. As expected in this learning problem with limited truth data, we observe that when the classifier is trained on a deterministic combination of 3,151 positive examples and 15,000 negative examples, there is strong evidence of classifier overfitting whereby the classifier simply learns the boundary between these two classes. Consequentially, the classical, supervised classifier learns previously observed IO behavior and is unable to recognize new IO accounts. This overfitting is observed both in Fig. S6 and Table S9, in which the English narrative network (*main paper*, Fig. 8) is labeled as entirely non-IO-like with the exception of the three known IO accounts within this network. Furthermore, in contrast to the balanced principal classifier features obtained using a semisupervised approach (Tables S6–S8), the features of the strictly supervised classifier are dominated by largest component of the training data (Fig. S1). For example, the most important supervised classifier feature is the Serbian language, and 8 out of the top-10 content features are related to Serbia, all corresponding to the fact that 55% of our training data is comprised of Twitter's Serbian IO dataset. We conclude from both classifier design principles and these results that a semisupervised approach is necessary to avoid overfitting in the IO classifier problem.

C.3. Sensitivity Analysis and Performance Statistics. Snorkel is used to provide additional, semisupervised training data by labeling IO positives. A Snorkel score is computed for each account, and in the *main paper*, all accounts that exceed a score of 70% are treated as IO truth within

classifier training. This 70% threshold is determined by both a sensitivity analysis that compares classifier performance across a range of metrics: precision–recall (P-R) at a fixed classifier score threshold (0.6 based on Fig. S9), area-under-the-PR-curve (AUPRC), the equal-error rate (EER, Table S5), and the fraction of training accounts that are labeled positively by Snorkel (Table S4). Performance is comparable across the three bracketed thresholds considered—50%, 70%, and 90%—though the 50% threshold has the worst performance for most of the statistics. Retaining all training data within cross-validation sets (contra omitting Snorkel positives) is most consistent with the objective of measuring classifier performance on actual accounts within a narrative network. The relative performance comparisons of Table S5 (using all data) and Table S4 show that a 70% threshold has the best AUPRC, best recall, with a small (0.5–1.5%) trade-off in EER and precision, and also the fraction of IO accounts at this threshold is consistent with published estimates of bot activity as discussed in C.1.

C.4. Feature Engineering. The IO classifier is trained on three categories of features: account behavior, languages used in tweets, and all 1- and 2-grams that appear more than 15 times across all account tweets in the training set. Initially, the feature set is composed of 17 behavioral features, 60 language features, and 1.8 million 1- and 2-grams. To limit the per account content used to generate these content features, a maximum of 10,000 randomly selected tweets are chosen from each account. The total number of tweets used in the English and French classifier is 40,155,545.

The standard machine-learning dimensionality reduction step is used to improve classifier performance in problems like this one that have a very large feature space relative to the number of training samples. Grid search optimization is used to determine the best-performing dimensionality reduction approach, and the relative importance of each feature. In the dataset used in the *main paper*, the Extra-Trees algorithm is used to reduce the feature space to 10 behavioral features (Table S6), 30 languages (Table S7), and 500 1- and 2-grams (Table S8). Additionally, the behavioral, language, and 1-and 2-gram feature spaces are each reduced independently of each other to ensure adequate representation by each feature category.

C.5. Classifier Scores versus Account Status. As noted in section C.2, the Random Forest with Extra Trees combination is the best performing classifier model. Its optimality is further justified upon examination of the distribution of the classifier scores across accounts in the French narrative network (Fig. S7). The accounts are divided by current (March 2020) status as reported by the Twitter API: active, suspended, and deleted. The suspended and deleted accounts are skewed toward higher classifier scores, showing a correlation between accounts detected by the classifier and behavior that results in suspension from Twitter.

Although both the Random Forest/Extra-Trees and xgBoost/Extra-Trees classifiers have near-identical performance in Fig. S3, the latter's performance on the French narrative network is not nearly as promising. In Fig. S8, the classifier scores over the French narrative network tend towards extremes for all three account status categories, in stark contrast to the more realistic distribution seen in Fig. S7.

C.6. Validation of Community-Based Proxy Truth. As discussed in the *main paper*, section Classifier Performance Comparisons, membership in the "Macron allegations" community of the French narrative network (*main paper*, Fig. 6) is used as a proxy for known IO accounts where independent truth is unavailable. To establish the narratives used by accounts in the Macron allegations community, topic modeling is performed on tweets from accounts in the community over the week preceding the media blackout (28 April to 5 May 2017). A selection of the generated topics is given in Table S10. Three representative tweets from each topic are shown in Table S11, which illustrate the stance of accounts within this community. Topic modeling is performed on the tweets in the pro-Macron and pro-Abstention communities over the same time period. A selection from those topics is given in Table S12 and Table S13, respectively.

To validate this proxy assumption and quantify its accuracy in the absence of the underlying truth, we hypothesize that: 1) the Macron allegations narrative is used by actual IO accounts in the 2017 French election; and 2) the distribution of known IO accounts is higher in IO narrative networks. The first hypothesis is confirmed by numerous independent news reports (14, 19, 20, 36) and direct observation (3, 16, 21, 37–40). Though we do not have the ability to independently establish the validity of the second hypothesis, we can show that the validity of our results are consistent with this hypothesis.

The distribution of our classifier scores is computed across both the "Macron allegations" and "pro-Macron"/"pro-abstention" communities in the French narrative network (Fig. S9). There is a distinct disparity between these histograms. Classifier scores of accounts in the Macron allegations community are relatively small at lower (non-IO-like) scores, and rise sharply to very high relative frequencies above a classifier score of 0.6—the expected range of IO-like scores. In contrast, classifier scores across the other communities are more evenly distributed and slightly skewed towards lower (non-IO-like) scores, also expected for narrative network communities dominated by pro-Macron activity.

D. Network Potential Outcome Framework for Causal Inference. Interference takes place in causal inference when the treatment applied to one unit affects the outcomes of other units due to their interactions and influence. An example of rising importance is treating individuals on a social network. This section introduces the mathematical framework of causal inference under network interference (41), used in the *main paper*, section Impact Estimation. Network potential outcomes are the fundamental quantity used to capture various types of causal effects such as network impact in Eq. [1] of the main paper. Realistically, many network potential outcomes will be unobserved and a complete randomization over the different treatment exposures is infeasible, making the estimation of the causal estimands challenging. Bayesian imputation of the missing network potential outcomes provides a natural conceptual solution. Theory for this imputation is developed based on the critical assumptions of unconfounded treatment assignment and an unconfounded influence network, leading to the key ignorability condition for the treatment exposure mechanism. Driven by this theory, a rigorous design and analysis procedure is proposed for causal inference under network interference.

D.1. Introduction. Interference, in the context of causal inference, refers to the situation when the outcome of a "unit" is affected not only by its own treatment, but also by the treatment of other units. Interference on a network of influence, known as network interference, is of rising importance. Common examples are experiments and observational studies on a social network where treatment effects propagate through peer influence, spread of know'ledge, or social benefits, etc. This phenomenon is also known as spillover effects and social contagions. The application areas are numerous, e.g., public health, education, and policy (42–45); social media and marketing (46–49); network security (4, 50); and economics (51–53).

Traditionally, interference has been viewed as a nuisance in causal experiments, and earlier works propose experimental designs that render the interference effects ignorable on restricted, simple block structures (54, 55). More recent work detects and estimates interference effects on networks,

many of them building on Rubin's potential outcome framework (56, 57). To detect network interference effects, Bowers et al. (58) propose hypothesis testings using potential outcome models with specific primary and peer effects, and Athey et al. (59) propose exact p-value tests by constructing artificial experiments on the original experimental units such that the null hypothesis is sharp. To estimate network interference effects, Aronow and Samii (60) propose inverse probability weighting when the probability of the specific exposure condition can be computed, Ugander et al. (61) develop a cluster randomization approach that leads to a closed-form solution on the probabilities of specific neighborhood exposures, and Sussman and Airoldi (62) propose exclusion restrictions on the potential outcomes and derive design conditions that lead to unbiased estimators. Li and Wager demonstrate practicality of non-parametric estimators of average primary and peer effects through random graph asymptotic analysis (63).

Social confounders present another source of challenge for causal inference under network interference. Early work by Manski (53) demonstrates unidentifiability of the peer effects in the presence of social confounders under linear outcome models. Recent works in causal inference show that confounding social covariates lead to unidentifiability and biased estimates of causal effects (64), especially on social networks (65, 66), and how longitudinal studies (67, 68) and design of experiment for specific peer effects (69) provide a way forward.

D.2. Definitions. The key quantities are the potential outcomes of each unit in the study under different treatment conditions. They serve as the basic building blocks for causal inference in the potential outcome framework. Most existing works assume an absence of interference and a simple binary treatment assignment, which is the Stable Unit Treatment Value Assumption (SUTVA) (70). Under SUTVA, the potential outcomes for each unit i are a function of its own treatment and are denoted as $Y_i(Z_i)$, where Z_i is a binary indicator for whether the treatment is assigned to unit i. The potential outcomes of all N units in an experiment, $\boldsymbol{Y}$, can be partitioned into two vectors of N components: $\boldsymbol{Y}(0)$ for all outcomes under control and $\boldsymbol{Y}(1)$ for all outcomes under treatment. $\boldsymbol{Y}$ can also be partitioned according to whether it is observed. In an experiment, a unit is either under control or under treatment. Therefore, half of all potential outcomes are observed, denoted as $\boldsymbol{Y}_{\text{obs}}$. The other half of the potential outcomes are unobserved, denoted as $\boldsymbol{Y}_{\text{mis}}$. As a result, causal inference is fundamentally a missing data problem, with a rich body of work on rigorous design and analysis for estimating and imputing the missing outcomes (57).

Under network interference, the potential outcome definitions need to be generalized to encompass the different treatment exposure conditions via the network, starting with the units:

Definition 1 (Finite Population on a Network of Influence)**.** *The study takes place on a finite population of N units where their influence on each other is represented as a $N \times N$ influence matrix $\boldsymbol{A}$. Each element of the influence matrix, $A_{ij} \in \mathbb{R}$, represents the strength of influence unit i has on unit j. One may visualize this population network as a graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, of which the node set $\mathcal{V}$ consists of the units of the study ($|\mathcal{V}| = N$) and the edge set $\mathcal{E}$ represents the none-zero entries of the influence matrix $\boldsymbol{A}$.*

The outcomes of each unit not only depend on its own treatment but also on exposure to treatments on other units propagated through the influence network.

Definition 2 (Network Potential Outcomes)**.** *Under network interference, the outcomes of a unit i, change according to its exposure to the treatment on the finite population $\boldsymbol{Z}$, through the network of influence $\boldsymbol{A}$. The network potential outcomes of i are denoted as $Y_i(\boldsymbol{Z}, \boldsymbol{A})$.*

Sometimes, it is clearer to denote the treatment on certain units separately. For example, one may want to denote the treatment on unit i itself. In such cases, the following notational convention is adopted: $Y_i(\boldsymbol{Z}, \boldsymbol{A}) \equiv Y_i(Z_i, \boldsymbol{Z}_{-i}, \boldsymbol{A})$, where $\boldsymbol{Z}_{-i}$ is the treatment vector excluding the ith element.

Definition 3 (Network Potential Outcome Sets)**.** *The entire set of network potential outcomes for unit i is $\mathbb{Y}_i = \{\, Y_i(\boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{A} = \boldsymbol{a}) \,\}$ for all $\boldsymbol{z} \in \mathcal{Z}$, $\boldsymbol{a} \in \mathcal{A}$, where $\mathcal{Z}$ is the set of all possible assignments on the closed neighborhood of i and $\mathcal{A}$ is the set of all possible influence networks on the closed neighborhood of i. The set of all network potential outcomes on the finite population with N units is $\mathbb{Y} = \{\, \mathbb{Y}_1, \mathbb{Y}_2, \ldots, \mathbb{Y}_N \,\}$. $\mathbb{Y}$ can also be partitioned based on whether it is observed. In an experiment, for each unit, only one of the numerous possible neighborhood treatment in $\mathcal{Z}$ is realized. The set of observed network potential outcomes, typically of size N, is denoted as $\mathbb{Y}_{\text{obs}}$. Most of the network potential outcomes will be unobserved. The set of unobserved outcomes, its size depending on the size of each unit's closed neighborhood, is denoted as $\mathbb{Y}_{\text{mis}}$.*

Lastly, covariates $\boldsymbol{X}$ on the network potential outcome units play a vital role in principled design and analysis for estimating and imputing the unobserved outcomes. For k-dimensional covariates on the units, the matrix $\boldsymbol{X}$ has $N \times k$ elements. Under the scope here, the treatment vector $\boldsymbol{Z}$ is limited to binary indicators for treatment versus control, but can be easily generalized for multi-level treatments.

D.3. Causal Estimands Using Network Potential Outcomes. The network potential outcomes serve as the building blocks for defining appropriate causal estimands to answer various causal questions under network interference. In addition to the causal estimand in *main paper*, Eq. [1], this section gives more example causal estimands, each focusing on quantifying the effect of a particular kind of exposure to treatment. Many more causal estimands may be defined using the network potential outcomes, but these demonstrate the flexibility of the network potential outcomes in expressing causal quantities under network interference.

