Health Behavior Research

Volume 4
Number 2 *Special Issue*

Article 4

April 2021

# The importance of interdisciplinary frameworks in social media mining: An exploratory approach between Computational Informatics and Social Network Analysis (SNA)

Danny Valdez
*Indiana University School of Public Health*, danvald@iu.edu

Meg Patterson
*Texas A&M University*, megpatterson@tamu.edu

Tyler Prochnow MEd
*Baylor University*, tyler_prochnow1@baylor.edu



## Recommended Citation

Valdez, Danny; Patterson, Meg; and Prochnow, Tyler MEd (2021) "The importance of interdisciplinary frameworks in social media mining: An exploratory approach between Computational Informatics and Social Network Analysis (SNA)," *Health Behavior Research*: Vol. 4: No. 2. https://doi.org/10.4148/2572-1836.1098

# The importance of interdisciplinary frameworks in social media mining: An exploratory approach between Computational Informatics and Social Network Analysis (SNA)


## Abstract

Social media content is one of the most visible sources of big data and is often used in health studies to draw inferences about various behaviors. Though much can be gleaned from social media data and mining, the approaches used to collect and analyze data are generally strengthened when examined through established theoretical frameworks. Health behavior, a theory driven field, encourages interdisciplinary collaboration across fields and theories to help us draw robust conclusions about phenomena. This pilot study uses a combined computer informatics and SNA approach to analyze information spread about mask-wearing as a personal mitigation effort during the COVID-19 pandemic. We analyzed one week's worth of Twitter data ($n$ = 10,107 tweets across 4,289 users) by using at least one of four popular mask-support hashtags (e.g., #maskup). We calculated network-measures to assess structures and patterns present within the Twitter network, and used exponential random graph modeling (ERGM) to test factors related to the presence of retweets between users. The pro-mask Twitter network was largely fragmented, with a select few nodes occupying the most influential positions in the network. Verified accounts, accounts with more followers, and those who generated more tweets were more likely to be retweeted. Contrarily, verified accounts and those with more followers were less likely to retweet others. SNA revealed patterns and structures theoretically important to how information spreads across Twitter. We demonstrated the utility of an interdisciplinary collaboration between computer informatics and SNA to draw conclusions from social media data.



## Keywords

Social media, Informatics, Social Network Analysis; COVID-19

## Acknowledgements/Disclaimers/Disclosures

The authors have no conflict of interest to report.

# The Importance of Interdisciplinary Frameworks in Social Media Mining: An Exploratory Approach Between Computational Informatics and Social Network Analysis (SNA)

**Danny Valdez, Ph.D.***
**Megan S. Patterson, Ph.D.**
**Tyler Prochnow, M.Ed.**

## Abstract

Social media content is one of the most visible sources of big data and is often used in health studies to draw inferences about various behaviors. Though much can be gleaned from social media data and mining, the approaches used to collect and analyze data are generally strengthened when examined through established theoretical frameworks. Health behavior, a theory driven field, encourages interdisciplinary collaboration across fields and theories to help us draw robust conclusions about phenomena. This pilot study uses a combined computer informatics and SNA approach to analyze information spread about mask-wearing as a personal mitigation effort during the COVID-19 pandemic. We analyzed one week's worth of Twitter data ($n$ = 10,107 tweets across 4,289 users) by using at least one of four popular mask-support hashtags (e.g., #maskup). We calculated network-measures to assess structures and patterns present within the Twitter network, and used exponential random graph modeling (ERGM) to test factors related to the presence of retweets between users. The pro-mask Twitter network was largely fragmented, with a select few nodes occupying the most influential positions in the network. Verified accounts, accounts with more followers, and those who generated more tweets were more likely to be retweeted. Contrarily, verified accounts and those with more followers were less likely to retweet others. SNA revealed patterns and structures theoretically important to how information spreads across Twitter. We demonstrated the utility of an interdisciplinary collaboration between computer informatics and SNA to draw conclusions from social media data.

*Corresponding author can be reached at: danvald@iu.edu

## Introduction

In the big data era, there are abundant resources that can be mined to study nuanced aspects of human behavior. Social media, defined as websites where users post shareable, personal content in real-time (Wang & Wei, 2012), is one of the most visible sources of such information. Indeed, the diachronic nature of social media feeds afford opportunities to study in-the-moment portrayals of emerging social phenomena (Valdez et al., 2020). And, as much of social media constitute part of the public domain, social media also represents a free source of abundant data that can be studied and analyzed in tandem using quantitative, qualitative, or mixed methodologies.

Historically, we have relied on computational informatics methods to collect and analyze large quantities of social media data to draw inferences about human behavior. This family of techniques uses computer algorithms to read, and learn from, text content to draw conclusions about themes (i.e., topic models), opinions (i.e., sentiment analysis), and demographic characteristics of users (i.e., classifiers) based on patterns within social media posts and other content (Simms et al., 2017). While the processes used to procure data and methods

used to analyze them efficiently summarize abundant information, they are generally strengthened when used alongside established theoretical frameworks and methodologies that emphasize the interconnectedness of social media data between posts and users.

