Identifying preferred routes of sharing information on social networks

Rozhin Mohammadikian[1,*], Parsa Bigdeli[1], Behrouz Askari[1], and G.Reza Jafari[1,2,3,+]

[1]Department of Physics, Shahid Beheshti University, Evin, Tehran 1983969411, Iran
[2]Center for Communications Technology, London Metropolitan University, London N7 8DB, UK
[3]Chandigarh Group of Colleges Jhanjeri, Sahibzada Ajit Singh Nagar, Punjab, India
[*]rozhinmkian99@gmail.com
[+]g_jafari@sbu.ac.ir

ABSTRACT

The spread of information has become faster and wider than ever with the advent of social network platforms. The question raised in this study is whether information dissemination in social networks is random or follows a discernible structure. Our results from real-world hashtag data suggest that the spread of hashtags is not random and follows specific patterns. This study proposes two preferential models to explore how news spreads on social media. Specifically, we examine global and local preferential selection models and demonstrate that information dissemination aligns with these patterns. According to these two models, information flows are distributed through specific paths on networks. This suggests that new information tends to propagate along the same paths as previous news, with the specific pathways varying depending on the type of content. Finally, an examination of the propagation of political hashtags on Twitter confirms the existence of these paths that also emerge from the two preferential models.

1 Introduction

The flow of information between users influences not only their knowledge but also their decision-making processes and judgment. Today, a large part of this information flow is carried out through social networks, which have become an inseparable part of daily life. From social marketing[1] and preventing the spread of misinformation[2], to predicting political elections[3–5] and natural disaster alarm[6,7], understanding the mechanisms of information propagation within cyberspace has consequently become an important area of multidisciplinary research. Although users can theoretically connect on most contemporary online social platforms, these connections are far from uniform. Pre-existing offline social networks are the primary source of this heterogeneity, largely influenced by the geographical distribution of users[8,9]. However, additional factors, such as the online visibility of certain users (e.g., celebrities, political figures, or media companies) or shared interests that connect individuals regardless of their offline proximity or relationship, can create differences between cyberspace networks and their offline counterparts. Moreover, in both networks, connections are utilized with varying frequency depending on the topic of communication. A person may be more likely to share messages or engage with certain individuals on specific subjects while turning to different contacts for other subjects. This tendency forms a set of preferences in content sharing for each user, referred to as the underlying diffusion network[10], which are considered to be topic-dependent. The news spread can then be modeled as a biased random walk on this topic-sensitive underlying network[11].

Understanding user preferences can help predict how news spreads on a given social networking platform, enabling applications such as targeted marketing and fake news prevention. Numerous studies have investigated various aspects of this phenomena[12], ranging from theoretical modeling[13] and measuring[14] information diffusion, to implementing computational and machine learning methods that could be used for practical use[15,16]. However, the literature reveals a notable gap in studies focused on modeling the emergence of user preferences. In this study, following a review of the phenomenology of news sharing, we propose that the formation of individual preferences can be governed by two distinct dynamics, which we refer to as the global and local preferential models. These models are inspired by the seminal work of Barabási and Albert[17], which will generate an underlying network that contains sharing preferences. Sharing of information is then modeled as a biased random walk on this underlying network. We then incorporate two measures, namely a modified weighted Jaccard index and functional similarity, to evaluate the presence of user preferences in news propagation over the networks generated by these models. As a case study, we apply our framework to political hashtags from the X platform (formerly Twitter) to demonstrate and validate our approach.

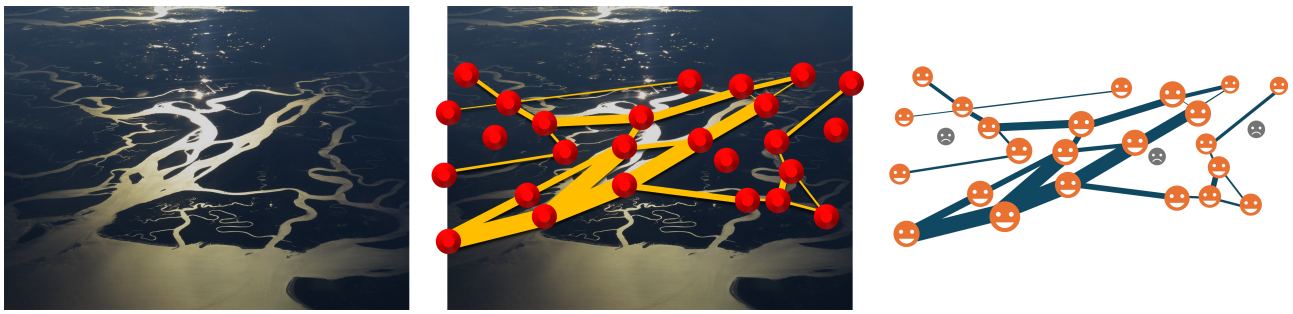


Figure 1. One case of preferred routes in nature is riverbeds[18], which form over time after many repetitions. The more water streams down a path, the more indented the riverbed will become. We take this idea to explain the formation of preferred routes of information sharing on social media.

2 Mathematical Model

To construct the model, we begin by examining empirical studies on the motivations and dynamics of social interactions. Rather than capturing the full behavioral complexity, we identify and incorporate key mechanisms that are most relevant for formal modeling. Social and group living has long been recognized as a survival strategy among many species, including humans, providing advantages in reproduction, offspring care, foraging, and predator avoidance[19]. In the modern context, however, the motivations for human social interaction have evolved beyond basic survival and reflect higher-order psychological needs. This shift is particularly evident in the use of social media. Wong et al.[20] argue that content sharing on social networks is primarily driven by the desire to inform and entertain an audience but also stems from deeper motivations such as maintaining relationships, influencing others, asserting individuality, and affiliating with social groups. According to their findings, social media engagement is largely shaped by an intrinsic need to share. They also mention that sharing content is influenced by three additional factors: (i) the recipient, (ii) the relationship between them, and (iii) the content, all of which we aim to incorporate into our model. The recipient and the strength of the relationship between the recipient and the sender determine the frequency—or equivalently, the probability—by which the content is shared. The strength of this relationship can arise from multiple causes, developing over time either internally through social media interactions or externally through offline connections.