Primary Causal Effect Estimands: If the primary treatment causal effect is the quantity of interest (i.e., want to separate it from the peer influence effects), an appropriate set of conditional causal estimands for each unit i are:

$$\xi_i(\boldsymbol{z}) \equiv Y_i(Z_i = 1, \boldsymbol{Z}_{-i} = \boldsymbol{z}, \boldsymbol{A}) - Y_i(Z_i = 0, \boldsymbol{Z}_{-i} = \boldsymbol{z}, \boldsymbol{A}) \quad [1]$$

where $\boldsymbol{z} \in \mathcal{Z}_{-i}$, is a member of the set of all possible assignments on $\boldsymbol{Z}_{-i}$. This set of conditional estimands capture the causal effect of the treatment on unit i conditioning on a particular treatment assignment $\boldsymbol{z}$ on the other units. This estimand focuses on the causal effect of receiving the treatment itself, by fixing the exposure through the influence network (i.e., the peer effect). In the absence of any exposure to treatment on peers, $\xi_i(\boldsymbol{0})$, the classical (i.e., under SUTVA) unit level causal effect, $Y_i(1) - Y_i(0)$, is recovered. If one wants to estimate the average primary treatment causal effect on unit i under all possible neighborhood treatment, the causal estimand becomes:

$$\xi_i^{\text{ave}} \equiv \frac{1}{2^{N-1}} \sum_{\boldsymbol{z} \in \mathcal{Z}_{-i}} \xi_i(\boldsymbol{z}) \quad [2]$$

For a population of size N, the average primary treatment causal effect on the population is simply:

$$\xi^{\text{ave}} \equiv \frac{1}{N} \sum_{i=1:N} \xi_i^{\text{ave}} \quad [3]$$

k **Treated Neighbor Causal Effect Estimands:** If the peer influence effect is the focus of interest, a natural quantity to consider is the causal effect of having k of the neighbors of unit i treated. Assuming unit i has at least k neighbors, an appropriate set of conditional causal estimands are:

$$\delta_{i,k}(z) \equiv \binom{|\mathcal{N}_{-i}|}{k}^{-1} \sum_{\boldsymbol{z} \in \mathcal{Z}^k} Y_i(Z_i = z, \boldsymbol{Z}_{-i} = \boldsymbol{z}, \boldsymbol{A}) - Y_i(Z_i = z, \boldsymbol{Z}_{-i} = \boldsymbol{0}, \boldsymbol{A}) \quad [4]$$

where $z \in \{0, 1\}$ are the possible assignments on unit i, $\mathcal{N}_{-i}$ the open neighborhood of i, and $\mathcal{Z}^k$ the set of all treatment assignments where exactly k of i's neighbors are treated (i.e., $\sum \boldsymbol{Z}_{\mathcal{N}_{-i}} = k$). This set of conditional estimands capture the average causal effect on unit i for having k of its neighbors treated, while conditioning the treatment assignment z on i itself. Similar to the previous example, the k treated neighbor causal effect averaged over unit i's own treatment can be expressed in the following estimand:

$$\delta_{i,k}^{\text{ave}} \equiv \frac{1}{2} \sum_{z=\{0,1\}} \delta_{i,k}(z) \quad [5]$$

The population here is a bit more nuanced, because not all units have at least k neighbors and therefore can not possibly receive such peer treatment. Therefore, the population average effect should only be averaged over the units that have at least k neighbors. Defining $\mathcal{V}_{\geq k}$ to be the set of units with at least k neighbors, the average k treated neighbor causal effect on the population is:

$$\delta_k^{\text{ave}} \equiv \frac{1}{|\mathcal{V}_{\geq k}|} \sum_{i \in \mathcal{V}_{\geq k}} \delta_{i,k}^{\text{ave}} \quad [6]$$

The idea of capturing causal peer effects based on the number of neighbors being treated has been proposed by other work. Ugander et al. (61) define a similar condition called the "absolute k-neighborhood exposure" where unit i meets this neighborhood treatment condition if i is treated and at least k of i's neighbors are treated. Adopting Ugander et al.'s peer treatment condition gives the following estimand for unit i:

$$\tilde{\delta}_{i,k} \equiv \left[\sum_{l=k}^{|\mathcal{N}_{-i}|} \binom{|\mathcal{N}_{-i}|}{l} \right]^{-1} \sum_{l=k}^{|\mathcal{N}_{-i}|} \sum_{\boldsymbol{z} \in \mathcal{Z}^l} Y_i(Z_i = 1, \boldsymbol{Z}_{-i} = \boldsymbol{z}, \boldsymbol{A}) - Y_i(Z_i = 1, \boldsymbol{Z}_{-i} = \boldsymbol{0}, \boldsymbol{A}) \quad [7]$$

One may want to know the causal effect of having a certain fraction of the neighbors treated (e.g., 30% of the neighbors treated), instead of the absolute number of neighbors. Ugander et al. (61) define a version of such treatment condition called the "fractional q-neighborhood exposure". A causal estimand for fractional neighborhood treatment can be defined by simply mapping the fractional criterion to an absolute number for each unit i. For example, a q% neighborhood treatment for unit i could map to a k neighbor treatment with $k = \lceil \frac{q}{100} |\mathcal{N}_{-i}| \rceil$. In any case, the individual k treated neighbor causal estimand in equation (4) serves as the basic building block for these types of neighborhood treatment causal estimands.

Influence Network Manipulation Causal Estimands: Sometimes, one may be able to manipulate the influence network to achieve the desired outcome. Causal effects of network manipulation can be expressed using network potential outcomes as building blocks. This may seem counterintuitive at first because the influence network itself is not a "treatment". However, under network interference, the influence network leads to exposure to treatment on peers, so manipulating the influence network alters the "social treatment". The influence network and the treatment assignment together can be viewed as an assignment of "social treatment". Consider the causal estimand below on the average total effect of manipulating the influence network from $\boldsymbol{A}$ to $\boldsymbol{A}'$, given a particular treatment assignment $\boldsymbol{z}$,:

$$\zeta_{\boldsymbol{A}}(\boldsymbol{z}) \equiv \frac{1}{N} \sum_{i \in 1:N} Y_i(\boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{A}') - Y_i(\boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{A}) \quad [8]$$

This estimand may quantify the effect of weakening the disinformation network through account suspension and warning. This estimand highlights the flexibility and expressiveness of the network potential outcomes as the basic building block for causal inference under network interference.

D.4. Bayesian Imputation of Missing Outcomes and Unconfoundedness Assumptions. Under network interference, most of the potential outcomes in the causal estimands will be unobserved. Furthermore, a complete randomized assignment of the different exposures to neighborhood treatment is typically infeasible, as the structure of the influence network impacts each unit's chance to receive a certain exposure (71). These make the estimation of causal estimands challenging under network interference. A natural solution is to perform Bayesian imputation of missing potential outcomes. Also known as the Bayesian predictive inference for causal effects, this method has been well established for causal inference in the absence of interference (i.e., SUTVA holds) (72, 73).

As the potential outcomes serve as the building blocks for each causal estimand, computing the posterior distribution of the unobserved potential outcomes also gives the posterior distribution of any causal estimands. In the regular case under SUTVA, the posterior distribution of interest is $P(\boldsymbol{Y}_{\text{mis}} | \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y}_{\text{obs}})$. Inference of the unobserved potential outcomes from the observed potential outcomes, the unit covariates $\boldsymbol{X}$, and the treatment assignment vector $\boldsymbol{Z}$, is typically done with a potential outcome model. Modeling and inference of this posterior distribution is greatly simplified when the treatment assignment mechanism can be ignored (i.e., $\boldsymbol{Z}$ can be dropped from the posterior). Rubin shows how the unconfounded treatment assignment assumption leads to this ignorability (72, 73).

In the case under network interference, the posterior distribution is on the expanded sets of network potential outcomes and includes the influence network: $P(\mathbb{Y}_{\text{mis}}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{A}, \mathbb{Y}_{\text{obs}})$. Similar to the regular case under SUTVA, this posterior can also be greatly simplified when the neighborhood treatment mechanism can be ignored (i.e., both $\boldsymbol{Z}$ and $\boldsymbol{A}$ can be dropped from the posterior). This section shows how the assumptions of unconfounded treatment assignment and unconfounded influence network lead to this more extended ignorability. Under network interference, although Bayesian imputation of the missing network potential outcomes offers a practical solution, the procedure needs to respect the key unconfoundedness assumptions in order to avoid incorrect causal estimates.

Assumption 1 (Unconfounded Treatment Assignment Assumption Under Network Interference)**.** *Conditional on the relevant unit covariates $\boldsymbol{X}$ and the influence network $\boldsymbol{A}$, the treatment assignment $\boldsymbol{Z}$ is independent from the potential outcomes $\mathbb{Y}$:*

$$P(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{A}, \mathbb{Y}) = P(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{A}). \quad [9]$$

This assumption can be met in real-world experiments through complete randomization of $\boldsymbol{Z}$ or including possible confounders in the conditional unit covariates $\boldsymbol{X}$. Sometimes, experiments on a social network target units with certain network characteristics for treatment, in order to achieve a desirable overall outcome. For example, a researcher may target the most influential units (e.g., high degree nodes) in order to maximize peer influence effects. Assumption 1 holds under such treatment assignment strategies because the treatment $\boldsymbol{Z}$ only depends on the influence network, $\boldsymbol{A}$, and relevant unit covariates $\boldsymbol{X}$. This paper estimates the impact of each unit as a potential source (i.e., being treated), so the treatment assignment does not depend on the potential outcomes, satisfying Assumption 1.

Assumption 2 (Unconfounded Influence Network Assumption Under Network Interference)**.** *Conditional on the relevant unit covariates $\boldsymbol{X}$, the influence network $\boldsymbol{A}$ is independent from the potential outcomes $\mathbb{Y}$:*

$$P(\boldsymbol{A}|\boldsymbol{X}, \mathbb{Y}) = P(\boldsymbol{A}|\boldsymbol{X}). \quad [10]$$

This assumption is met if the formation of the influence network has no correlation with the potential outcomes. However, this is often not true in real-world experiments (66). Intuitively, correlation between the potential outcomes and the influence network may arise from certain characteristics of a unit that are correlated with both its outcomes as well as its relationships with other units on the network. Activity level and group memberships are examples of such characteristics. For example, an account's node degree on the retweet network may be positively correlated with its potential outcomes (i.e., tweet count on the IO narrative) because hubs in the influence network may tweet more in general. Similarly, an account's membership to an IO community may be positively correlated with its potential outcomes. Therefore, meeting Assumption 2 will likely require including such confounding characteristics in the conditional unit covariates $\boldsymbol{X}$. A method to achieve this will be formally proposed in Theorem 2, but first, we introduce the ignorability condition.

Theorem 1 (Ignorable Treatment Exposure Mechanism Under Network Interference)**.** *If the unconfounded treatment assignment assumption and the unconfounded influence network assumption are both met, the treatment exposure mechanism is ignorable and does not enter the posterior distribution of the missing potential outcomes:*

$$P(\mathbb{Y}_{\text{mis}}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{A}, \mathbb{Y}_{\text{obs}}) = P(\mathbb{Y}_{\text{mis}}|\boldsymbol{X}, \mathbb{Y}_{\text{obs}}). \quad [11]$$

Theorem 1 is proved via factorization and application of Assumptions 1 and 2:

$$\begin{aligned}
P(\mathbb{Y}_{\text{mis}}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{A}, \mathbb{Y}_{\text{obs}}) &= \frac{P(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{A}, \mathbb{Y})P(\mathbb{Y}|\boldsymbol{X}, \boldsymbol{A})}{\int P(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{A}, \mathbb{Y})P(\mathbb{Y}|\boldsymbol{X}, \boldsymbol{A})\, d\mathbb{Y}_{\text{mis}}} \\
&= \frac{P(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{A})P(\mathbb{Y}|\boldsymbol{X}, \boldsymbol{A})}{P(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{A}) \int P(\mathbb{Y}|\boldsymbol{X}, \boldsymbol{A})\, d\mathbb{Y}_{\text{mis}}} \\
&= \frac{P(\mathbb{Y}|\boldsymbol{X}, \boldsymbol{A})}{\int P(\mathbb{Y}|\boldsymbol{X}, \boldsymbol{A})\, d\mathbb{Y}_{\text{mis}}} \\
&= P(\mathbb{Y}_{\text{mis}}|\boldsymbol{X}, \boldsymbol{A}, \mathbb{Y}_{\text{obs}}) \\
&= \frac{P(\boldsymbol{A}|\boldsymbol{X}, \mathbb{Y})P(\mathbb{Y}|\boldsymbol{X})}{\int P(\boldsymbol{A}|\boldsymbol{X}, \mathbb{Y})P(\mathbb{Y}|\boldsymbol{X})\, d\mathbb{Y}_{\text{mis}}} \\
&= \frac{P(\boldsymbol{A}|\boldsymbol{X})P(\mathbb{Y}|\boldsymbol{X})}{P(\boldsymbol{A}|\boldsymbol{X}) \int P(\mathbb{Y}|\boldsymbol{X})\, d\mathbb{Y}_{\text{mis}}} \\
&= \frac{P(\mathbb{Y}|\boldsymbol{X})}{\int P(\mathbb{Y}|\boldsymbol{X})\, d\mathbb{Y}_{\text{mis}}} \\
&= P(\mathbb{Y}_{\text{mis}}|\boldsymbol{X}, \mathbb{Y}_{\text{obs}}). \quad [12]
\end{aligned}$$

Finally, Assumption 2 can be satisfied by conditioning on network parameters through parametric modeling.