Social Network Analysis (SNA) is both a method and a theory used to understand the interconnectedness of people, data, and systems (Valente, 2010). The premise of SNA is to measure how structures, patterns, and positions within a network relate to various outcomes, including: information spread (Bueno, 2015; Hambrick, 2012); participation in certain behaviors (e.g., smoking, physical activity, contraception use, or gaming; Boulay and Valente, 2005; Patterson et al., 2019; Prochnow et al., 2020; Valente et al., 2013); and disease transmission (Emch et al., 2012; Klovdahl, 1985). Social media websites, which by nature connect users to one another in an online space, are strong examples of social networks in practice. Whereas informatics is used to consolidate and explore social media data, SNA provides a structure with which to better understand it (Patterson et al., 2019). Thus, the theoretical structure of SNA can add to and inform the robust data collection efforts of data mined from social media.

Herein we discuss the utility of SNA and social media mining in practice. Although this interdisciplinary collaboration between social media mining and SNA is not new, it is perhaps less critically studied using examples that pertain exclusively to Health Behavior and health messaging (see Xu and Li, 2013 for insights into social media mining and SNA). Throughout 2020, the COVID-19 pandemic has dominated much health-related discourse. This discourse has been played out through social media as users share pertinent content about the pandemic, such as news and personal perspectives, among other content. By combining the data collection abilities afforded through social media mining and calculating network structures through SNA, we can glean important insights into how pandemic-related information (in our example, facemask use) spreads within networks.

## The COVID-19 pandemic, facemasks, and information spread

The COVID-19 pandemic led to unprecedented mitigation efforts to curb the spread of SARS- CoV-2. According to the Centers for Disease Control and Prevention (CDC), one of the most effective personal mitigation efforts is the use of personal face coverings (i.e. facemasks). When used correctly, facemasks can prevent 17-45% of COVID-19-related deaths (Eikenberry et al., 2020). Indeed, at the time of writing, mask wearing has proven to be one of the strongest and easiest mitigators of COVID-19 infection, which may have stunted some of the recent upticks in COVID-19 cases (Peeples, 2020). Multiple companies, private businesses, and even state and local governments have created facemask-wearing guidelines to promote continuous and appropriate mask use to reduce COVID-19 spread.

The effectiveness and simplicity of facemask use has led to much activism present on social media lobbying for widespread adoption of mask use, including at times in which wearing a facemask is not required (Sobowale et al., 2020). For example, during later months of the COVID-19 pandemic numerous health professionals, celebrities, and other influencers used social media to disseminate information about the importance of consistent and correct facemask use (Ahmed et al., 2020). These included specific hashtags about facemask wearing, news content related to facemask wearing, photos of individuals in facemasks,

among dissemination of other mask related information.

Social media is widely known to influence health behavior (Centola, 2013; Nelon et al., 2020). This phenomenon likely transcends into public health crises, such as the COVID-19 pandemic, where individuals are seeking pandemic-related news at increased volumes (see Bento et al., 2020). With facemask wearing constituting one of the more recent social media trends emerging during the COVID-19 pandemic, facemask use is an ideal example to explore and discuss the integration of SNA, social media analytics, and health behavior— particularly with regard to how the sharing of pro-facemask-wearing information is spread online. This can be accomplished by analyzing the network structures of data collected through social media mining. For many, these two fields of study are well known. However, we present a brief primer of each to situate our study in the theories in which it is grounded.

## A Short Primer on Network Theory

Social network analysis (SNA) represents a theoretical framework and a methodology that studies the interconnectedness of people and systems (Borgatti et al., 2018). Network theory posits that the way entities are connected and structured within a network drives outcomes as much, if not more, than individual-level attributes (Borgatti et al., 2018). There are several structural indices computed via network data that, according to network theory, have implications on individual-, group-, and network-level outcomes. For example, at the individual-level, a person positioned more centrally in a network likely receives more social support and social capital, and faces greater social constraints, as compared to someone more peripheral in the network (Valente, 2010). Similarly, nodes structurally positioned between dense subgroups within networks have strong influence due to their control of what information moves from one group to another (Freeman, 1978). They serve as gatekeepers and bridges for information spread. At the group-level, clusters or cliques (i.e., densely connected subgroups within a larger network) can impact information and behavior spread. Information is passed, and behaviors are adopted, quickly within clusters, but may be slower to move outside of clusters to other parts of the network (Granovetter, 1985). Finally, at the network-level, densely connected networks will spread information quicker than a more fragmented network (Valente, 2005). Therefore, understanding the various properties of a network, and how networks are structured, could have implications on how public health information is spread through social media networks, as well as inform future programmatic or intervention strategies (Valente, 2012; Valente et al., 2015).