We aim to model the former mechanism using our proposed framework, while treating the latter as a predefined input parameter, since external social connections often form long before an individual is eligible to use social media. Although these external relationships can evolve after joining a platform, we assume such changes are relatively slow and can be neglected in our dynamic formulation. Overall, the key point is that these relationships are both user-specific and time-dependent. As highlighted in the third point above, the strength of a relationship can vary depending on the content being shared. According to Berger et al.[21], sharing practically useful information is motivated either by altruistic intentions, such as helping others, or by the desire to enhance one's social image, for example, to appear knowledgeable to colleagues. Picone et al.[22] reinterpret this latter motivation as a form of self-publication, drawing an analogy to journalism in which individuals prioritize the interests of their audience and selectively share topics aligned with those interests. This perspective supports the assumption that each social media user has context-dependent sharing preferences. For instance, a user may exhibit distinct sharing behaviors depending on whether the content is political, entertainment-related, or scientific. Consequently, the network of shared content can be decomposed into topical multiplexes that evolve independently and simultaneously. This concept is known as *thematic multiplex*[23].

Lastly, we also take into consideration the fact that some users are more prominent due to various reasons, such as being a celebrity or a political figure. Such prominence increases their visibility within the network and makes them more likely to be involved in social interactions[24]. It is important to clarify that when referring to the roles of sender and recipient, we mean the temporal sequence in which individuals are exposed to a given piece of content. These roles can manifest differently across platforms. For instance, in messaging systems like email or direct messaging features of platforms such as Instagram, the sender possesses the content before the recipient and decides whether or not to share it. In contrast, on microblogging platforms such as X, the user who originally posts the content (e.g., the one being retweeted) is the sender, and the one who retweets it is the recipient, even though in this case, it is the recipient who actively decides to share it.

2.1 Global Preference Evolution Dynamics

We consider a base network G, modeled as a weighted, undirected graph of size N, where nodes represent users. The degree k of a node quantifies its prominence, and the edge weight L between two nodes, an integer, represents the number of social interactions (termed *events*) between them. Following the framework of Rabbani et al.[25], the dynamics proceed by selecting a sender node at random to initiate an event, capturing user activity. The sender is more likely to interact with a recipient of higher node degree, reflecting the increased prominence and hence visibility of such users. Once the recipient is chosen, the interaction is recorded by incrementing the corresponding edge weight by x, which is typically set to one when following the definition of the number of interactions, but it can also be set to other values to emphasize the influence of interactions. Accordingly, the probability of a node participating in an event, either as the sender or the recipient, at time t_n is:

$$\Pi_{k_i}(t_n) = \frac{1}{N} + \frac{k_i(t_n)}{\sum_{i=1}^{N} k_i(t_n)} \tag{1}$$

If m nodes participate at each time step, the probability of event participation of node i becomes $m\Pi_{k_i}$. In the continuum limit of time, the evolution of each node's weighted degree follows:

$$\frac{dk_i(t)}{dt} = \frac{m}{N} + \frac{mk_i(t)}{2mt + m_0} \tag{2}$$

Here, m_0 denotes the total initial weights of the links, and $2mt$ accounts for the cumulative increase in total node degrees over time. This dynamic is inspired by the Barabási-Albert (BA) preferential attachment model but differs in some aspects. Unlike the BA model, our network maintains a fixed number of nodes. Moreover, on the occurrence of each social event, it is the weight of the corresponding link that changes explicitly, not the number of neighbors of the selected nodes, although the change in the weight of the links consequently results in changing the weighted degrees of each node.

To compute the change in the weight of a link, we consider the probability that an event occurs between nodes i and j, accounting for both cases in which either node may act as the sender and the other as the recipient. Assuming that each node is equally likely to initiate an event, the combined probability of interaction between them is given by the average of the two directional probabilities. Each of these probabilities is proportional to the prominence (i.e., degree) of the initiating node and normalized over the total degree in the system. Thus, the interaction probability becomes:

$$\begin{aligned} \Pi_{L_{ij}}(t) &= \frac{1}{2}\left(\frac{1}{N} \times \frac{k_i(t)}{(2mt + m_0)} + \frac{1}{N} \times \frac{k_j(t)}{(2mt + m_0)}\right) \\ &= \frac{(k_i(t) + k_j(t))}{2N(2mt + m_0)} \end{aligned} \tag{3}$$

When m interactions occur at each time step, this leads to the following rate of change for the link weight in the continuum limit:

$$\frac{dL_{ij}}{dt} = \frac{m(k_i(t) + k_j(t))}{2N(2mt + m_0)}. \tag{4}$$

This result indicates that the rate of increase in a link's weight is directly proportional to the average prominence of its two end nodes.

2.2 Local Preference Evolution Dynamics

In the evolution dynamics of global preference, each node evaluates the prominence of all other nodes based on their weighted degrees, which are assumed to be globally visible across the network. This allows a node to initiate interaction with a highly prominent node even in the absence of prior connections. However, in many real-world settings, social interactions are influenced by shared history between individuals. To account for this, we introduce an alternative mechanism, which we refer to as the *local preference* model.