Theorem 2 (Unconfounded Influence Network by Conditioning on Network Parameters)**.** *The unconfounded influence network assumption in Assumption 2, $P(\boldsymbol{A}|\boldsymbol{X}, \mathbb{Y}) = P(\boldsymbol{A}|\boldsymbol{X})$, is met if:*

1. *The distribution of the influence network $\boldsymbol{A}$ can be characterized by a model H_G with nodal parameters $\boldsymbol{X}_G$ and population parameters $\boldsymbol{\Theta}_G$: $\boldsymbol{A} \sim H_G(\boldsymbol{X}_G, \boldsymbol{\Theta}_G)$;*

2. *The potential outcomes* $\mathbb{Y}$ *correlate with the influence network* $\boldsymbol{A}$ *only through a subset of the nodal parameters* $\tilde{\boldsymbol{X}}_G \in \boldsymbol{X}_G$ *and population parameters* $\tilde{\boldsymbol{\Theta}}_G \in \boldsymbol{\Theta}_G$*;*
3. *The unit covariates* $\boldsymbol{X}$ *contain these network parameters* $\tilde{\boldsymbol{X}}_G$ *and* $\tilde{\boldsymbol{\Theta}}_G$*.*

Theorem 2 leverages the assumption that the influence network $\boldsymbol{A}$ can be characterized by a model with nodal (i.e., unit-specific) parameters $\boldsymbol{X}_G$ (e.g., expected degree, community membership, position in the latent space, etc.) and population parameters $\boldsymbol{\Theta}_G$ (e.g., inter-community interaction, sparsity, etc.). Most if not all of the currently well-known network models in the network inference community can be described in this way, including the latent space models (74), the latent class models such as the membership blockmodels (75–77), degree distribution models (78, 79), and the graphon (80). Typically, only a subset of the nodal parameters, $\tilde{\boldsymbol{X}}_G$, may be correlated with the potential outcomes, like the unit-specific characteristics such as activity level and community membership in the IO narrative influence network. Some of the population parameters, $\tilde{\boldsymbol{\Theta}}_G$, like sparsity, may be correlated with the potential outcomes as well. Intuitively, conditioning on these parameters breaks the correlation between the potential outcomes and the influence network, therefore meeting the unconfounded influence network assumption. Often, these network model parameters are not readily observed, but they can be estimated from the social network data collected in the experiment. This is similar in spirit to work by Frangakis and Rubin (81) on latent ignorability where the treatment mechanism is ignorable by conditioning on the latent compliance covariate.

D.5. Theory to Practice: Design and Analysis Under Network Interference. Driven by the theoretical framework developed above, the experimental and analytical procedure for Bayesian imputation of missing potential outcomes and causal estimands is summarized in the following steps.

1. Define the population, the treatment, and the network potential outcomes. Collect prior information on the underlying influence network. This can be from data on interactions between the units (e.g., emails, tweets, phone calls, etc.) or a survey on the social network (e.g., list of relationships). One may only have partial or prior information on the influence network, in which case the network will need to be imputed during analysis.
2. Propose an appropriate network model, such as the ones mentioned in the previous section (74–80), and estimate the model parameters using the observed influence network or the prior distribution on the influence network.
3. Propose an appropriate potential outcome model including all the possible confounding covariates, guided by Assumptions 1 and 2, and Theorem 2, in order to meet the ignorable treatment exposure mechanism condition specified in Theorem 1. Some of these may be the estimated network parameters from the previous step and the possible treatment assignment confounders. Perform Bayesian inference to compute the posterior distribution of the potential outcome model parameters, jointly with the influence network if it is not fully known, as typically is the case. Weakly informative priors on the model parameters have shown to improve convergence stability while minimizing any bias on the posterior distribution (82). Some model parameters such as propagated effects on the network (each γ_k in *main paper*, Eq. [2]) should respect social phenomenologies such as decaying exposure effects with each additional hop in the propagation. This can be accomplished with truncated priors to restrict the feasible parameter range. Lastly, adequacy of the potential outcome model in describing the observed outcomes can be evaluated via statistical tests such as the posterior predictive check (83).
4. Using the potential outcome model and the parameter posterior distribution from the step above, impute the missing network potential outcomes in the causal estimands of interest. This will finally provide estimates on the desired causal estimands. Typically, one would want to quantify the uncertainty on the estimated causal estimand. This can be accomplished by multiple imputation (84) of the missing potential outcomes, by imputing them with independent samples of model parameters from their posterior distribution.

This procedure accounts for potential confounders through covariate adjustment, where accuracy depends on the adequacy of the potential outcome model. Additional robustness to model mis-specification can be achieved through balancing of confounding covariates across different treatment exposure conditions in the causal estimand. This can be done via treatment design in experiments or matching in observational studies, as a desirable future expansion on this framework. Propensity score matching for multiple treatment conditions such as the case here with numerous treatment exposure conditions through the network is a challenging and relevant research topic with recent work by Forastiere et al. (85) and Han and Rubin (86).

E. Impact Estimation on the #MacronLeaks Narrative. Further support of this framework's efficacy is provided by its application to a well-known, highly visible, and distinctive IO narrative network defined by a simple Twitter hashtag, #MacronLeaks (36, 40, 87) (Fig. S10). Vertex color in Fig. S10 indicates the number of times each account tweets the hashtag #MacronLeaks; vertex size is the in-degree, i.e., the number of retweets received by each account. Compared to the financial allegation narrative detected via topic modeling in the *main paper*, this narrative is on a related but more focused event of the leaking of candidate Macron's emails. While many IO narratives do not align nicely with a hashtag and need to be detected via topic modeling, the #MacronLeaks narrative is an instance where the hashtag became a prominent signature for the narrative. The hashtag was widely used in many tweets and retweets, as reflected by the higher magnitude impact statistics in Table S14, compared to Table 1 in the main paper. Among the accounts with high estimated impact, there exists independent confirmation of the prominence of the two accounts @wikileaks and @JackPosobiec in pushing the #MacronLeaks narrative (36, 87). The relatively high causal impact of these accounts shown in Table S14 is consistent with independently reported articles about this narrative. Further, a new finding is the high-impact spreading of this IO narrative by the known IO account @Pamela_Moore13 (37–39).

Comparison between impact statistics in Table S14 highlights the advantage of causal impact estimation in quantifying impact over activity- and topologically based statistics such as retweet (RT) volume and PageRank centrality (88, 89). The limitation of activity count statistics as indicators for impact is seen in @UserA and @UserC, both having high tweet and retweet counts but little impact due to their positions on and connectivities to the network with low centrality. Follower count and PageRank centrality clearly highlight @wikileaks's impact on the network, which is consistent with its high causal impact. However, accounts such as @JackPosobiec, @UserB, and known IO account @Pamela_Moore13 have only a medium level of PageRank centrality, but high causal impact. @JackPosobiec, being one of the earliest participant, has been reported as a key source in

pushing the #MacronLeaks narrative (36, 87). @UserB serves as a bridge into the predominantly French-speaking subgraph (the cluster seen in the middle of Fig. S10). Causal impact is able to detect these prominent accounts that do not stand out in other impact statistics by modeling the narrative propagation on the network. Unlike existing propagation methods on network topology (90) alone, causal inference accounts also for the observed outcomes at each node.

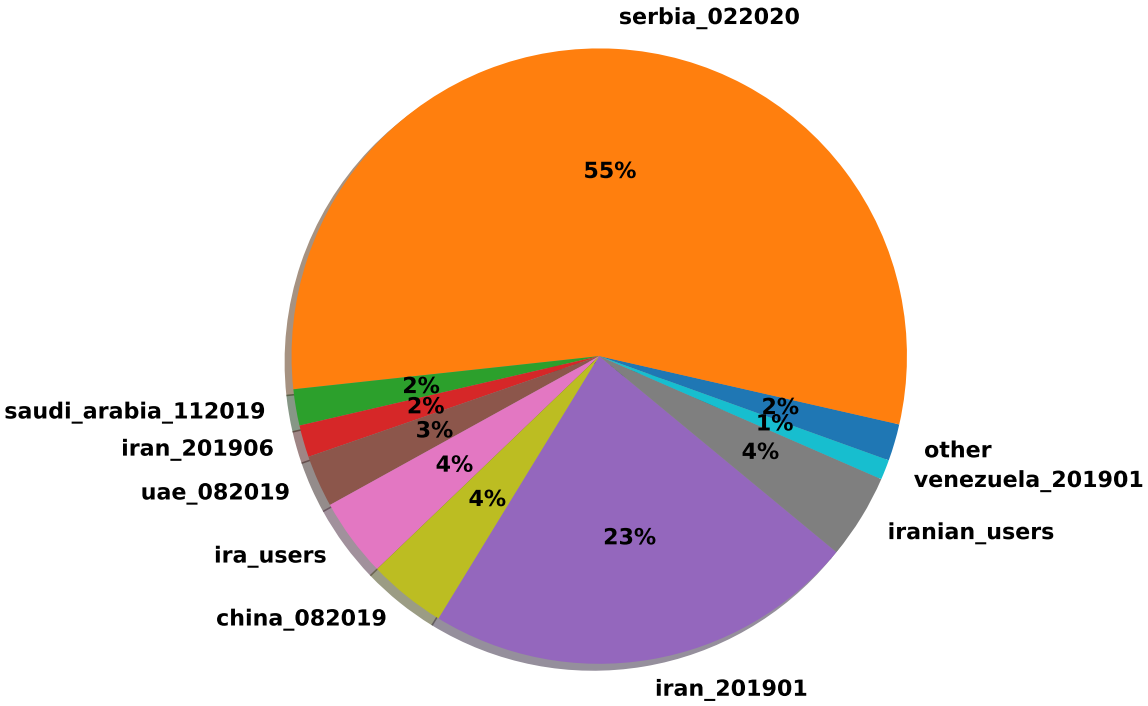
F. Influential IO-like Content 2017 vs. 2020. Many of the high-impact accounts with behaviors and content similar to known IO accounts (upper-right corner of *main paper*, Fig. 9) have been suspended by Twitter since 2017. However, several remain actively engaged at the time of writing in narratives also used by IO accounts in 2020 (12, 91, 92). For example, high-impact, IO-like accounts that posted on the 2017 French election narrative network (*main paper*, Figs. 2 and 6) are actively posting three years later on COVID-19 conspiracy theories. Additional validation of this paper's approach is suggested by examining how the content of such active accounts has been used by known IO accounts. Inspecting a small-but-representative sample of 3 active accounts appearing in the upper-right corner of *main paper*, Fig. 9 shows that content from these accounts has been used by known IO accounts hundreds of times (9 in one case, 472 in another, and 589 in another) for each of these active accounts (39). Furthermore, we observe that our classifier is robust to whether or not these accounts are directly retweeted by known IO accounts. Only 5 tweets from these 3 representative accounts (1, 1, and 3 tweets for each account) were included in our classifier as "positive" examples used by known IO accounts, whereas 555 tweets (total for all 3) were retweeted by accounts included as negative examples in the training data. Because the classifier training data is comprised of 40 million tweets, the presence of 5 tweets from these specific accounts does not have a large effect on their positive classification, especially that their content appears 555 times in non-IO "negative" training data. Finally, note that we do not assert or imply that these are IO accounts, but merely observe that their content and behavior is quantifiably similar to known IO accounts, independent of their authenticity. This observation is consistent with the fact that content from these specific accounts has also been used hundreds of times (as retweets) by known IO accounts (39, 93).

Table S1. Select English topics.