## A Short Primer on Social Media Mining

Social media mining refers to the collection and analysis of data derived from social networking websites (e.g., Twitter, Instagram, Facebook, and others). Much of the data collection is done using computer code which elicits data through a website's Application Programming Interface (API). The API can generate information such as User ID, the text used to comprise a post, friend counts, and the number of times a social media post was shared with others. Data collected through an API are similarly analyzed using computer code and algorithms that explore and draw inferences about the collected data. To date, Twitter remains a prominent data source to study various population level phenomena (e.g., Karami et al., 2020). This is due, in part, to Twitter's user agreement, which makes posts written by users part of the public domain

(see developer.twitter.com), and the platform's commonplace use for relaying current-events information (Moon & Hadley, 2014). Using social media mining collection and analysis, Twitter data has been used widely to study various psychosocial phenomena including to predict the stock market (Bollen et al., 2011), map mood during natural disasters (Cho et al., 2013), and infer mental health status about the COVID-19 pandemic (Valdez et al., 2020).

### Present Study

Social network theory emphasizes the importance of accurate and responsible public health messaging due to the ways networks serve as a mechanism for information spread (Valente, 2010). Social media mining affords researchers the data necessary to draw conclusions about networks through an SNA approach. By identifying the ways tweets with pro-mask rhetoric are structured and patterned, we gain insight into how Twitter can be leveraged to increase positive messaging, and as a result, improve public health behaviors. We seek to answer three research questions:

1. What are the basic network structures related to the retweeting of pro-mask tweets?
2. How might these structures be related to information spread?
3. What are the implications and applications of an interdisciplinary collaboration between social media mining and SNA?

## Methods

### Data Collection

This is a pilot investigation testing a combined social media mining and SNA approach. As such, we collected one week's worth of tweets pertaining to support for protective mask wearing in the United States through Twitter's API (August 7-August 14) to create a manageable dataset. To procure our data, we began by identifying tweet IDs comprised of individual Twitter posts containing one, or a combination of, popularly used pro mask-related hashtags identified by Google Trend data:

1. #maskup*,
2. #maskssavelives*,
3. #wearadamnmask*,
4. #maskitorcasket*

(endpoint: GET statuses/show/id). Note, asterisks were added at the end of each hashtag during the search query to include wildcards—i.e., accidental misspellings of a hashtag that likely intend to convey the same meaning (e.g., #masksup versus #maskup).

Using a proprietary Decahose provided by the Indiana University Network Science Institute (IUNI) that nets users 10% of total tweets through Twitter's API, we downloaded each tweet, as well as the standard metadata provided by Twitter including: the user who posted the content; the number of times a tweet was retweeted; the total number of followers; the total number of tweets per account; and the verified status (i.e., Twitter's process of ensuring accounts originate from specific people, such as celebrities). See https://iuni.iu.edu/projects for more information.

We then removed non-English tweets ($n$ = 324) and potential bots ($n$ = 74). Our final sample was comprised of ($n$ = 4,289) users, and ($n$ = 10,107) tweets. Note, we evaluated accounts for possibly bot-like behavior using a proprietary algorithm Botometer (previously Bot or Not), which provides a bot-score based on written patterns within a specific account. We then used our hashtag data to generate an edgelist, which reorganized our data by use of one (or a

combination) of the included hashtags, the associated tweet, and the extent to which a source tweet was retweeted by Person A (the follower) from Person B (the author).

## Network Operationalization

After collecting the data, we created a network using our social media data by connecting accounts (nodes) that retweeted others using one of the aforementioned pro-mask hashtags. To operationalize the network, a directed connection was defined as the action of retweeting another account. In other words, the connection would be directed from the account which *made the retweet* to the account which *made the original tweet* to understand the action of actively spreading or retweeting pro-mask sentiments. Further, the nodes in this network would then be individual Twitter users/accounts, while the connections would be the act of retweeting.

## Data Analysis

**Descriptive statistics.** Sample characteristics, including means, standard deviations, and frequencies, were calculated for account followers, tweets, and verification status (i.e., whether an account was deemed "verified" by Twitter) using RStudio. Network descriptive statistics including centrality measures, group measures, and network-level measures were computed using the igraph package (Csárdi & Nepusz, 2006). Centrality is a property of a node's position within a network (Borgatti et al., 2018), and was measured in the form of in-degree, out-degree, betweenness, closeness, and eigenvector centrality in this study. Group measures, which mathematically identify groups or communities of nodes (Valente, 2010), included a k-core analysis and community detection using modularity. Finally, network-level measures included density, centralization, and transitivity, and represent overall patterns present within the entire network (Valente, 2010). See Table 1 for definitions of each network measure, as well as results from these data.