While in the global preference model, a node's decision to interact was based on the degrees of other nodes, in the local preference model, this decision depends on the existing link weights. If node i is randomly selected as the sender at time step t_n, the probability of its interaction with node j is proportional to the weight of their link, $L_{ij}(t_n)$, normalized by the sum of the weights of all links connected to node i, i.e., $L_{ij}(t_n)/k_i(t_n)$. Here, $k_i(t_n)$ denotes the weighted degree of node i, defined as the sum of the weights of its links to neighboring nodes. To account for events initiated by either node i or node j, we symmetrize the probability by averaging the two possibilities. The resulting marginal probability that a link is selected for an event is:

$$\Pi_{L_{ij}}(t_n) = \frac{L_{ij}(t_n)}{2N}\left(\frac{1}{k_j(t_n)} + \frac{1}{k_i(t_n)}\right) \tag{5}$$

Due to the nature of this dynamic, which dictates that the evolution of each link's weight is proportional to itself, all links with zero weight will remain zero indefinitely. This is inconsistent with real-world social networks, where any two individuals can potentially connect, albeit mostly with low probability. To address this, we add a small constant L_0 to all possible links in the network, ensuring weak but nonzero interaction probabilities. This approach resembles the modification introduced by Price et al.[26], a precursor to the BA model. We choose $L_0 \ll x$, where x the link weight increment is (e.g., for $x = 1$, we set $L_0 = 0.01$). Consequently, the network becomes effectively fully connected. For notational simplicity, we redefine the adjusted link weights as $L'_{ij} := L_{ij} + L_0$ for all $i, j \in G$. This leads to the following rate of change for the link weight in the continuum limit:

$$\frac{dL'_{ij}}{dt} = \frac{L'_{ij}(t_n)}{2N}\left(\frac{1}{k_j(t_n)} + \frac{1}{k_i(t_n)}\right). \tag{6}$$

Accordingly, the probability of node i being a participant of an event, either as the sender or the recipient, is:

$$\begin{aligned}\Pi_{k_i}(t_n) &= \frac{1}{2}\left(\frac{1}{N}\times 1 + \frac{1}{N}\times\frac{L_{ij}(t_)}{k_j(t_n)} + \frac{1}{N}\times\frac{L_{il}(t_n)}{k_l(t_n)} + ...\right)\\ &= \frac{1}{2}\left(\frac{1}{N} + \langle\frac{L_{ij}(t_n)}{k_j(t_n)}\rangle_i\right).\end{aligned} \tag{7}$$

The $\langle\frac{L_{ij}(t_n)}{k_j(t_n)}\rangle_i$ term refers to the average of $\frac{L_{ij}(t_n)}{k_j(t_n)}$ all j; naturally, those that are not a neighbor of i will have a link weight of zero. Eq. 7 shows that in the local preference model, the probability of a node participating in an event as the recipient (the second term in parentheses) is determined by how favored the node is in the eyes of its neighbors, while the probability of being chosen as the sender (the first term in parentheses) remains the same for all nodes.

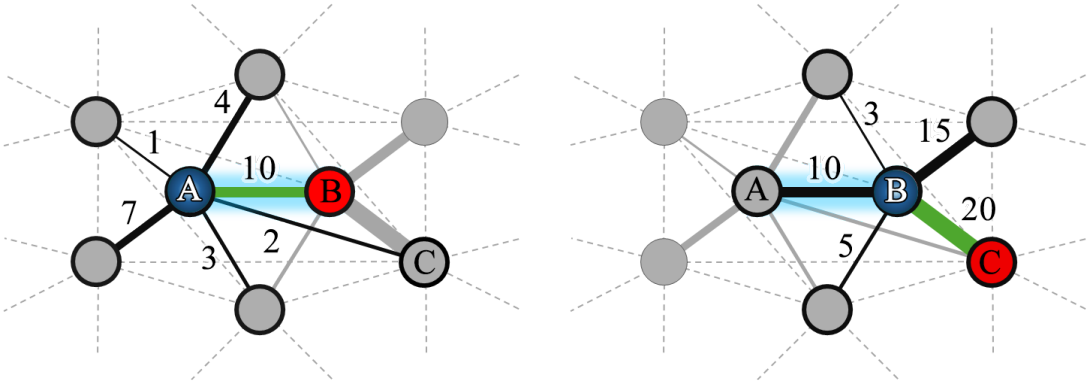
In the global preference model, all nodes determine their interaction preferences based on a common measurement, which is the degree of other nodes. As a result, node preferences are largely homogeneous across the network. The only source of variation arises from the constraint that self-loops are prohibited, meaning that no node can include itself in its own preference list. In contrast, the local preference model can introduce heterogeneity. Here, each node's preferences are shaped by the weights of its existing links, which is not necessarily the same as those of other nodes.

Also note that, for both models, the same link weight may lead to asymmetric preference decisions for the end nodes of the link; e.g., even if node B is the priority preference of node A, node A is not necessarily the priority preference of node B. This asymmetry in preference arises despite the underlying network being undirected and the adjacency matrix being symmetric. In practice, while the network topology is symmetric, the induced preference structure is not (see Fig. 2). This observation justifies the use of an undirected adjacency matrix, which offers computational efficiency, even when modeling phenomena that exhibit inherently directed behavior.

3 Route Preference In Information Flow

We propose that the evolution of social networks can follow a mechanism incorporating both the global and local dynamics introduced in Sections 2.1 and 2.2. However, it is not possible to directly verify this through comparing the networks that the models generate with the real-world underlying preference network for direct validation of this hypothesis, as the latter is not accessible to us; at most, we can acquire a part of it (e.g. the follower-followee network, or friendship network). But even this will not suffice; indeed, for most *opinion formation* social interactions, social threads often involve people who don't follow each other[27]. Furthermore, studies have shown that additional factors, including the writing style and sentiment of shared content, significantly influence interaction patterns[28,29]. As a result, the follower-followee structure becomes a poor predictor of the likelihood of interaction. For example, on the X platform (formerly Twitter), only approximately 10% of retweet interactions involving hashtags occur between users who are connected through a follower-follower relationship[30]. Instead, what is readily observable is the flow of events between users, which we will acquire from the real world social networks.