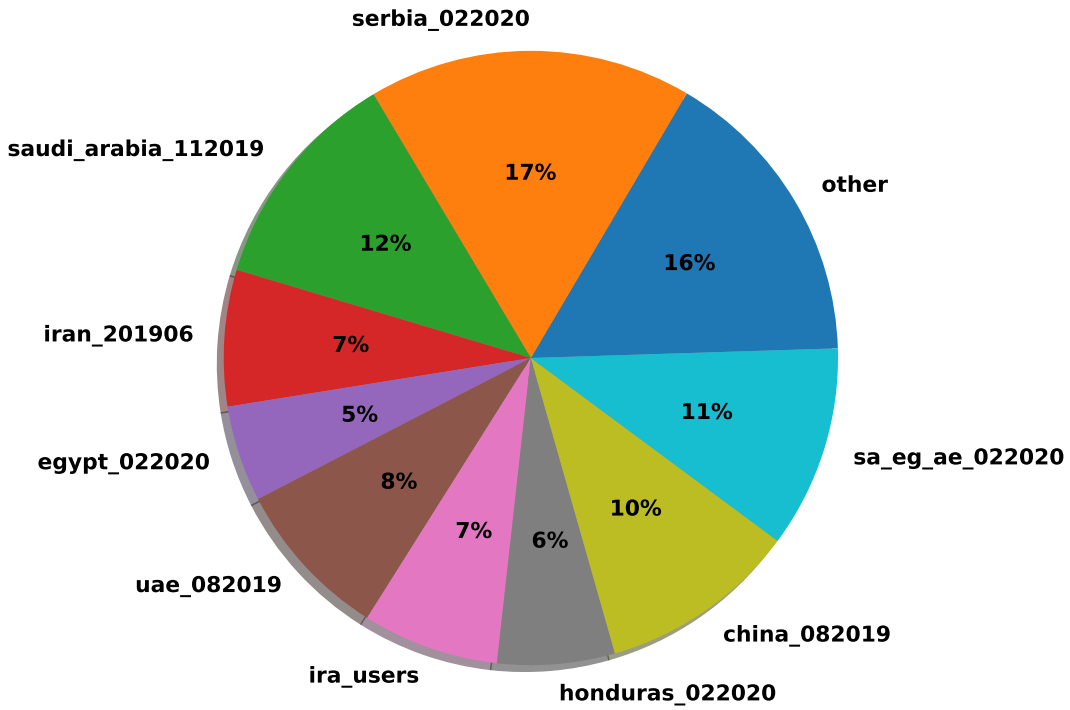
Topic 1 *Anti-Macron, financial allegations*	**Topic 2** *May 1st protest in Paris*	**Topic 3** *Brexit*	**Topic 4** *Russia and Middle East*
macron	police	brexit	syria
tax	antifa	news	russia
documents	paris	win	sptnkne.ws
emmanuel	day	amp	syrian
evasion	mayday2017	twitter.com	isis
engaging	on.rt.com	britain	russian
putin	protesters	brussels	turkey
trump	breaking	theresa	sputniknews.com
huge	video	ukip	military
disobedientmedia.com	anti	labour	amp
on.rt.com	amp	macron	attack
prove	arrested	election	army
busted	officers	europe	war
news	news	pen	israel
cheat	violent	on.rt.com	news
rally	violence	ge2017	putin
breaking	texas	telegraph.co.uk	killed
phone	portland	juncker	on.rt.com
harrisburg	left	:united_kingdom:	ukraine
nato	pen	politics	nato

Table S2. Select French topics.

Topic 1 *Anti-Macron, financial allegations*	**Topic 2** *Final debate of 2017 election*	**Topic 3** *Police violence at May 1st protests*	**Topics 4** *Islamophobia during election*
macron	macron	paris	macron
compte	débat	crs	uoif
journaliste	2017ledébat	policiers	youtube.com
fiscale	marine	gauche	watch
plainte	pen	mai	islamistes
offshore	debat2017	1er	voter
:red_circle:	soir	blessés	soutenu
mort	france	1ermai	jamaismacron
documents	plateau	extrême	france
macrongate	menace	policier	musulmans
emmanuel	quitter	macron	islamiste
porte	2017ledébat	brûlé	soutien
twitter.com	debat	police	:france:
bahamas	brigitte	france	emmanuel
évasion	mlp	:france:	jamais
preuves	lepen	molotov	:thumbs_down:
france	tours	violences	juifs
pen	faire	théo	frères
fraude	bout	ordre	lyon
société	heure	cgt	2017lédebat



(A)



(B)

Fig. S1. (A) Fraction of our 3,151 accounts by origin from Twitter's datasets (39) used for classifier training and testing. (B) Account fraction by origin of all released Twitter datasets by March 2020.

Table S3. IO account heuristics (27–30) as Snorkel labeling functions.

```
# Account has no profile description
@labeling_function()
def profile_length(x):
    return IO if x.profile_length==0 else ABSTAIN

# Account has repeated interactions with external news sites
@labeling_function()
def external_news__interactions(x):
    return IO if x.num_external__news_interactions > 5 else ABSTAIN

# Account follows large number of accounts
@labeling_function()
def num_following(x):
    return IO if x.following_count > 3000 else ABSTAIN

# On average, every tweet includes a link
@labeling_function()
def num_links(x):
    return IO if x.avg_num_links > 1 else ABSTAIN

# Account tweeted in many languages
@labeling_function()
def many_langs(x):
    return IO if np.count_nonzero(x.num_langs_used) > 10 else ABSTAIN

# Account rarely favorited tweets
@labeling_function()
def few_faves(x):
    return IO if x.num_faves <20 else ABSTAIN

# Account favorited large number of tweets
@labeling_function()
def too_many_faves(x):
    return IO if x.num_faves >30000 else ABSTAIN

#Account tweeted in an undetermined language often
@labeling_function()
def many_und_tweets(x):
    return IO if x.und > 0.05 else ABSTAIN

# Account has follow, following counts indicative of real user
@labeling_function()
def normal_people_ff_ratio(x):
    return REAL if x.follower_count < 500 and 0.75 < x.followers_following_ratio < 4 else ABSTAIN

# Account interacted with external news source once or never
@labeling_function()
def no_external_news__interactions(x):
    return REAL if x.num_news__news_interactions < 2 else ABSTAIN

#Account seldom included links in tweets
@labeling_function()
def few_tweets_w_links(x):
    return REAL if 0.05 < x.ratio_tweets_w_links_all_tweets < 0.15 else ABSTAIN

# Number of likes by account in normal range
@labeling_function()
def normal_num_likes(x):
    return REAL if 500 < x.num_faves < 10000 else ABSTAIN

# Profile description of a normal length
@labeling_function()
def normal_profile_len(x):
    return REAL if x.profile_length > 50 else ABSTAIN

# Many legitimate organizations have large number of followers, don't want to classify them as IO~accounts
@labeling_function()
def org_num_followers(x):
    return REAL if x.follower_count > 60000 else ABSTAIN
```

Table S4. Proportion of training data subsets labeled as IO accounts by Snorkel heuristics.

	50% threshold	70% threshold	90% threshold
French election, English	38%	32%	23%
French election, French	34%	28%	22%
Topic and language neutral	30%	15%	8%

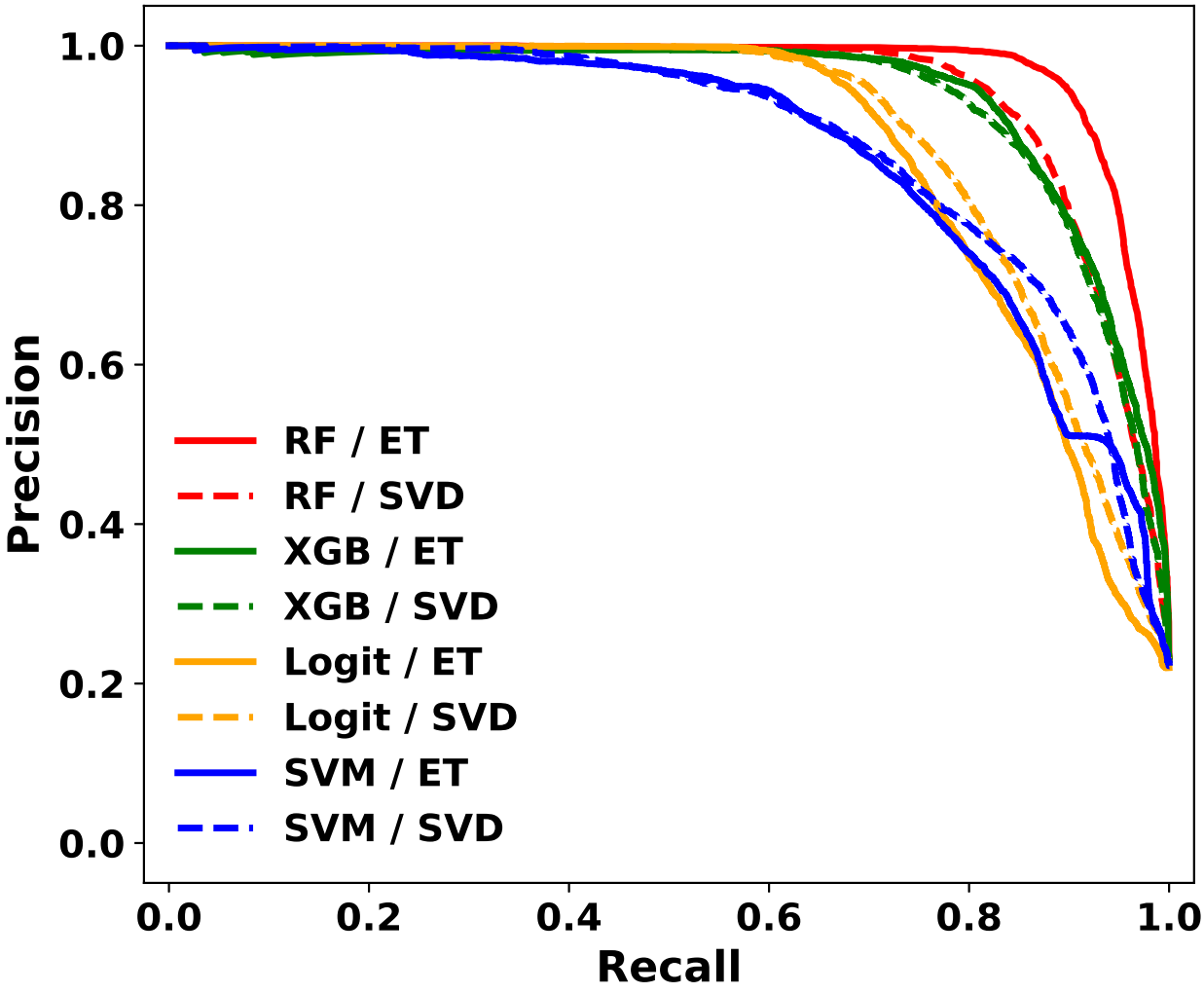


Fig. S2. Classifier P-R performance across many classifier algorithms. Validation set selected from known IO accounts and negatively labeled Snorkel data.

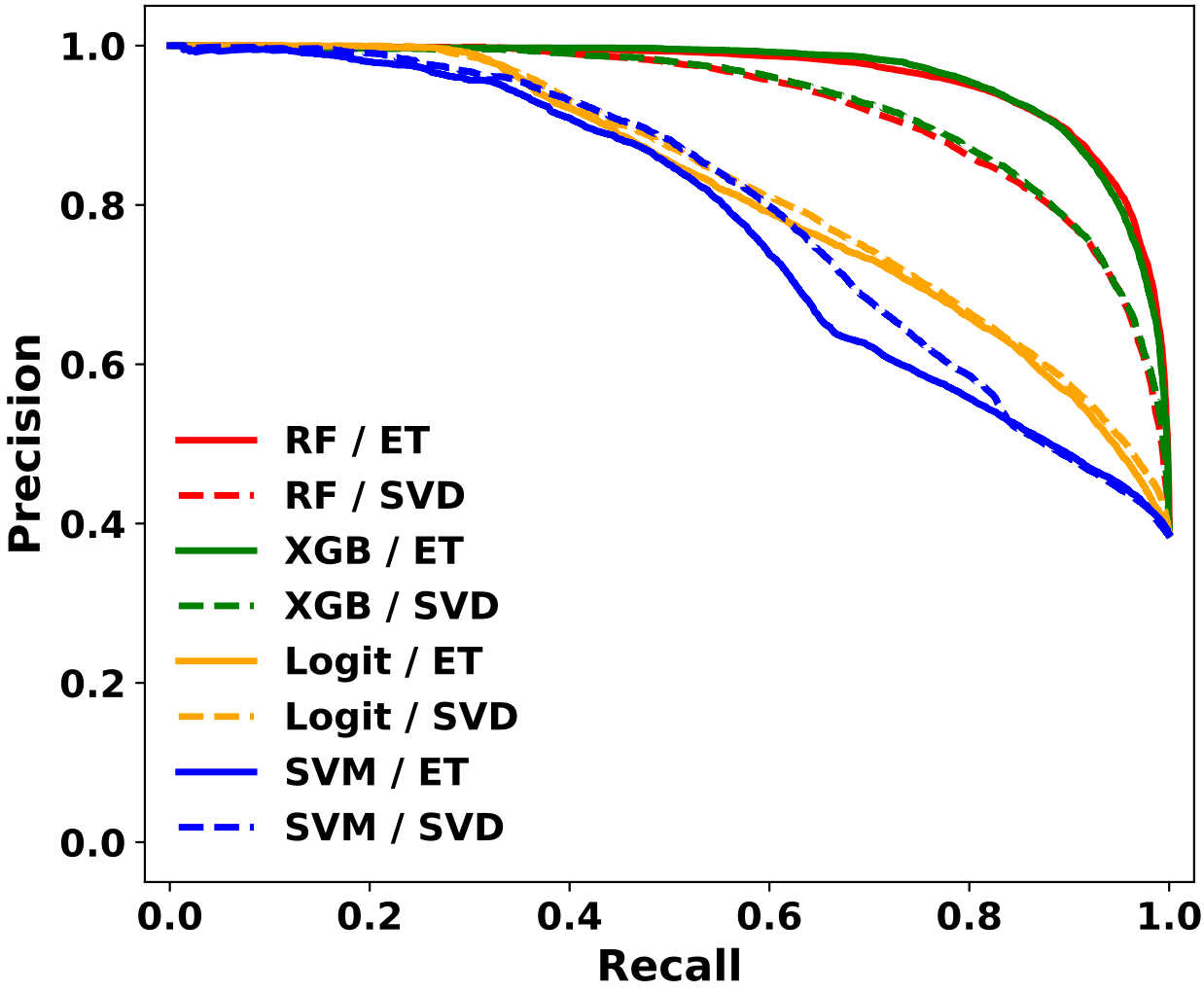


Fig. S3. Classifier P-R performance across many classifier algorithms. Validation set selected from all data.