**Exponential random graph models.** We used exponential random graph modeling (ERGM) to further understand the action of retweeting within this network. Through the use of iterative Marcov chair Monte Carlo algorithms, ERGMs approximate the maximum likelihood estimates for the log-odds of associations between given factors (structural and attribute-related factors such as density or number of followers) and tie presence within networks. To do so, ERGMs use the empirical network data combined with certain researcher assigned parameters to simulate other networks, which are then compared to the original empirical network and used to determine statistical significance, serving as a nonparametric approach to handling dependence within data (Lusher, Koskinen, & Robins, 2013). ERGMs return parameter estimates (PE) and standard errors (SE) for each factor entered into the model. A factor is deemed significant at a $p < .05$ level if the PE is greater than two times the SE. Cleaning and management of network data, along with ERGM analyses, were completed using the statnet package in R Studio (Handcock et al., 2019).

**ERGM model specification.** Network structure parameters were added to model density (edges or connections in the network), reciprocated connections (connections that are mutually shared between two individuals), and transitive triads (three individuals connected to each other). Sender and receiver covariates were added for number of followers and number of tweets to determine significant associations between those variables and either retweeting someone (sending) or being retweeted (receiving). Similarly, factors were added to

Table 1

*Centrality, Group-level, and Network-level Characteristics*

| Term | Definition | Mean or Network Score | SD |
|---|---|---|---|
| **Centrality** | | | |
| *In-degree* | The number of times a node (user) was retweeted (connection established by a user retweeting the node's content) | $M = 0.73$ | 3.52 |
| *Out-degree* | The number of times a node retweeted information out in their network (connection established by the node retweeting another user's content) | $M = 0.73$ | 0.47 |
| *Betweeness* | Measures how often a node falls along the shortest path between two other nodes. | $M = 0.05$ | 1.20 |
| *Closeness* | A reverse measure of distance from the center of the network. Nodes with higher closeness scores are more reachable. | $M < 0.001$* | < 0.01* |
| *Eigenvector Centrality* | How connected someone is to the most central/popular/powerful people in a network | $M < 0.001$* | 0.02 |
| **Group-level** | | | |
| *k-cores* | *k*-cores identify densely connected "cores" of nodes. The k-value represents the densest core, with lower *k*-cores representing less connected (and more peripheral) nodes | $K = 3$ ($n = 11$ nodes in 3*k* core) | n/a |
| *Community Detection* | Identifies groups of connected/clustered nodes within the overall network; modularity reflects how fragmented a network is, with higher modularity scores revealing more fragmented networks | 1,175 communities<br>$M = 3.59$ nodes<br>Modularity = 0.99 | 6.62 |
| **Network Level** | | | |
| *Density* | The proportion of ties that exist in a network compared to the total possible number of connections in a network | < 0.001* | |
| *Centralization* | A measure of network structure; higher centralization indices indicate a more "hierarchical" structure across the network | 0.03 | |
| *Transitivity* | The proportion of all triangles in a network; indicates clustering in the network, with higher transitivity scores meaning more nodes have connections in common | < 0.001* | |

*Note.* *values are less than 0.001 and may be a reflection of the disconnected nature of the network at large.

determine whether being a verified account increased the odds of retweeting or being retweeted. A homophily (sameness) term was added to determine whether accounts retweeted others who matched their verified status more than would be by chance. Lastly, an in-degree popularity term was added to model the propensity of being retweeted because many others have already retweeted the account.

## Results

### Sample Characteristics

Overall, 5.2% ($n$ = 221) of users ($n$ = 4,289) were verified. Number of followers per user ranged from 0 to 3,604,035, with a mean of 13,937.2 followers (SD = 111,220.6) and median of 970 followers. Number of tweets per user ranged from 1 to 1,621,861, with a mean of 51,057.7 (SD = 106,307.9) and median of 14,855 tweets. Users were retweeted in this sample 0.73 times on average (SD = 3.51), with one user being retweeted as many as 108 times. Nearly three quarters (70.8%, $n$ = 2,984) of the network's users were never retweeted. Users retweeted an average of 0.73 (SD = 0.47) tweets and a max of four tweets in this sample, suggesting there were not users tweeting large amounts of content out, but there were users whose content was disseminated more often. Of those whose content was retweeted in this network ($n$ = 1,243), 15% were verified ($n$ = 186), their number of followers ranged from 0 to 3,604,035 (mean = 39,072.3, median = 3,235, SD = 201,406.1), and they tweeted a mean 37,878.90 tweets (median = 11,042, range = 1-1,233,437, SD = 85,048.6). Of those who retweeted content out ($n$ = 3,046), 1.1% ($n$ = 35) were verified, number of followers ranged from 0 to 582,524 (mean = 3,680.2, median = 590, SD = 22,604.3), and users tweeted an average of 56,435.7 tweets (median = 16,682.5, SD = 113,422.0, range = 1-1,621,861). See Table 2 for all sample characteristics.