In conclusion, we assume that there is an effective underlying network, to which we do not have access, and that observable social events are spread by the preferences embedded within this network, analogous to the trajectory of a biased random walker on a weighted network. If the underlying network is considerably preferential, then we expect the flow of information (corresponding to the walk of the random walker) to be carried out on preferred routes. As discussed in Section 2, these dominant routes are expected to vary across contexts, but within each context, the pieces of information will most likely spread on the same routes, rather than diffusing randomly on the network.



(a) The probability of traversing the undirected link A-B when the sender is node A is approximately 37%. Here, the preferred choice of node A is node B and not node C.

(b) The probability of traversing the undirected link A-B when the sender is node B is approximately 19%. Here, the preferred choice of node B is node C and not node A.

Figure 2. The weight of the undirected link A-B can render different preferences depending on the sender node, hence we are able to model a phenomenon that is inherently directed using an undirected network.

3.1 Real World Data Retrieval

As an example case, we used the hashtag retweet data from the Farsi X with political contexts[31]. These data were acquired by Mohammadi et al.[32] and further studied by Bigdeli et al.[33], spanning from April 29, 2021, to June 24, 2021 —7 weeks prior to Iran's 2021 presidential election. The trending hashtags were tracked daily, and 16 of them were used in this study. For inclusiveness, the hashtags chosen consisted of those specific to each of the two opposing parties and also included neutral hashtags used by both groups. Data retrieval was done such that every tweet containing a specific hashtag from a specified date was stored in an SQLite database, accessed through the X's Standard Search API. More specifically, each tweet and retweet was saved along with the user ID of the retweeter and the user ID of the user being retweeted, along with the hashtags used in the body of the retweet and a timestamp of the retweet happening. The number of distinct retweets containing at least one of the election hashtags was 5701902, and they were posted by 140638 unique users collected through X platform's data, which at the time of collection was named Twitter.

A weighted, undirected retweet network was constructed from the data gathered for each hashtag, where nodes are the users and the link weights are the number of times a tweet containing that Hashtag has been shared between two users, in either directions. This retweet network can be seen as the trajectory of a biased random walk on the underlying network of preferences discussed previously in this section. Note that the choice of an undirected retweet network will not pose any problem for our analysis, as discussed in section 2.2. To further study the existence of preference, the two used measures will require us to perform pre-processing on the raw retweet data, the details of which are presented in Section 3.2.3.

3.2 Measures for Detecting Preference on Network

To assess the presence of preference in the underlying network structure, we require quantitative measures to extract meaningful patterns from the observable information flow. Since the underlying preference network is inaccessible, we instead examine the structure of social interactions, such as information propagation events, as indirect indicators of these hidden preferences. Specifically, we employ two complementary metrics: (1) a modified weighted Jaccard index, which quantifies the overlap in communication paths while accounting for interaction weights, and (2) a functional similarity measure, which captures the consistency in the influence of nodes during information spread. Together, these measures allow us to infer whether information flows consistently along preferred routes, revealing latent structural biases in the network.

3.2.1 Modified weighted Jaccard Index

Similar to the desire paths already introduced in the social sciences[34], we would like to study the overlap of the imprint of social interactions between social network users. If there exists preferred routes for the dissemination of news of a certain topic, then all hashtags relating to that topic will follow that route more or less. If the tweet is considered as a biased random walker, the footprints of the walkers corresponding to each hashtag must have a considerable overlap with those of another hashtag. Due to the inherent probabilistic nature of news dissemination, some randomness is inevitable, and non-preferred links will also be traversed. However, what is critical is the relative frequency with which preferred links are traversed as opposed to non-preferred ones. To identify and compare the preferred dissemination routes for a given topic, we quantify how often each link associated with a hashtag is traversed (i.e., the number of times news is shared through that link), as illustrated in Fig. 3.

The typical [unweighted] Jaccard index gives the ratio of the overlap of two sets to the union of them (i.e. $J(A,B) = |A \cap B|/|A \cup B|$), which has already been a staple in studying social networks[35,36]. Naturally, no repetition is considered if we want to calculate the Jaccard index of two multisets (a.k.a. sets but with arbitrary repetition of elements). Therefore, the unweighted Jaccard index will only take into account the links traversed without considering the number of repetitions on that link. The weighted Jaccard index[37] is a solution to measuring the overlap of multisets, where the frequency of the news sharing between each pair of users is taken into account. In order to embed the intuition behind the desire paths and the depth of footprints that retweets leave behind, we use a modification of this measure, which essentially gauges the percentage of all the steps taken in the sharing of two hashtags that have traversed mutual links. For two hashtags with adjacency matrices $\mathbf{M_1}$ and $\mathbf{M_2}$, this could be calculated as follows:

$$\tilde{J}_w(\mathbf{M_1},\mathbf{M_2}) = \frac{\sum_{i,j}\left[\left((M_{1,ij}+M_{2,ij}\right)H(M_{1,ij})H(M_{2,ij})\right]}{\sum_{i,j}\left[M_{1,ij}+M_{2,ij}\right]} \tag{8}$$

where $H(x)$ is the Heaviside function and $M_{1,ij}$ and $M_{2,ij}$ are the ijth component of $\mathbf{M_1}$ and $\mathbf{M_2}$ respectively. If these matrices are not of the same dimensions, which is often the case with real-world data, the data must undergo a pre-processing step, discussed in Section 3.2.3.