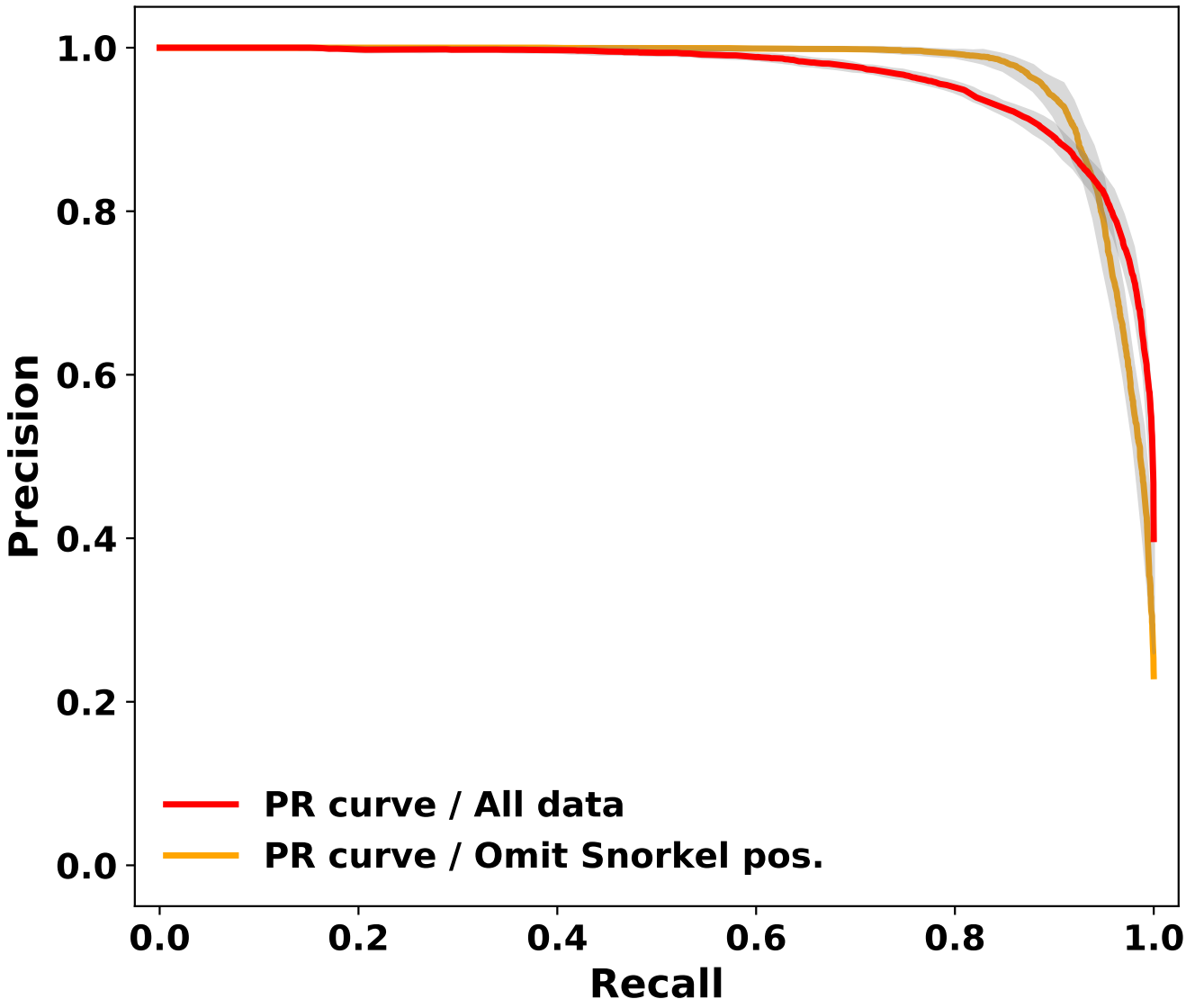


Fig. S4. RF / ET classifier precision–recall (P-R) performance with cross-validation error over twenty 90 : 10 splits; all data ( ), and Snorkel positives omitted ( ). Maximum standard deviation (gray region,) is 0.016 and 0.032, respectively.

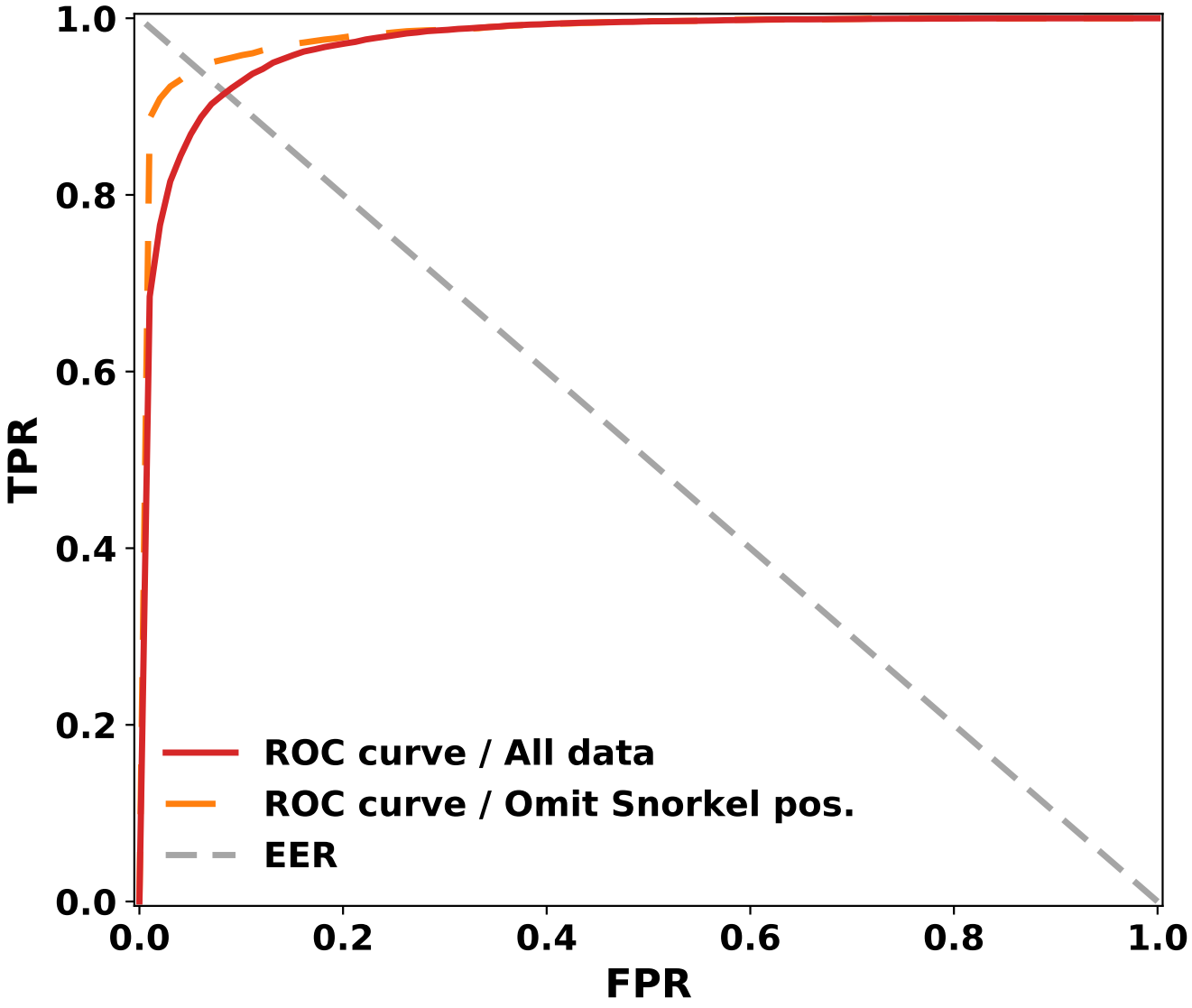


Fig. S5. RF / ET classifier receiver operating characteristic (ROC) performance: true positive rate (TPR, a.k.a. recall) versus false positive rate (FPR); all data (—), and Snorkel positives omitted (—). Classifier EER is determined by the point at which $1 - \text{TPR} = \text{FPR}$ (dashed gray curve, —).

Table S5. Sensitivity comparison, Snorkel thresholds for IO labeling and cross validation set selection.

	Precision	Recall	AUPRC	EER
50% threshold, all data	94.3%	78.7%	95.8%	10.5%
50% threshold, omit Snorkel positives	88.2%	91.6%	96.9%	6.1%
70% threshold, all data	95.6%	78.9%	96.4%	8.1%
70% threshold, omit Snorkel positives	91.6%	90.3%	96.6%	5.6%
90% threshold, all data	97.1%	73.5%	96.0%	7.6%
90% threshold, omit Snorkel positives	95.8%	90.5%	97.4%	5.0%

Table S6. Behavioral features in descending importance.

num_external_news_interactions	num_tweets_in_time_range	follower_count
avg_num_chars	sd_num_tweets_per_day	profile_length
avg_num_hashtag_chars	ratio_retweets_w_links_all_tweets	
num_faves	avg_num_tweets_per_day	

Table S7. Language features in descending importance (Twitter language codes).

en	de	ht	nl	pl	eu
sr	fr	tl	cs	no	ca
it	sl	ro	da	lt	fi
und	pt	ru	sv	lv	hu
in	es	et	tr	cy	hi

Table S8. Top 150 1- and 2-grams* (out of 500) in descending order of importance.

en:serbia	en:serbia␣austria	en:investigation
en:macron	en:vučić	en:nursultan
fr:lepen	en:aleksandar␣vučić	en:friends
en:france	en:serbian␣president	en:18
fr:macron	en:campaign	en:jeff␣sessions
en:president␣serbia	en:mueller	en:nursultan␣nazarbayev
en:lepen	en:country	en:smarttraffic
fr:france	en:government	en:jeff
en:belgrade	en:germany	en:robert␣mueller
fr:pen	en:sessions	en:foreign
en:serbian	en:attack	en:euro␣atlantic
en:serbia␣amp	en:israel	en:follow
en:aleksandar	en:meeting	en:terrorist
fr:marine	en:presidential	en:beautiful
en:election	en:merkel	en:girl
en:forum	en:aleksandar␣vucic	en:palace␣serbia
fr:fillon	en:nazarbayev␣55qsospayr	en:close
en:le	fr:merkel	en:congress
en:president	en:strike	en:presidentielle2017
fr:emmanuel	en:night	en:probe
fr:emmanuel␣macron	en:exercise␣organised	en:organised␣nato
en:french	en:55qsospayr	en:countries
en::serbia:	en:report	en:firing
en:trump	fr:lepen␣macron	en:arrested
en:amp	en:team	en:china
en:le␣pen	en:american	en:press
en:pen	en:french␣election	en:missile
en:vote	en:fbi	en:khashoggi
en:people	en:life	en:visit
en:2	en:emmanuel	en:presidential␣election
en:discussed	en:watch	en:syrian
fr:faut	en:emmanuel␣macron	en:nazarbayev
en:video	en:leaks	en:congress␣participants
en:day	en:war	en:minister
fr:macron␣lepen	en:hosting␣largest	en:heart
en:vucic	en:special	en:proof
en:win	en:nato␣euro	en:citizens
en:business␣forum	en:participants␣president	en:prince
en:austria	en:photo␣congress	en:share
fr:parti	en:atlantic␣di	en:crisis
en:russian	en:family	en:family␣photo
en:breaking	en:iran	en:president␣nursultan
en:syria	en:killed	en:turkey
en:obama	en:children	en:business
en:news	en:emails	en:participants
en:love	en:largest␣disaster	en:happening
en:support	en:kosovo	en:disaster␣response
en:live	en:robert	en:including
en:eu	en:saudi	en:france␣macron
en:amp␣hosting	en:attacks	fr:cqfd

*2-grams are represented by two words separated by the visible space character '␣'.