### Network Descriptives

**Centrality.** In this network, a user was retweeted (in-degree) 0.73 (SD = 3.52, range = 0-108) times on average and retweeted others in the network (out-degree) an average of 0.73 (SD = 0.47, range = 0-4) times. The majority of users in this network were never retweeted (70.8%, $n$ = 2,983), with 20.6% ($n$ = 869) of the network being retweeted once, and less than 1% being retweeted more than 12 times ($n$ = 39). More than a quarter (27.5%; $n$ = 1,159) of the network never retweeted another user (and therefore only provided content to the network), 71.4% ($n$ = 3,009) retweeted one user, and 1.1% ($n$ = 47) of the network retweeted more than one user. The average betweenness score for nodes in this network was 0.05 (SD = 1.20), and the mean scores for both closeness and eigenvector centrality were less than 0.00 (SD = 0.00, 0.02, respectively).

**Group-level measures.** Group-level measures revealed a largely fragmented network in this study. A k-core analysis revealed three "cores" within the data, with 11 nodes making up the largest (and densest) core (3k-core). Finally, a community detection analysis was conducted based on edge betweenness scores. This technique splits the network into mutually exclusive groups of connected nodes, where nodes can only belong to a single community (Valente, 2010). Community detection revealed 1,175 communities with modularity scores of 0.99. The majority of communities were sized at two nodes (68.3%, $n$ = 802), with the largest community consisting of 111 nodes. The average community consisted of 3.59 nodes (SD = 6.62).

Table 2

*Sample Characteristics for 4,289 Twitter Users*

| | % | *n* | Mean | Median | SD | Range |
|---|---|---|---|---|---|---|
| **Verified** | | | | | | |
| Overall | 5.2 | 221 | | | | |
| Source | 15 | 135 | | | | |
| Retweeter | 1.1 | 35 | | | | |
| **Num. of Followers** | | | | | | |
| Overall | | | 13,937.2 | 970 | 111,220.6 | 3,604,035 |
| Source | | | 39,072.3 | 3,235 | 201,406.1 | 3,604,034 |
| Retweeter | | | 3,680.2 | 590 | 22,604.3 | 582,524 |
| **Num. of Tweets** | | | | | | |
| Overall | | | 51,057.7 | 14,855 | 106,307.9 | 1,621,860 |
| Source | | | 37,878.9 | 11,042 | 85,048.56 | 1,233,436 |
| Retweeter | | | 56,435.7 | 16,682.5 | 113,422.0 | 1,621,860 |

*Note.* Source = people whose content was retweeted out ($n$ = 1,243); Retweeter = people who retweeted content out ($n$ = 3,046); SD = standard deviation

**Network-level measures.** Similar to group-level measures, network-level measures revealed a largely disconnected (and therefore fragmented) network. The density of this network was 0.0002 and transitivity score was 0.001, suggesting an overall sparsely connected network. The centralization index for this network was 0.025. See Table 2 for all network descriptive results.

## ERGM

Retweets were significantly more likely to occur between accounts of the same verification status (PE = 0.28, $p < .01$) and shared a connection to a third account (transitivity; PE = 0.65, $p < .01$). There was also a significant in-degree popularity term, meaning accounts were more likely to retweet other accounts that had been retweeted frequently by others (PE=0.45, p<.01). Accounts were more likely to retweet others if they were not verified (PE = -0.93, $p < .01$) and tweeted more often (PE = 0.0000009, $p < .01$); however, accounts were less likely to retweet if they had more followers (PE = -0.00002, $p < .01$). On the other hand, accounts were more likely to be retweeted if they were verified (PE = 0.89, $p < .01$), had more followers (PE = 0.0000004, $p < .01$), and tweeted more often (PE = 0.0000004, $p < .01$). Table 3 provides all ERGM terms, PE, SE, and $p$-values for the presence of a retweet between accounts.

## Discussion

This study explored an interdisciplinary collaboration between Computational Informatics (the field where social media mining is housed) and Social Network Analysis. Our intent was to support the combined use of both groups of methodologies in understanding the spread of health information and how that information can potentially lead to improved health behaviors. Our example, information spread about pro-mask wearing during the COVID-19 pandemic, underlines the need to understand the structural and positional mechanisms which may promote or constrict the sharing of health-related information (SNA) within abundantly available Twitter data (social media mining). Indeed, our results supply further understanding on the

Table 3

*ERGM Results for Retweeting and Being Retweeted*

| | PE (*SE*) | *p* |
|---|---|---|
| **Structural Terms** | | |
| Edges | -9.94 (0.005) | < .01* |
| Transitivity | 0.65 (0.005) | < .01* |
| In-degree Popularity | 0.45 (0.0008) | < .01* |
| Verification Homophily | 0.28 (0.006) | < .01* |
| **Sender Covariates** | | |
| Verification | -0.93 (0.003) | < .01* |
| Followers | -0.00002 (0.000001) | < .01* |
| Tweets | 0.0000009 (0.0000001) | < .01* |
| **Receiver Covariates** | | |
| Verification | 0.89 (0.009) | < .01* |
| Followers | 0.0000004 (0.00000005) | < .01* |
| Tweets | 0.00000004 (0.00000001) | < .01* |

basic network structures related to the retweeting of pro-mask twitter users, implicate how these structures may be related to information spread, and expand the implications and applications of this interdisciplinary collaboration of research.