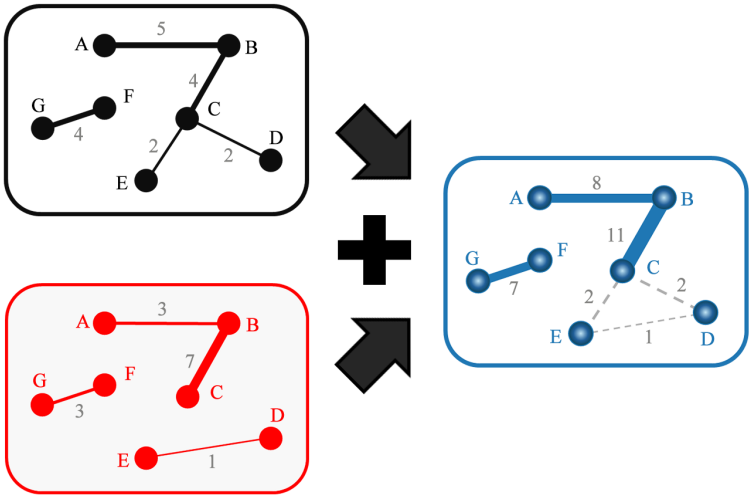


Figure 3. Overlap of retweet imprints of two hashtags: the red (lighter color) and the black (darker color) graphs on the left represent the sharing frequency of two distinct but contextually-similar hashtags between mutual users. The solid lines on the right show the mutual links used in the spread of both hashtags, while the dotted links represent those links used only for the sharing of one hashtag. For the example case of above, the modified weighted Jaccard index is calculated as $\tilde{J}_w(A,B) = \frac{8+7+11}{8+7+11+2+2+1} \approx 0.84$.

3.2.2 Functional Similarity

To further investigate the presence of preference within the underlying networks of contextually similar hashtags, we focus on the behavior of individual nodes when exposed to messages of related topics. That is, we aim to determine which nodes a given node is most likely to share a received message with. A suitable metric for this purpose is the cosine similarity[38,39]. Generally, the cosine similarity between two normalized vectors is given by their inner product, which quantifies the degree of alignment between the vectors in their respective vector space. When these vectors represent the sharing behavior of a specific node across two different networks, the cosine similarity reflects how similarly that node functions in those networks, hence also called the *functional similarity*. For the weighted, undirected network corresponding to hashtag m_1, represented by adjacency matrix $\mathbf{M_1}$,

the retweet state vector of node i is constructed as described in Eq. 9, and normalized by the norm of the space (inner product),

$$\left|p_i^{(m_1)}\right\rangle = \frac{1}{\sqrt{\sum_{j=1}^{N} M_{1,ij}^2}} \begin{pmatrix} M_{1,i1} \\ M_{1,i2} \\ \vdots \\ M_{1,iN} \end{pmatrix} \tag{9}$$

where $M_{1,ij}$ shows the frequency of retweets between node i and node j in hashtag m_1. The functional similarity of a node in the dissemination of two hashtags m_1 and m_2 with adjacency matrices $\mathbf{M_1}$ and $\mathbf{M_2}$ is then determined by

$$S_i^{m_1,m_2} = \left\langle p_i^{(m_1)} \middle| p_i^{(m_2)} \right\rangle = \frac{\sum_{j=0}^{N} M_{1,ij} M_{2,ij}}{\sqrt{\sum_{j=1}^{N} M_{1,ij}^2 \sum_{j=1}^{N} M_{2,ij}^2}} \tag{10}$$

Averaging over functional similarities of multiple nodes $\langle S_i^{m_1,m_2} \rangle$ results in an overall measurement of the existence of preference. Note that the dimension of the state vector $|p_i^{(m_1)}\rangle$ is by default the size (number of users) of the retweet network of hashtag m_1. However, most of the time modifications must be made for Eq. 10 to be applicable, which we will discuss in section 3.2.3. Additionally, the choice of whether to pick the ith row of an adjacency matrix or its ith column to construct the state vectors used in Eq. 10 is naturally irrelevant for undirected networks, such as the case of this study.

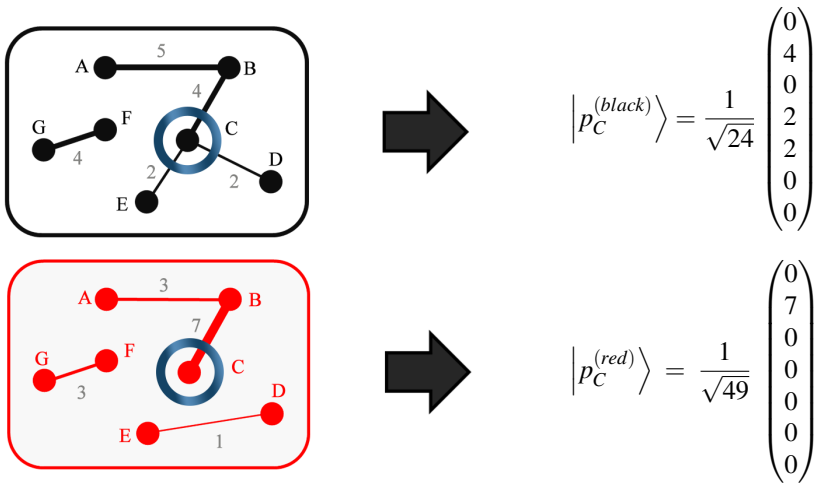


Figure 4. Retweet imprints of a single node in two hashtags: In sharing two different hashtags, a node of interest (marked with gradient color) can act with partial similarity, calculated by Eq 10. The functional similarity of the node C in the above example will be $S_C^{black,red} = \langle p_C^{(black)} | p_C^{(red)} \rangle \approx 0.82$. One way of measuring the overall similarity of two hashtags is to take the average of the functional similarities of all their nodes.

Similar measures based on inner product have been used for the study of retweet networks, such as *favoritism* of nodes[27,40]. However, favoritism is measured for a single node over one set of retweets, whereas functional similarity studies a single node over two sets of retweets (the difference might be analogized to the difference between autocorrelation and cross-correlation of series). The closer the functional similarity of a node is to unity over two hashtags, the more preferential the underlying network is.