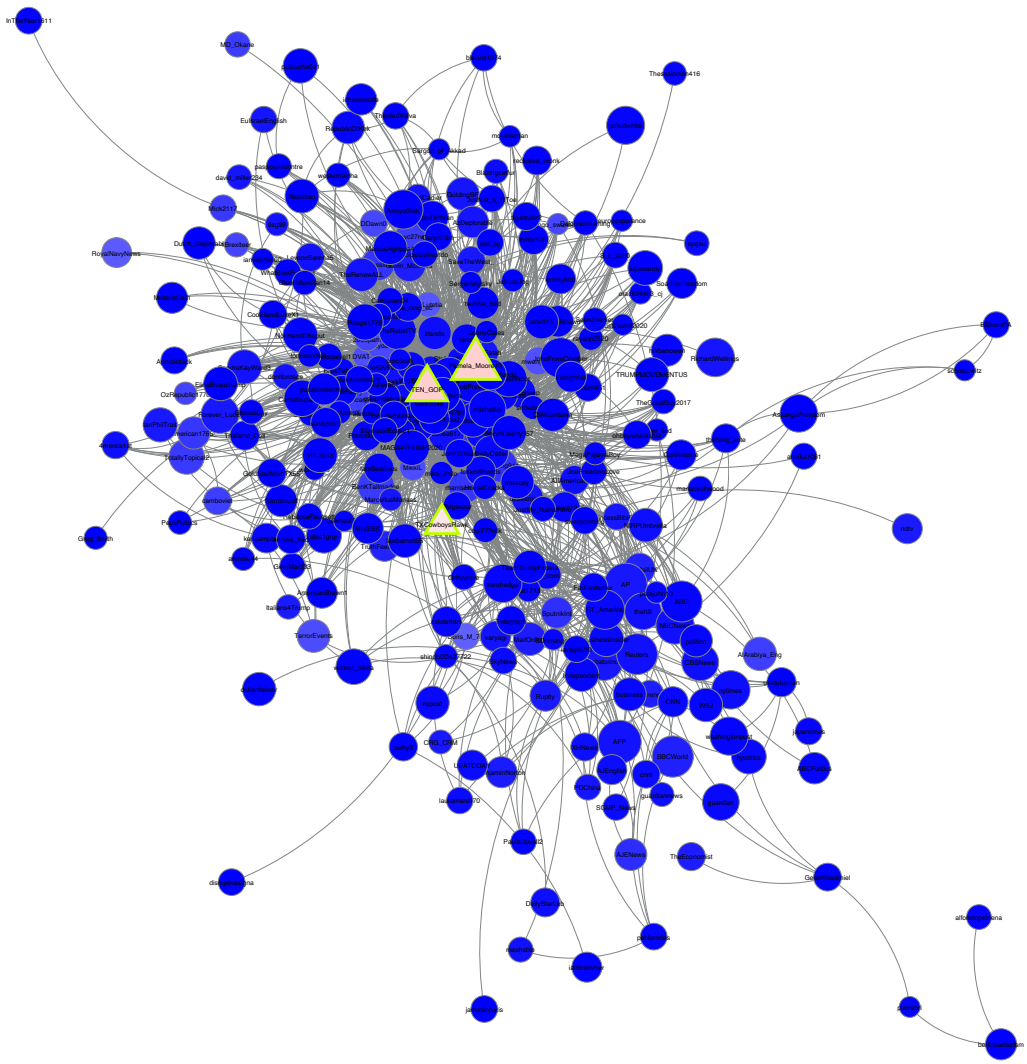


Fig. S6. Classification on English narrative network (*main paper*, Fig. 8) without Snorkel labeling functions; RF / ET classifier. Classifier overfitting without Snorkel is apparent because classifier scores for all accounts not within Twitter's known IO dataset are all clustered near zero. See also Table S9 for the feature importances without Snorkel.

Table S9. Top 36 classifier features* without Snorkel, descending order of importance. Note that classifier overfitting without Snorkel labeling functions is apparent from the prevalence of Serbian-related features that correspond to the dominance of known IO accounts in our training data from Twitter's Serbian dataset (Fig. S1).

lang:sr	en:macron	en:business␣forum
num_tweets_in_time_range	en::serbia:	en:forum
following_count	en:france	en:vučić
num_faves	en:serbian	en:discussed
en:serbia	follower_count	fr:lepen
sd_num_tweets_per_day	en:serbia␣amp	en:serbian␣president
lang:sl	lang:fr	fr:macron
lang:und	avg_num_chars	en:vucic
en:president␣serbia	en:aleksandar␣vučić	lang:es
ratio_retweets_w_links_all_tweets	en:belgrade	en:nursultan
ratio_retweets_all_tweets	en:aleksandar	profile_length
lang:en	avg_num_tweets_per_day	en:election

*2-grams are represented by two words separated by the visible space character '␣'.

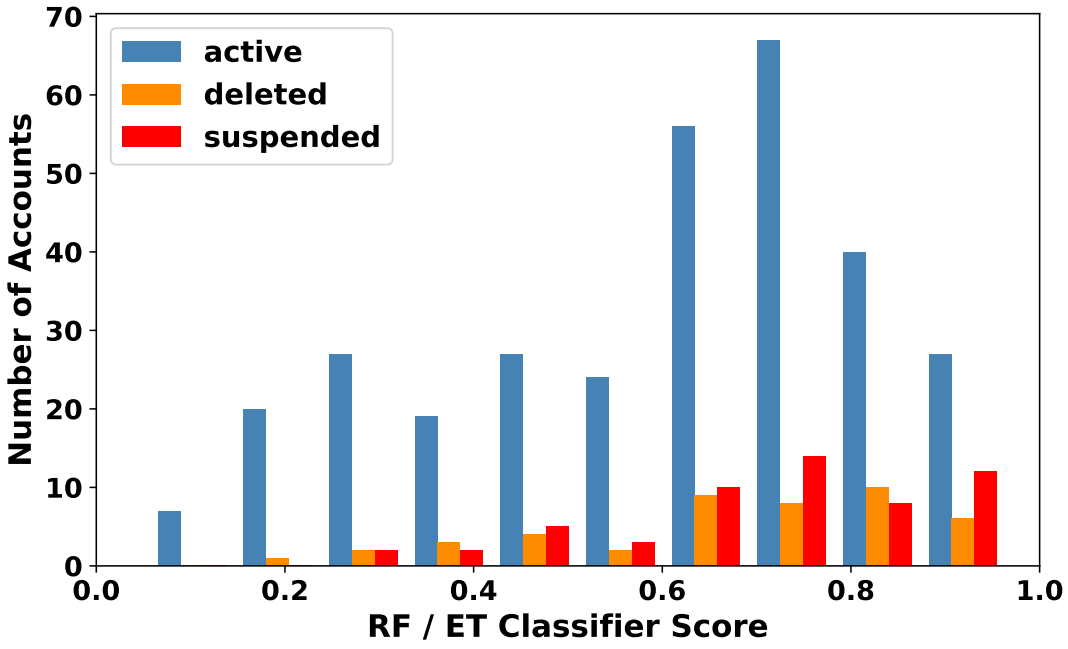


Fig. S7. Classifier score histograms (Random Forest/Extra-Trees) for active, suspended, and deleted accounts in the French narrative network.

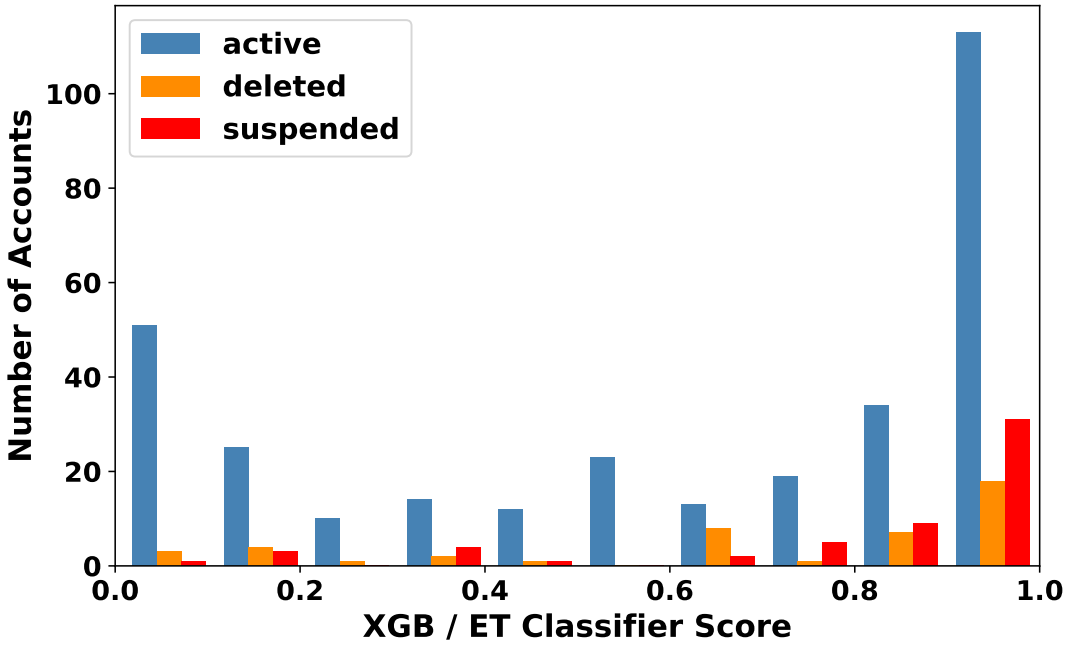


Fig. S8. Classifier score histograms (xgBoost/Extra-Trees) for active, suspended, and deleted accounts in the French narrative network.

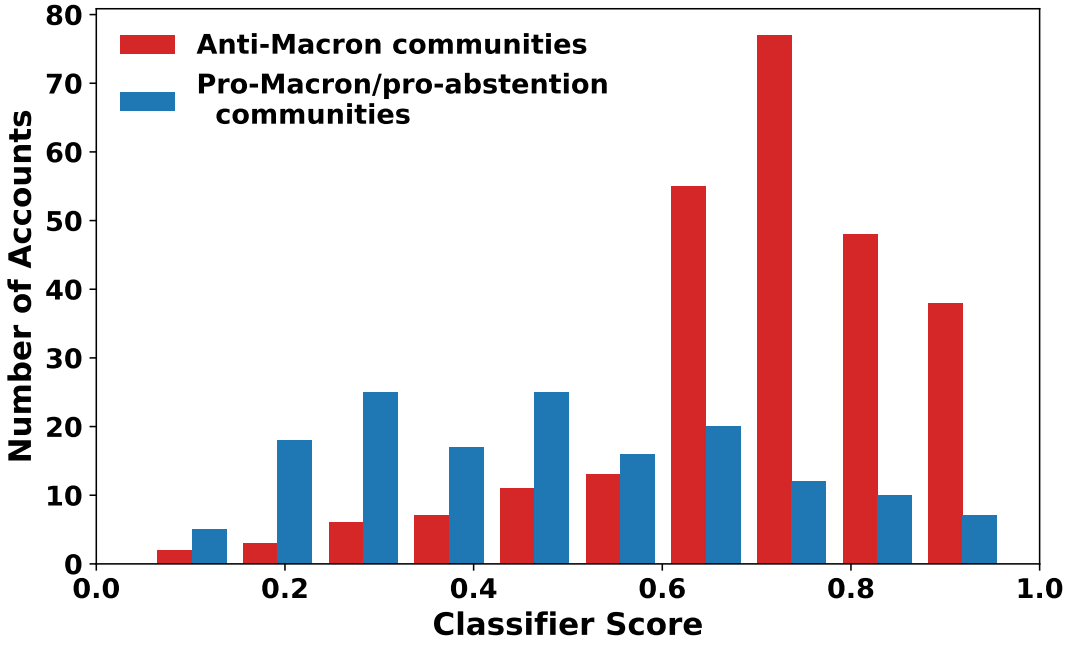


Fig. S9. Relative classifier score frequencies between communities (*main paper*, Fig. 6).

Table S10. Select French topics from the three Macron allegations communities.

Topic A *Macron, police violence at May 1st protests*	**Topic B** *Macron, immigration and Islam*	**Topic C** *Macron, financial allegations*
macron	macron	macron
policiers	france	france
gauche	français	soutien
paris	oradour	marie
crs	marine	emmanuel
france	shoah	uoif
extrême	loi	garaud
mai	veut	marine
1er	2017ledébat	pen
blessés	frontières	voter
1ermai	faire	compte
police	emmanuel	porte
policier	guerre	candidat
brûlé	campagne	plainte
twitter.com	immigration	fiscale
ordre	ans	appelle
guillon	islamiste	islamistes
mort	travail	hollande
:france:	étrangers	grande
forces	vote	2017ledébat

Table S11. Representative tweets of the Topics in Table S10.