Overall, this network was sparsely connected, decentralized, and disjointed. Because mean scores on each of the centrality measures were low, we can assume that most users' information is only reaching a small number of connected users. This differential sharing or disjointed network of sharing has previously been noted in social media virality (Goel et al., 2015). However, based on *k*-core analyses, as well as the centralization index of this network, results suggest a select few users had a much higher influence over the network, despite being unreachable by most nodes. These more popular nodes could serve as change agents, or opinion leaders, in the network (Valente & Pumpuang, 2007), and would be important to include if attempting to spread public health information across this network.

Community detection analysis revealed a vastly fragmented network, with a near maximum modularity score and several disconnected communities within the larger network. This means that structurally, information is likely "locked in" within those fragmented communities and is unlikely to spread beyond local contacts. Thus, programmatic efforts could focus on building connections between communities, and creating opportunities for information to spread across segmented groups (Granovetter, 1985; Valente et al., 2015). It is important to structurally identify subgroups of the network that are more cohesive,

understand the key players within them, and promote positive messaging and greater connection/access to them (Valente et al., 2015). Thus, understanding what factors related to the presence of ties within this network is a first step in building more connections among users.

Transitivity and in-degree popularity were significant factors related to retweets in this sample network. Both factors speak to the structural influences which may be at play regarding the sharing of information on Twitter. A significant transitive factor underlines clustering and local community findings, which emerge around the sharing of polarized hashtag content. Others who believe similarly may create an enclosed information sharing environment prone to echo chambers (Malik & Lee, 2020). Similarly, the in-degree popularity term may implicate the social influence inherent to popular accounts or popular tweets being retweeted by many people (Riquelme & González-Cantergiani, 2016), hence providing more impressions and more opportunity for others to see pro-mask content on their feed and retweet the content. Level of popularity is also used by Twitter to share more popular content to users. This factor may accentuate the impact of these posts; however, this effect cannot be determined in the analysis used. Understanding the impact of these echo chambers and proliferation of popular tweets and hashtags would provide further implications on the dissemination of information and misinformation in a pandemic setting.

Further, several characteristics of the accounts were significantly associated with the odds of retweeting or being retweeted by other accounts. While the verification process on Twitter is called into question at times, it played an important role in determining the retweet network seen here. Verified accounts were significantly more likely to be retweeted by others while significantly less likely to retweet others. This may show a propensity for verified accounts to curate and share original content instead of retweeting others' content. Similarly, accounts were less likely to retweet pro-mask content if they had more followers. These results show that verified accounts and accounts with more followers may serve an important role in spreading these hashtags (and related content) due to their status as opinion leaders (Riquelme & González-Cantergiani, 2016; Valente & Pumpuang, 2007).

Overall, when identifying key opportunities to create connections within this network that could enhance the spread of pro-mask content, identifying central nodes who are verified and have many followers could be helpful for content generation. Further, linking less influential nodes to these central players could drive the information being sent out, seeing as the influential nodes are not likely to retweet content, but be the source of retweeted information. Finally, finding opportunities to connect smaller, disjointed communities together could results in a greater spread of information, reducing the chance of information starting and stopping within isolated groups.

## Implications for Health Behavior Theory

Herein, we used SNA theory and methods to identify key constructs related to the spread of health information across a Twitter network, and suggested opportunities for intervention based on our theoretically informed results. The insights afforded by the analyses further reinforce that when mining social media data, regardless of the research question, one should consider the additional implementation of SNA theories and frameworks to understand the underlying interconnectedness of the data. Indeed, through interdisciplinary approaches, one can

draw more-accurate and generalizable conclusions than when exploring data from only one perspective. For example, without SNA, social media mining could be limited to exploratory analyses that only analyze the data at hand. By contrast, without social media mining, SNA may be missing an important puzzle piece that adds context to the derived network connections. Though both SNA and social media mining could undoubtedly stand on their own, integrated theoretical principles only stand to strengthen the merit of findings.

From a health perspective, this can help inform key outcomes intending to promote positive health behaviors across a spectrum of health issues. We reiterate that while this interdisciplinary approach is not new, it has been less critically assessed and evaluated from a purely health behavior perspective. And, in that small gap there remain exciting opportunities to further explore the potential unity between computational informatics and SNA in directions that exclusively pertain to health behavior science. Indeed, SNA and computational informatics combined can help health behaviorists understand nuanced aspects of human behavior through the effects of social interactions on these platforms.