3.2.3 Considerations

To apply these measures of preference to real-world data, we should make two considerations:

1. Two hashtag datasets collected over similar time spans will rarely produce retweet network adjacency matrices of the same size; that is, the number of users involved in the sharing of each hashtag can differ significantly. However, to apply

the modified weighted Jaccard index and functional similarity measures (Eqs. 8 and 10), the underlying networks must be of equal size. To assimilate the dimensions of two retweet networks, we can either expand their nodes to include the union of their users by adding isolated nodes or shrink them to sub-networks of them that include only the intersection of their users. The only difference is that the former will have additional zero terms in the summations of these measurements. The latter will eliminate the users who had only participated in the propagation of one of the two hashtags. In fact this makes our study more sensible, since our goal is to find whether a user will share similar topics from the same old route or not. So the user should participate in sharing both of them in the first place before we can even begin to study the overlap or similarity of the propagation routes.

2. Discrepancy between the total number of retweets (i.e. the total number of walks of the biased random walker on the underlying network) of two hashtags can alter the results obtained from Eq. 8. To avoid this unwanted effect, we normalize the weights of all the links (i.e. the components of the adjacency matrix) to the total number of weights. This ensures that variations in $\tilde{J}_w$ only reflect the strength and the structure of sharing preferences between nodes rather than their overall activities. For functional similarity, no such normalization is required beyond what is already performed in Eq. 9, as the state vectors are already normalized.

3.3 Simulation

Before performing the simulation, we note that some hashtags contain mutual retweets, meaning individual retweets may include multiple hashtags simultaneously. Including these retweets in the analysis of preference introduces sampling bias, as such retweets artificially inflate the similarity measures due to repetition. To prevent this bias, we divide each hashtag's retweet data into two temporal halves (the first and second halves of the 7-week period) and compare each pair of hashtags using data from different halves. This approach doubles the number of compared pairs from 120 to 240, with two data pairs for each hashtag pair corresponding to the differing temporal halves.

For the simulation, two initial underlying networks of $N = 400$ were evolved under two rules: the global preference dynamic (section 2.1) and the local preference dynamic (section 2.2). For the initial network of the global preference model, $m_0 = 5$ links with weights of unity were randomly added to a weighted, undirected null network, while for the local preference model, the initial network was a fully connected network with link weights of unity. These two networks evolved independently to generate the underlying network on which we later simulate biased random walks, corresponding to the propagation of hashtags. We selected an arbitrary evolution time for each case, guided primarily by the need to obtain meaningful results in the subsequent analysis. We chose $t = 300$ for the global preference model and $t = 500$ for the local preference model. At each timestep, $m = 3$ links were chosen to be increased by $x = 10$ units. The choice of the parameter x is relative to the initial weight of the links. As a null hypothesis representing the absence of sharing preferences, we also considered a third underlying network, fully connected, with homogeneous link weights. In this network, no node has any preferential priority over others.

With an adiabatic assumption that the changes of the preferences of the underlying networks happens on a greater timescale than the typical time between the rise and decay of hashtags[41], we assume the changes in individual preferences caused by the sharing of news of specific single hashtags (for example, in a time span of 7 weeks, are negligible compared to the overall time span of users' presence, which has a median of 3-5 years[42]. With that in mind, we simulated hashtag retweets on the mentioned three networks without the retweets affecting the preferences of nodes (i.e., the weight of the links is not changed). The retweet simulation is essentially a biased random walk starting from an initial node and spreading based on each node's preferences. In other words, the probability p of the random walker jumping to a neighboring node is proportional to the weight of the links of the node it is on.

Another assumption discussed in Section 2 is that hashtags with similar topics disseminate on the same underlying network. Therefore, we repeated the simulation of the biased random walks on each of the three underlying networks for a particular number of times (128 times, chosen so that 8 repetitions were accounted for the 16 corresponding hashtags). For each repetition, we added noise to the jumping probabilities from each node. We obtained the noisy jumping probabilities with a noise intensity η from a simple mapping:

$$f : [0,1] \to [0,1], \quad \forall \eta \in [0,1] : \quad f(p) = p(1-\eta) + \frac{1}{N-1}\eta \tag{11}$$

This mapping is such that for $\eta = 0$ the probabilities of jumping are strictly the same as the probabilities induced by the underlying network, and for $\eta = 1$, the probabilities are strictly equal with no preference. Also note that the $N-1$ in the denominator accounts for a possibility of traverse to all the nodes of the network except the one the walker is standing on. For each of the simulation repetitions, η was picked from a gaussian distribution with a mean of 0.3 and a standard deviation of 0.15, where the values were bounded to the interval of $[0,1]$.

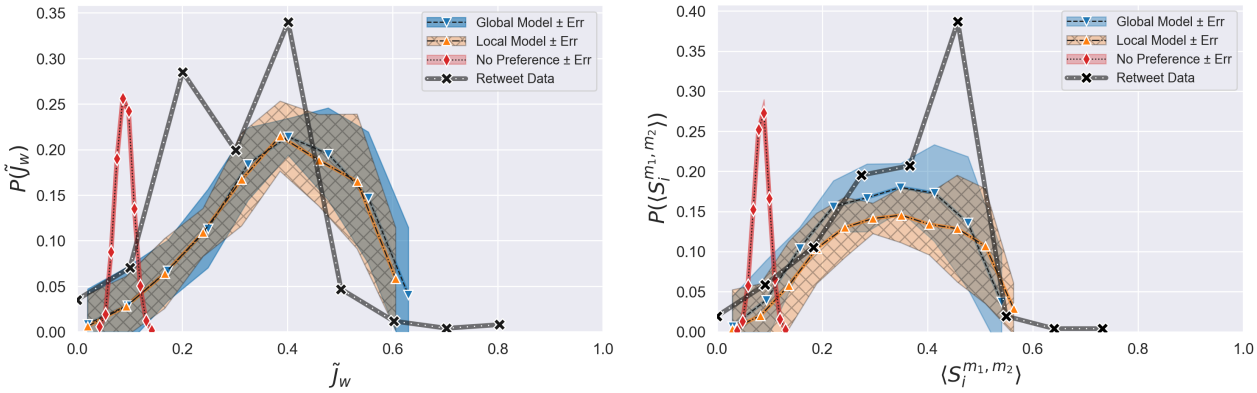
To better mimic the real world, we restricted the initiating nodes of the random walks, meaning that only a certain subset of nodes started the chain of retweets by tweeting an original tweet. A quarter of the nodes were chosen as the initiating subset of