Topic A *Macron, police violence at May 1st protests*	Topic B *Macron, immigration and Islam*	Topic C *Macron, financial allegations*
 **[id 535820307] Rottweiller83** @rottweiller83 · 5:26 AM - 28 April 2017 · Retweeted Du matériel #ToutSaufMacron a été volé lors d'une agression ultra violente de 2 jeunes par un gang pro-Macron. Nous allons porter plainte :france:	 **[id 799704104004100096] M-J :latin_cross: :star_of_david: :france:** @Cal_369 · 2:20 PM - 30 April 2017 @EmmanuelMacron Rothschild : appels de fond pour vous financer. Caché au public ! #MacronNon http://www.valeursactuelles.com/politiqu politiques-quand-rothschild-sponsorise-macron-72133	 **[id 4856695311] Alain Thomas** @AlainThomas1 · 1:15 PM- 28 April 2017 · Retweeted Du très très lourd sur @Macron , banquier pourri , "enflure bancaire" financé par Goldman Sachs ! Gravissime ! TWEET ET RETWEET
 **[id 789035544021983232] Julia :france: :pig: :sun: :butterfly:** @mamititi31 · 7:33 PM - 1 May 2017 · Retweeted Les vrais facistes sont ceux qui,à l'extrême gauche,manifestent et cassent en ce moment,refusant le résultat du 1er tour des présidentielles	 **[id 2983945431] l'oranaise** @L_oranaise_ · 4:09 PM - 4 May 2017 · Retweeted voilà ce qui nous attend avec #Macron et après ca voile obligatoire pour les femmes et jeunes filles? :angry_face: :angry_face: :angry_face:	 **[id 768475921112104960] l'oranaise** @Pascal Azoulay · 8:25 AM - 4 May 2017 · Retweeted #MacronGates : Voici 1 copie pour un chèque déposé par Macron dans la banque Nevis as an Offshore Asset Prot C'est bien
 **[id 775389437672947717] Frexit_2017** @avril_sylvie · 7:33 PM - 1 May 2017 · Retweeted La porte-parole de #Macron appelle à la violence contre Marine et nos forces de l'ordre. Effarant. 	 **[id 304858372] MAXIMUS DECIMUS** @lorquaphilip · 7:33 PM - 1 May 2017 · Retweeted Après avoir nié l'existence de la culture française, Macron nie désormais l'existence de la France.	 **[id 799704104004100096] M-J :latin_cross: :star_of_david: :france:** @Cal_369 · 1:00 AM - 4 May 2017 @JackPosobiec #2017LeDebat C'est donc là que sont fric non déclaré à pris la fuite ! https://www.les-crises.fr/emmanuel-macron-36-millions-deuros-de-revenus-cumules-patrimoine-negatif/

Table S12. Select French topics from the pro-Macron/anti-Le Pen/anti-abstention community.

Topic A *pro-Macron and anti-Le Pen*	**Topic B** *pro-Macron and anti-abstention*	**Topic C** *Final election debate: lead-up and event*
ensemble	macron	pen
france	voter	2017ledebat
veux	vote	marine
macron	pen	macron
projet	tour	lepen
europe	appelle	débat
république	mélenchon	programme
pays	faire	2017ledébat
national	france	euro
français	insoumis	mlp
mai	blanc	debat2017
jevotemacron	abstention	mme
politique	mai	projet
macrondirect	marine	soir
liberté	lepen	jevotemacron
:france:	faut	jamais
porte	emmanuel	france
macronpresident	1er	madame
2017ledébat	twitter.com	faire
jt20h	dimanche	twitter.com

Table S13. Select French topics from the pro-abstention community.

Topic A *Support for Jean-Luc Mélenchon*	**Topic B** *Support of abstention*	**Topic C** *Criticism of Macron and Le Pen*
mélenchon	twitter.com	macron
jean	macron	pen
france	faire	marine
mai	voter	utm
insoumis	vote	emmanuel
franceinsoumise	tour	article
1er	mélenchon	source
législatives	insoumis	fillon
1ermai	sansmoile7mai	perdu
insoumise	pen	débat
tour	faut	politique
luc	lepen	discours
youtu.be	oui	france
rdls26	jlm	twitter
paris	mlp	candidat
youtube.com	veut	campaign
melenchon	abstention	social
legislatives2017	blanc	medium
watch	électeurs	lafarge
présidentielle	voix	bit.ly

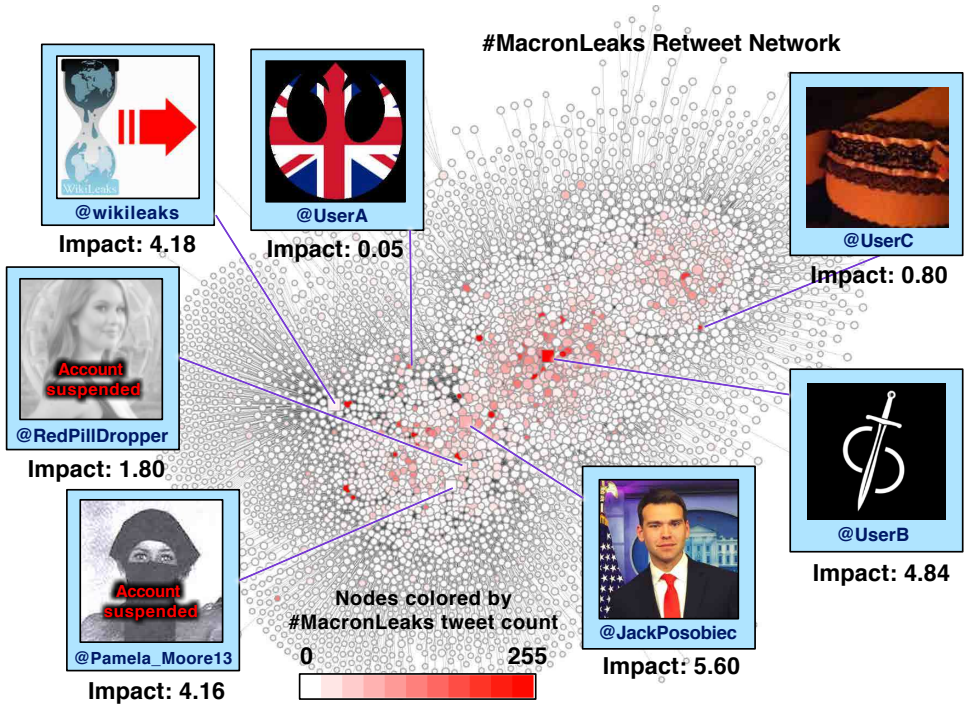


Fig. S10. Causal impact of highlighted accounts on the #MacronLeaks narrative network. Vertices are accounts and edges are retweets. Vertex color indicates the number of tweets and the vertex size corresponds to in-degrees. Causal impact is the average number of additional tweets generated by an user's participation (*main paper*, Eq. [1]). Image credits: Twitter/wikileaks, Twitter/RedPillDropper, Twitter/Pamela_Moore13, Twitter/JackPosobiec.

Table S14. Comparison of impact statistics between accounts on the #MacronLeaks narrative network: tweets (T), total retweets (TRT), most retweeted tweet (MRT), followers (F), first tweet time on 5 May, PageRank centrality (PR), and causal impact* (CI).

Screen name	T	TRT	MRT	F	1st time	PR	CI*
@JackPosobiec	95	47k	5k	261k	18:49	667	**5.60**
@RedPillDropper	32	8k	3k	8k	19:33	44	**1.80**
@UserA†	256	59k	8k	1k	19:34	7	**0.05**
@UserB†	260	54k	8k	3k	20:25	424	**4.84**
@wikileaks	25	63k	7k	5515k	20:32	9294	**4.18**
@Pamela_Moore13	4	4k	2k	54k	21:14	294	**4.16**
@UserC†	1305	51k	8k	< 1k	22:16	6	**0.80**

*Estimate of the causal estimand in *main paper*, Eq. [1]
†Anonymized screen names of currently active accounts

References

1. N. Confessore, G. J. Dance, R. Harris, M. Hansen (2018) The follower factory. *The New York Times*. https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html. Accessed 27 January 2018.
2. L. Fan, W. Wu, X. Zhai, et al. (2014) Maximizing rumor containment in social networks with constrained time. *Soc. Netw. Anal. Min.* 4:214. doi:10.1007/s13278-014-0214-4.
3. R. S. Mueller, III (2019) *Report On The Investigation Into Russian Interference In The 2016 Presidential Election*. (U. S. Department of Justice) Vol. 1. https://www.justsecurity.org/wp-content/uploads/2019/04/Muellller-Report-Redacted-Vol-I-Released-04.18.2019-Word-Searchable.-Reduced-Size.pdf. Accessed 1 March 2020.
4. D. Shah, T. Zaman (2011) Rumors in a network: Who's the culprit? *IEEE Trans. Information Theory* 57(8):5163–5181. doi:10.1109/TIT.2011.2158885.
5. L. G. Stewart, A. Arif, K. Starbird (2018) Examining trolls and polarization with a retweet network in *Proc. ACM WSDM; MIS2: Misinformation and Misbehavior Mining on the Web*. http://snap.stanford.edu/mis2/files/MIS2_paper_21.pdf. Accessed 1 March 2020 doi:10.475/123_4.
6. M. Tambuscio, G. Ruffo, A. Flammini, F. Menczer (2015) Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks in *Proc. ACM WWW*. doi:10.1145/2740908.2742572.
7. H. Zhang, H. Zhang, X. Li, M. T. Thai (2015) Limiting the spread of misinformation while effectively raising awareness in social networks in *Computational Social Networks, CSoNet 2015*, Lecture Notes in Computer Science, eds. M. Thai, N. Nguyen, H. Shen. (Springer) Vol. 9197. doi:10.1007/978-3-319-21786-4_4.
8. Col. S. G.. Chekinov, Lt. Gen. S. A.. Bogdanov (2013) On the nature and content of new generation warfare. *Military Thought* 10:13–24. http://www.eastviewpress.com/Files/MT_FROM%20THE%20CURRENT%20ISSUE_No.4_2013.pdf. Accessed 1 March 2017.
9. V. Pugačiauskas (2015) In the post-Soviet propaganda sphere. *J. Baltic Security* 1(1):127–133. http://www.baltdefcol.org/files/files/JOBS/JOBS.01.1.pdf. Accessed 1 March 2017.
10. NATO StratCom (2015) Analysis of Russia's information campaign against Ukraine, (NATO Strategic Communications Centre of Excellence), Technical report. https://www.stratcomcoe.org/download/file/fid/1886. Accessed 1 March 2017.
11. K. Giles (2016) Handbook of Russian information warfare, (NATO Defense College), Technical Report 9. http://www.ndc.nato.int/download/downloads.php?icode=506. Accessed 1 March 2017.
12. J. Kao, M. S. Li (2020) How China built a Twitter propaganda machine then let it loose on coronavirus. *ProPublica*. https://www.propublica.org/article/how-china-built-a-twitter-propaganda-machine-then-let-it-loose-on-coronavirus. Accessed 1 April 2020.
13. V. Veebel (2016) Estonia confronts propaganda: Russia manipulates media in pursuit of psychological warfare. *Concordiam per J. European Security and Defense Issues* 7:14–19. https://www.marshallcenter.org/mcpublicweb/mcdocs/files/College/F_Publications/perConcordiam/pC_V7_SpecialEdition_en.pdf. Accessed 1 March 2017.
14. J. Borger (2017) US official says France warned about Russian hacking before Macron leak. *The Guardian*. https://www.theguardian.com/technology/2017/may/09/us-russians-hacking-france-election-macron-leak. Accessed 1 January 2018.
15. C. Nyst, N. Monaco (2018) State-sponsored trolling: how governments are deploying disinformation as part of broader digital harassment campaigns, (Institute for the Future), Technical report. http://www.iftf.org/fileadmin/user_upload/images/DigIntel/IFTF_State_sponsored_trolling_report.pdf. Accessed 1 March 2019.
16. E. Birnbaum (2019) Mueller identified 'dozens' of US rallies organized by Russian troll farm. *The Hill*. https://thehill.com/policy/technology/439532-mueller-identified-dozens-of-us-rallies-organized-by-russian-troll-farm. Accessed 18 May 2019.
17. S. Kargar, A. Rauchfleisch (2019) State-aligned trolling in Iran and the double-edged affordances of Instagram. *new media & society* 21(7):1506–1527. doi:10.1177/1461444818825133.
18. T. Rid (2020) *Active Measures: The Secret History of Disinformation and Political Warfare*. (Farrar, Straus and Giroux, New York NY).
19. L. Dearden (2017) Emmanuel Macron launches legal complaint over offshore account allegations spread by Marine Le Pen. *The Independent*. https://www.independent.co.uk/news/world/europe/french-presidential-election-latest-emmanuel-macron-legal-complaint-marine-le-pen-offshore-account-a7717461.html. Accessed 1 April 2020.
20. J. McAuley, I. Stanley-Becker (2017) Macron campaign says its emails have been subjected to 'massive, coordinated' hacking. *The Washington Post*. https://www.washingtonpost.com/world/macron-campaign-says-its-emails-have-been-subjected-to-massive-coordinated-hacking/2017/05/06/368c0460-31e1-11e7-a335-fa0ae1940305_story.html. Accessed 1 March 2020.
21. M. S. Schmidt, N. Perlroth (2020) U. S. charges Russian intelligence officers in major cyberattacks. *The New York Times*. https://www.nytimes.com/2020/10/19/us/politics/russian-intelligence-cyberattacks.html. Accessed 19 October 2020.
22. M. Akbarpour, M. O. Jackson (2018) Diffusion in networks and the virtue of burstiness. *Proc. Natl. Acad. Sci. U.S.A.* 115(30):E6996–E7004. doi:10.1073/pnas.1722089115.
23. G. Pennycook, D. G. Rand (2019) Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc. Natl. Acad. Sci. U.S.A.* 116(7):2521–2526. doi:10.1073/pnas.1806781116.
24. N. S. Contractor, L. A. DeChurch (2014) Integrating social networks and human social motives to achieve social influence at scale. *Proc. Natl. Acad. Sci. U.S.A.* 111 (Supplement 4)(49):13650–13657. doi:10.1073/pnas.1401211111.
25. A. K. McCallum (2002) Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu. Accessed 1 January 2018.
26. A. Ratner, et al. (2017) Snorkel: Rapid training data creation with weak supervision in *Proc. VLDB Endowment*. Vol. 11, pp. 269–282. doi:10.14778/3157794.3157797.
27. D. Costa-Roberts (2018) How to spot a Russian bot. *Mother Jones*. https://www.motherjones.com/media/2018/08/how-to-identify-russian-bots-twitter. Accessed 1 March 2020.
28. J. Im, et al. (2019) Still out there: Modeling and identifying Russian troll accounts on Twitter. arxiv:1901.11162.
29. S. Zannettou, et al. (2019) Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the web in *Proc.*