## Limitations

Our work is subject to limitations. First, our intentionally restrictive example was used to simply illustrate the benefit of combining SNA and social media mining frameworks. As such, the network displayed in our study only comprises retweets from one week when a series of pro-mask hashtags were particularly popular in social media spaces (Heverin & Zach, 2010). Because of the limited scope of our study, these results *do not include* anti-mask stances; an equally feverous movement on social media encouraging people to forgo mask-wearing practices. We also note that our study, as with any study using Twitter data, is subject to the limitations inherent to social media data, including a bias within key demographic information (Gore et al., 2015), and the temporal and spatial patterns that affect how information is relayed on social media (Emch et al., 2012). However, these limitations do not diminish the importance of our work. Rather, they create further opportunities to conduct more expansive studies using a combined informatics and SNA approach that expand on our original study and move into new domains entirely. Future research should consider expanding the parameters of our original study by including a broader date range, additional hashtags that represent counter-mask movements, and other forms of gauging interaction on Twitter (e.g., mentions, impressions, likes). Additionally, future research should continue to push the methodological boundaries between these new fields by exploring deeper hypothesis-driven and theory-guided research questions about health behaviors expressed on social media, including added content analysis components such as topic models (Blei et al., 2003) and sentiment analysis (Hutto & Gilbert, 2014).

## Discussion Questions

1. What are the implications for engaging in social media research without a pre-determined theoretical framework to guide the study?
2. How similar/different would an anti-mask network analysis be? What are the underlying reasons for those similarities and/or differences?
3. How can a content analysis of our dataset strengthen the findings of the study?

### Acknowledgments

The authors have no conflicts of interest to report.

## References


Ahmed, W., Vidal-Alaball, J., Lopez Segui, F., & Moreno-Sánchez, P. A. (2020). A social network analysis of tweets related to masks during the COVID-19 pandemic. *International Journal of Environmental Research and Public Health*, *17*(21), 8235. https://doi.org/10.3390/ijerph17218235

Bento, A. I., Nguyen, T., Wing, C., Lozano-Rojas, F., Ahn, Y.-Y., & Simon, K. (2020). Evidence from Internet search data shows information-seeking responses to news of local COVID-19 cases. *Proceedings of the National Academy of Sciences*, *117*(21), 11220–11222. https://doi.org/10.1073/pnas.2005335117

Blei, D. M., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2018). *Analyzing social networks*. Sage. Retrieved August 28, 2020, from http://journals.openedition.org/lectures/24709

Boulay, M., & Valente, T. W. (2005). The selection of family planning discussion partners in Nepal. *Journal of Health Communication*, *10*(6), 519–536. https://doi.org/10.1080/10810730500228789

Bueno, N. P. (2015). Are opinion leaders important to spread information to cope with extreme droughts in (all) irrigation systems? A network analysis. *Scientometrics*, *105*(2), 817–824. https://doi.org/10.1007/s11192-015-1734-z

Centola, D. (2013). Social media and the science of health behavior. *Circulation*, *127*(21), 2135–2144. https://doi.org/10.1161/CIRCULATIONAHA.112.101816

Cho, S. E., Jung, K., & Park, H. W. (2013). Social media use during Japan's 2011 earthquake: How Twitter transforms the locus of crisis communication. *Media International Australia*, *149*(1), 28–40. https://doi.org/10.1177/1329878X1314900105

Csárdi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems, 1695*(5), 1-9.

Eikenberry, S. E., Mancuso, M., Iboi, E., Phan, T., Eikenberry, K., Kuang, Y., Kostelich, E., & Gumel, A. B. (2020). To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic. *Infectious Disease Modelling*, *5*, 293–308. https://doi.org/10.1016/j.idm.2020.04.001

Emch, M., Root, E. D., Giebultowicz, S., Ali, M., Perez-Heydrich, C., & Yunus, M. (2012). Integration of Spatial and Social Network Analysis in Disease Transmission Studies. *Annals of the Association of American Geographers, 105*(5), 1004–1015. https://doi.org/10.1080/00045608.2012.671129

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239. https://doi.org/10.1016/0378-8733(78)90021-7

Goel, S., Anderson, A., Hofman, J., & Watts, D. J. (2015). The structural virality of online diffusion. *Management Science*, *62*(1), 180–196. https://doi.org/10.1287/mnsc.2015.2158

Gore, R. J., Diallo, S., & Padilla, J. (2015). You are what you tweet: Connecting the geographic variation in America's obesity rate to Twitter content. *PloS One*, *10*(9), e0133505. https://doi.org/10.1371/journal.pone.0133505

Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. *American Journal of Sociology*, *91*(3), 481–510.

Hambrick, M. E. (2012). Six degrees of information: Using social network analysis to explore the spread of information within sport social networks. *International Journal of Sport Communication*, *5*(1), 16–34. https://doi.org/10.1123/ijsc.5.1.16

Handcock, Mark S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N., Bender-deMoll, S., & Morris, M. (2019). Package 'statnet'. https://cran.r-project.org/web/packages/statnet/statnet.pdf

Heverin, T., & Zach, L. (2010). Twitter for city police department information sharing. *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1–7. https://doi.org/10.1002/meet.14504701277

Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216-225. Article 1. https://ojs.aaai.org/index.php/ICWSM/article/view/14550

Karami, A., Shah, V., Vaezi, R., & Bansal, A. (2020). Twitter speaks: A case of national disaster situational awareness. *Journal of Information Science*, *46*(3), 313–324. https://doi.org/10.1177/0165551519828620

Klovdahl, A. S. (1985). Social networks and the spread of infectious diseases: The AIDS example. *Social Science & Medicine*, *21*(11), 1203–1216. https://doi.org/10.1016/0277-9536(85)90269-2

Lusher, D., Koskinen, J., & Robins, G. (Eds.). (2013). *Exponential random graph models for social networks: Theory, methods, and applications* (Vol. 35). Cambridge University Press.