the network, corresponding to the 25% of the users of X who publish 95% of the original tweets on this platform[43]. Finally, we used a different number of total steps for the random walk of each hashtag. The number of steps for each simulation ensemble was chosen so that the distribution of the ratio of total retweets to the number of participating users of the hashtags of the real world data was satisfied.

After performing the retweet simulation as such, it is time to apply the preference existence measures mentioned in sections 3.2.1 and 3.2.2 on each pair of the hashtags both for real data (which have been halved as explained above) and their simulated counterparts separately.

4 Results and Discussion

The results of the modified weighted Jaccard index and mean functional similarity of nodes are depicted in Fig. 6. This result shows the probability distribution of each measure when applied to all the available pairs of networks in the case of real data, along with three cases of retweet simulations conducted on the three underlying networks of section 3.3. We used Sturge's rule to determine the number of bins in the figures.



(a) Probability distribution of Jaccard index of pairs of hashtags **(b)** Probability distribution of mean functional similarities of nodes in pairs of hashtags

Figure 5. Result of measures of the existence of preference in the dissemination of tweets.

The sampling errors corresponding to the three simulation cases are shown with colored areas (where the error corresponding to the case with no preference is so low that it is hardly visible).

As is seen, both the model and the retweet data span a range of values that the model with no preference cannot acquire in both the Jaccard index and mean node similarity. This means that indeed there exist sharing preferences among users in this example social network case. Moreover, the preferential models can capture the overall trend of the preference measures in the retweet data. The deficiencies in the distributions of the models compared to the data can be attributed to the models' abstraction of the phenomena of information sharing, where it lacks to encompass the full complexity of this social-technological process, such as the presence of modular structures in the network of users[44], or the heterogeneity of activity rate for the sender nodes[45,46].

It is worth noting that between the two measures, the average functional similarities of the models' simulations is more comparable to the data than the results of the modified weighted Jaccard index. This is to be expected, since the latter relies on all the steps of the sharing of one hashtag, while the node similarity only studies how each individual node behaves regardless of what happens in the next steps (e.g. whether the next node will choose the more likely choice of sharing or not).

5 Conclusion

In this study, we found that the spreading of news on social networks is not random and that there are specific preferential paths for this news flow. Knowing these paths will play an important role in managing or limiting its dissemination. We introduced two mathematical models of local and global preferences to evolve networks with no preference into networks that exhibit preferential behaviors similar to those of a real-world case (e.g., retweets of hashtags on the X social platform). We measured preferential behavior using a modified weighted Jaccard index and mean node similarity. A sample of retweet data of trending hashtags confirmed the effectiveness of these two models and the preferential dissemination of news on social networks. The results of the simulated models, although differs in details from the real world data, captures the essence of the existence of

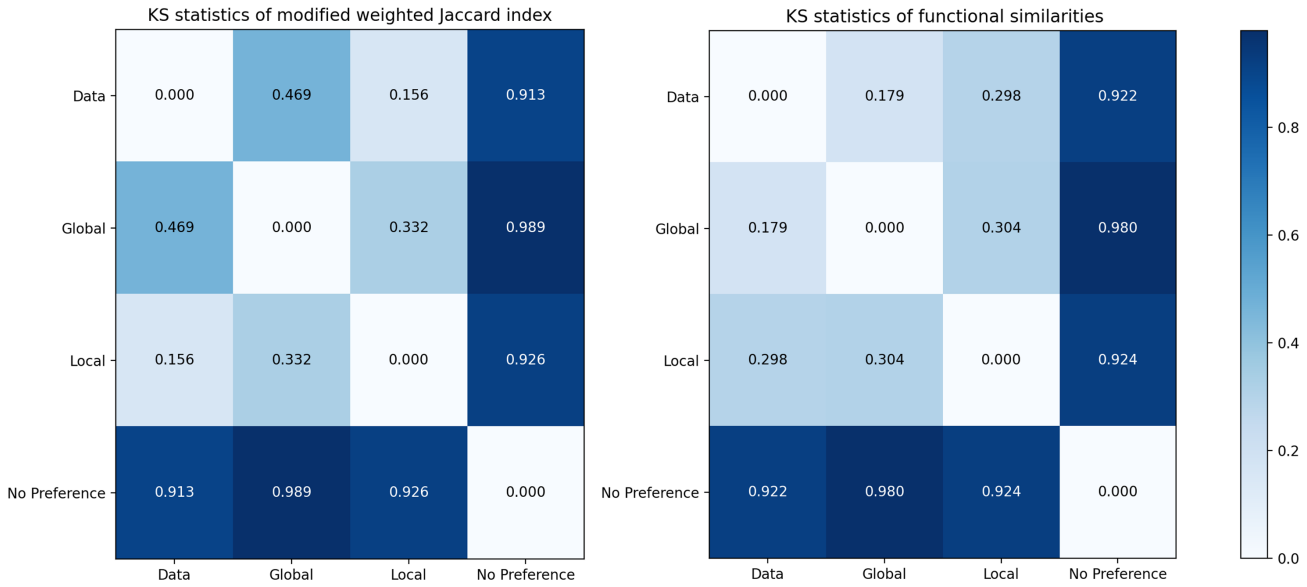


Figure 6. Statistics of the Kolmogorov-Smirnov test for the measures of the preference of existence. The less the value of the statistics, the more the two distributions are alike.

preference in news dissemination seen in real world data, which significantly differs from that of a no preference model. This suggest that the mechanism of the birth of preferences for social users can follow the dynamics of the introduced models. It is important to emphasize that these preferred transmission paths are dependent on the type of news. A user's choice of recipient may vary based on the content; for example, one contact may be preferred for political news, while another is favored for humorous content.