2019 World Wide Web Conf. pp. 218–226. doi:10.1145/3308560.3316495.
30. L. Luceri, S. Giordano, E. Ferrara (2020) Don't feed the troll: Detecting troll behavior via inverse reinforcement learning in *Proc. 2020 Intl. Conf. Web and Social Media (ICWSM)*. arxiv:2001.10570.
31. O. Varol, E. Ferrara, C. A. Davis, F. Menczer, A. Flammini (2017) Online human-bot interactions: Detection, estimation, and characterization in *Proc. 11th Intl. AAAI Conf. Web and Social Media (ICWSM 2017)*. pp. 280–289. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/viewPaper/15587. Accessed 1 January 2018.
32. A. Liaw, M. Wiener (2002) Classification and regression by RandomForest. *R news* 2(3):18–22.
33. T. Chen, C. Guestrin (2016) Xgboost: A scalable tree boosting system in *Proc. ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining (KDD)*. pp. 785–794. doi:10.1145/2939672.2939785.
34. F. Pedregosa, et al. (2011) Scikit-learn: Machine learning in Python. *J. Machine Learning Research* 12:2825–2830. arxiv:1201.0490.
35. P. Geurts, D. Ernst, L. Wehenkel (2006) Extremely randomized trees. *Machine Learning* 63(1):3–42. doi:10.1007/s10994-006-6226-1.
36. A. Marantz (2017) The far-right American nationalist who tweeted #MacronLeaks. *The New Yorker*. https://www.newyorker.com/news/news-desk/the-far-right-american-nationalist-who-tweeted-macronleaks. Accessed 1 January 2018.
37. US House Permanent Select Committee on Intelligence (2017) HPSCI minority exhibits during open hearing, memorandum. https://democrats-intelligence.house.gov/uploadedfiles/hpsci_minority_exhibits_memo_11.1.17.pdf. Accessed 1 January 2018.
38. US House Permanent Select Committee on Intelligence (2017) Exhibit of the user account handles that Twitter has identified as being tied to Russia's "Internet Research Agency.". https://democrats-intelligence.house.gov/uploadedfiles/exhibit_b.pdf. Accessed 1 January 2018.
39. V. Gadde, Y. Roth (2018) Enabling further research of information operations on Twitter. https://blog.twitter.com/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.html. Accessed 1 January 2020.
40. S. T. Smith, E. K. Kao, D. C. Shah, O. Simek, D. B. Rubin (2018) Influence estimation on social media networks using causal inference in *Proc. 2018 IEEE Statistical Signal Processing Workshop (SSP)*. pp. 28–32. doi:10.1109/SSP.2018.8450823.
41. E. K. Kao (2017) Causal inference under network interference: A framework for experiments on social networks. *Harvard University*. arxiv:1708.08522.
42. G. Basse, A. Feller (2016) Analyzing multilevel experiments in the presence of peer effects. arxiv:1608.06805.
43. D. A. Kim, et al. (2015) Social network targeting to maximise population behaviour change: a cluster randomised controlled trial. *The Lancet* 386(9989):145–153. doi:10.1016/S0140-6736(15)60095-2.
44. S. C. Mednick, N. A. Christakis, J. H. Fowler (2010) The spread of sleep behavior influences drug use in adolescent social networks. *PLoS One* 5(3):e9775. doi:10.1371/journal.pone.0009775.
45. M. E. Sobel (2006) What do randomized studies of housing mobility demonstrate? *J. American Statistical Association* 101(476):1398–1407. doi:10.1198/016214506000000636.
46. H. Gui, Y. Xu, A. Bhasin, J. Han (2015) Network A/B testing: From sampling to estimation in *Proc. 24th Intl. Conf. World Wide Web*. (International World Wide Web Conferences Steering Committee), pp. 399–409. doi:10.1145/2736277.2741081.
47. R. M. Bond, et al. (2012) A 61-million-person experiment in social influence and political mobilization. *Nature* 489(7415):295–298. doi:10.1038/nature11421.
48. E. Bakshy, D. Eckles, R. Yan, I. Rosenn (2012) Social influence in social advertising: Evidence from field experiments in *Proc. 13th ACM Conf. Electronic Commerce*. pp. 146–161. doi:10.1145/2229012.2229027.
49. B. M. Parker (2011) Design of network experiments. http://www.newton.ac.uk/programmes/DAE/seminars/090111301.pdf. Accessed 1 January 2013.
50. K. Coronges, et al. (2012) The influences of social networks on phishing vulnerability in *Proc. 45th Intl. Conf. System Science (HICSS)*. (IEEE), pp. 2366–2373. doi:10.1109/HICSS.2012.657.
51. A. Banerjee, A. G. Chandrasekhar, E. Duflo, M. O. Jackson (2013) The diffusion of microfinance. *Science* 341(6144):1236498. doi:10.1126/science.1236498.
52. D. Acemoglu, A. Ozdaglar, A. ParandehGheibi (2010) Spread of (mis)information in social networks. *Games and Economic Behavior* 70(2):194–227. doi:10.1016/j.geb.2010.01.005.
53. C. F. Manski (1993) Identification of endogenous social effects: The reflection problem. *Rev. Economic Studies* 60(3):531–542. doi:10.2307/2298123.
54. O. David, R. A. Kempton (1996) Designs for interference. *Biometrics* 52(2):597–606. doi:10.2307/2532898.
55. J. M. Azais, R. A. Bailey, H. Monod (1993) A catalogue of efficient neighbour-designs with border plots. *Biometrics* 49(4):1252–1261. doi:10.2307/2532269.
56. D. B. Rubin (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educational Psychology* 66(5):688–701. doi:10.1037/h0037350.
57. G. W. Imbens, D. B. Rubin (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences*. (Cambridge University Press). doi:10.1017/CBO9781139025751.
58. J. Bowers, M. M. Fredrickson, C. Panagopoulos (2013) Reasoning about interference between units: A general framework. *Political Analysis* 21(1):97–124. doi:10.1093/pan/mps038.
59. S. Athey, D. Eckles, G. W. Imbens (2018) Exact *p*-values for network interference. *J. Am. Stat. Assoc.* 113(521):230–240. doi:10.1080/01621459.2016.1241178.
60. P. M. Aronow, C. Samii (2017) Estimating average causal effects under general interference, with application to a social network experiment. *Ann. Appl. Stat.* 11(4):1912–1947. doi:10.1214/16-AOAS1005.
61. J. Ugander, B. Karrer, L. Backstrom, J. Kleinberg (2013) Graph cluster randomization: Network exposure to multiple universes in *ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining (KDD)*. pp. 329–337. doi:10.1145/2487575.2487695.

62. D. L. Sussman, E. M. Airoldi (2017) Elements of estimation theory for causal effects in the presence of network interference. arxiv:1702.03578.
63. S. Li, S. Wager (2020) Random graph asymptotics for treatment effect estimation under network interference. arxiv:2007.13302.
64. T. J. VanderWeele, E. J. Tchetgen, M. E. Halloran (2014) Interference and sensitivity analysis. *Statist. Sci.* 29(4):687–706. doi:10.1214/14-STS479.
65. E. L. Ogburn, T. J. VanderWeele (2014) Causal diagrams for interference. *Statist. Sci.* 29(4):559–578. doi:10.1214/14-STS501.
66. C. R. Shalizi, A. C. Thomas (2011) Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research* 40(2):211–239. doi:10.1177/0049124111404820.
67. S. Aral, L. Muchnik, A. Sundararajan (2009) Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Natl. Acad. Sci. U.S.A.* 106(51):21544–21549. doi:10.1073/pnas.0908800106.
68. A. J. O'Malley, F. Elwert, J. N. Rosenquist, A. M. Zaslavsky, N. A. Christakis (2014) Estimating peer effects in longitudinal dyadic data using instrumental variables. *Biometrics* 70(3):506–515. doi:10.1111/biom.12172.
69. D. Eckles, R. F. Kizilcec, E. Bakshy (2016) Estimating peer effects in networks with peer encouragement designs. *Proc. Natl. Acad. Sci. U.S.A.* 113(27):7316–7322. doi:10.1073/pnas.1511201113.
70. D. B. Rubin (1980) Randomization analysis of experimental data: The Fisher randomization test comment. *J. Am. Stat. Assoc.* pp. 587–589. doi:10.2307/2287653.
71. P. Toulis, E. Kao (2013) Estimation of causal peer influence effects in *Proc. 30th Intl. Conf. Machine Learning*. pp. 1489–1497.
72. D. B. Rubin (1991) Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics* 47(4):1213–1234. doi:10.2307/2532381.
73. D. B. Rubin (1978) Bayesian inference for causal effects: The role of randomization. *Ann. Stat.* pp. 34–58. https://www.jstor.org/stable/2958688.
74. P. D. Hoff, A. E. Raftery, M. S. Handcock (2002) Latent space approaches to social network analysis. *J. Am. Stat. Assoc.* 97(460):1090–1098. doi:10.1198/016214502388618906.
75. Y. J. Wang, G. Y. Wong (1987) Stochastic blockmodels for directed graphs. *J. American Statistical Association* 82(397):8–19. doi:10.1080/01621459.1987.10478385.
76. E. M. Airoldi, D. M. Blei, S. E. Fienberg, E. P. Xing (2008) Mixed membership stochastic blockmodels. *J. Machine Learning Research* 9(1981–2014):3.
77. E. K. Kao, S. T. Smith, E. M. Airoldi (2019) Hybrid mixed-membership blockmodel for inference on realistic network interactions. *IEEE Trans. Network Science and Engineering* 6(3):336–350. doi:10.1109/TNSE.2018.2823324.
78. W. Aiello, F. Chung, L. Lu (2001) A random graph model for power law graphs. *Experimental Mathematics* 10(1):53–66. doi:10.1080/10586458.2001.10504428.
79. F. Chung, L. Lu (2002) The average distances in random graphs with given expected degrees. *Proc. Natl. Acad. Sci. U.S.A.* 99(25):15879–15882. doi:10.1073/pnas.252631999.
80. L. Lovász, B. Szegedy (2006) Limits of dense graph sequences. *J. Combinatorial Theory, Series B* 96(6):933–957. doi:10.1016/j.jctb.2006.05.002.
81. C. E. Frangakis, D. B. Rubin (1999) Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* 86(2):365–379. doi:10.1093/biomet/86.2.365.
82. A. Gelman, A. Jakulin, M. G. Pittau, Y. Su (2008) A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* 2(4):1360–1383.
83. A. Gelman, X. Meng, H. Stern (1996) Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* pp. 733–760.
84. D. B. Rubin (2004) *Multiple imputation for nonresponse in surveys*, Wiley Classics. (John Wiley & Sons, Hoboken NJ) Vol. 81.
85. L. Forastiere, E. M. Airoldi, F. Mealli (2020) Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association* pp. 1–18.
86. S. Han, D. B. Rubin (2020) Contrast specific propensity scores. arxiv:2007.01253.
87. E. Ferrara (2017) Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday* 22(8–7). https://firstmonday.org/ojs/index.php/fm/article/download/8005/6516. Accessed 1 July 2020.
88. A. N. Langville, C. D. Meyer (2005) A survey of eigenvector methods for web information retrieval. *SIAM review* 47(1):135–161. doi:10.1137/S0036144503424786.
89. T. P. Peixoto (2014) The graph-tool Python library. http://graph-tool.skewed.de. Accessed 1 March 2017.
90. S. T. Smith, E. K. Kao, K. D. Senne, G. Bernstein, S. Philips (2014) Bayesian discovery of threat networks. *IEEE Trans. Signal Proc.* 62(20):5324–5338. doi:10.1109/TSP.2014.2336613.
91. J. Donati (2020) U.S. adversaries are accelerating, coordinating coronavirus disinformation, report says. *The Wall Street Journal*. https://www.wsj.com/articles/u-s-adversaries-are-accelerating-coordinating-coronavirus-disinformation-report-says-11587514724. Accessed 21 April 2020.
92. J. Donovan (2020) Here's how social media can combat the coronavirus 'infodemic'. *MIT Technology Review*. https://www.technologyreview.com/2020/03/17/905279/facebook-twitter-social-media-infodemic-misinformation/. Accessed 1 April 2020.
93. D. Linvill, P. Warren (2020) The Russia Tweets. https://russiatweets.com. Accessed 1 April 2020.