Malik, P., & Lee, S. (2020). Follow me too: Determinants of transitive tie formation on Twitter. *Social Media + Society 6*(3), 1-12. https://doi.org/10.1177/2056305120939248

Moon, S. J., & Hadley, P. (2014). Routinizing a new technology in the newsroom: Twitter as a news source in mainstream media. *Journal of Broadcasting & Electronic Media*, *58*(2), 289–305. https://doi.org/10.1080/08838151.2014.906435

Nelon, J. L., Moscarelli, M., Stupka, P., Sumners, C., Uselton, T., & Patterson, M. S. (2020). Does scientific publication inform public discourse? A case study observing social media engagement around vaccinations. *Health Promotion Practice*, 1524839919899925. https://doi.org/10.1177/1524839919899925

Patterson, M. S., Gagnon, L. R., Vukelich, A., Brown, S. E., Nelon, J. L., & Prochnow, T. (2019). Social networks, group exercise, and anxiety among college students. *Journal of American College Health*, 1–9. https://doi.org/10.1080/07448481.2019.1679150

Patterson, Megan S., Prochnow, T., & Goodson, P. (2019). The spread and utility of social network analysis across a group of health behavior researchers. *Health Behavior*

*Research*, *2*(4). https://doi.org/10.4148/2572-1836.1063

Peeples, L. (2020). Face masks: What the data say. *Nature*, *586*(7828), 186–189. https://doi.org/10.1038/d41586-020-02801-8

Prochnow, T., Patterson, M. S., & Hartnell, L. (2020). Social support, depressive symptoms, and online gaming network communication. *Mental Health and Social Inclusion*, *24*(1), 49–58. https://doi.org/10.1108/MHSI-11-2019-0033

Riquelme, F., & González-Cantergiani, P. (2016). Measuring user influence on Twitter: A survey. *Information Processing & Management*, *52*(5), 949–975. https://doi.org/10.1016/j.ipm.2016.04.003

Simms, T., Ramstedt, C., Rich, M., Richards, M., Martinez, T., & Giraud-Carrier, C. (2017). Detecting cognitive distortions through machine learning text analytics. *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, 508–512. https://doi.org/10.1109/ICHI.2017.39

Sobowale, K., Hilliard, H., Ignaszewski, M. J., & Chokroverty, L. (2020). Real-time communication: Creating a path to COVID-19 public health activism in adolescents using social media. *Journal of Medical Internet Research*, *22*(12), e21886. https://doi.org/10.2196/21886

Valdez, D., ten Thij, M., Bathina, K., Rutter, L. A., & Bollen, J. (2020). Social media insights into US mental health during the COVID-19 pandemic: Longitudinal analysis of Twitter data. *Journal of Medical Internet Research*, *22*(12), e21418. https://doi.org/10.2196/21418

Valente, T. W. (2005). Network models and methods for studying the diffusion of innovations. In P. J. Carrington, J. Scott, & S. Wasserman (Eds.), *Models and methods in social network analysis* (pp. 98–116). Cambridge University Press. https://doi.org/10.1017/CBO9780511811395.006

Valente, T. W. (2010). *Social networks and health: Models, methods, and applications*. Oxford University Press.

Valente, T. W. (2012). Network interventions. *Science, 337*(6090), 49-53. https://doi.org/10.1126/science.1217330

Valente, T. W., Fujimoto, K., Soto, D., Ritt-Olson, A., & Unger, J. B. (2013). A comparison of peer influence measures as predictors of smoking among predominately Hispanic/Latino high school adolescents. *Journal of Adolescent Health*, *52*(3), 358–364. https://doi.org/10.1016/j.jadohealth.2012.06.014

Valente, T. W., Palinkas, L. A., Czaja, S., Chu, K.-H., & Brown, C. H. (2015). Social network analysis for program implementation. *PLOS ONE*, *10*(6), e0131712. https://doi.org/10.1371/journal.pone.0131712

Valente, T. W., & Pumpuang, P. (2007). Identifying opinion leaders to promote behavior change. *Health Education & Behavior*, *34*(6), 881–896. https://doi.org/10.1177/1090198106297855

Wang, X., Yu, C., & Wei, Y. (2012). Social media peer communication and impacts on purchase intentions: A consumer socialization framework. *Journal of Interactive Marketing*, *26*(4), 198–208. https://doi.org/10.1016/j.intmar.2011.11.004

Xu, G., & Li, L. (2013). *Social Media Mining and Social Network Analysis: Emerging Research: Emerging Research*. IGI Global.