Future works may expand or combine global and local preferential models to better resemble a case of real world interest. The choice of the evolution time of the underlying network is another case that can be studied further and can potentially give insight into the age of the underlying network. It is also possible to study how the underlying network of two hashtags with considerably different topics differs from one another using the same measures of preference existence.

References

1. Yang, J., Yao, C., Ma, W. & Chen, G. A study of the spreading scheme for viral marketing based on a complex network model. *Phys. A: Stat. Mech. its Appl.* **389**, 859–870 (2010).
2. Gausen, A., Luk, W. & Guo, C. Can we stop fake news? using agent-based modelling to evaluate countermeasures for misinformation on social media. In *ICWSM Workshops* (2021).
3. Hu, H. Competing opinion diffusion on social networks. *Royal Soc. Open Sci.* **4**, 171160 (2017).
4. Lang, N., Wang, L. & Zha, Q. Opinion dynamics in social networks under competition: the role of influencing factors in consensus-reaching. *Royal Soc. Open Sci.* **9**, 211732 (2022).
5. Lewis-Beck, M. S. Election forecasting: Principles and practice. *The Br. J. Polit. Int. Relations* **7**, 145–164 (2005).
6. Kim, J. & Hastak, M. Social network analysis: Characteristics of online social networks after a disaster. *Int. journal information management* **38**, 86–96 (2018).
7. Zhang, N., Huang, H. & Su, B. Comprehensive analysis of information dissemination in disasters. *Phys. A: Stat. Mech. its Appl.* **462**, 846–857 (2016).
8. Grabowicz, P. A., Ramasco, J. J., Gonçalves, B. & Eguíluz, V. M. Entangling mobility and interactions in social media. *PloS one* **9**, e92196 (2014).
9. Crandall, D. J. *et al.* Inferring social ties from geographic coincidences. *Proc. Natl. Acad. Sci.* **107**, 22436–22441 (2010).
10. Poux-Médard, G., Velcin, J. & Loudcher, S. Dirichlet-survival process: Scalable inference of topic-dependent diffusion networks. In *European Conference on Information Retrieval*, 562–570 (Springer, 2023).

11. Molaei, S., Babaei, S., Salehi, M. & Jalili, M. Information spread and topic diffusion in heterogeneous information networks. *Sci. reports* **8**, 9549 (2018).

12. Kurka, D. B., Godoy, A. & Von Zuben, F. J. Online social network analysis: A survey of research applications in computer science. *arXiv preprint arXiv:1504.05655* (2015).

13. Li, M., Wang, X., Gao, K. & Zhang, S. A survey on information diffusion in online social networks: Models and methods. *Information* **8**, 118 (2017).

14. Gómez-Gardeñes, J. & Latora, V. Entropy rate of diffusion processes on complex networks. *Phys. Rev. E* **78**, 065102 (2008).

15. Firdaus, S. N., Ding, C. & Sadeghian, A. Retweet prediction based on topic, emotion and personality. *Online Soc. Networks Media* **25**, 100165 (2021).

16. Bunyamin, H. & Tunys, T. A comparison of retweet prediction approaches: the superiority of random forest learning method. *TELKOMNIKA (Telecommunication Comput. Electron. Control.* **14**, 1052–1058 (2016).

17. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *science* **286**, 509–512 (1999).

18. Petra4711. Swamp, florida, wetland image. https://pixabay.com/photos/swamp-florida-wetland-river-delta-2248571/, Accessed May 20, 2025 (2017).

19. Rubenstein, D. I. On predation, competition, and the advantages of group living. In *Social behavior*, 205–231 (Springer, 1978).

20. Wong, L. Y. & Burkell, J. Motivations for sharing news on social media. In *Proceedings of the 8th International conference on social media & society*, 1–5 (2017).

21. Berger, J. & Milkman, K. L. What makes online content viral? *J. marketing research* **49**, 192–205 (2012).

22. Picone, I., De Wolf, R. & Robijt, S. Who shares what with whom and why?: News sharing profiles amongst flemish news users. In *The Future of Journalism: Risks, Threats and Opportunities*, 118–129 (Routledge, 2020).

23. Hanteer, O. & Rossi, L. An innovative way to model twitter topic-driven interactions using multiplex networks. *Front. big Data* **2**, 9 (2019).

24. Wasserman, S. Social network analysis: Methods and applications. *The Press. Synd. Univ. Camb.* 173 (1994).

25. Rabbani, F., Khraisha, T., Abbasi, F. & Jafari, G. R. Memory effects on link formation in temporal networks: A fractional calculus approach. *Phys. A: Stat. Mech. its Appl.* **564**, 125502 (2021).

26. Price, D. J. D. S. Networks of scientific papers: The pattern of bibliographic references indicates the nature of the scientific research front. *Science* **149**, 510–515 (1965).

27. Kwak, H. What is twitter, a social network or a news media? *Dep. Comput. Sci. KAIST* (2010).

28. Jiménez-Zafra, S. M., Sáez-Castillo, A. J., Conde-Sánchez, A. & Martín-Valdivia, M. T. How do sentiments affect virality on twitter? *Royal Soc. Open Sci.* **8**, 201756 (2021).

29. Stieglitz, S. & Dang-Xuan, L. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *J. management information systems* **29**, 217–248 (2013).

30. Bastos, M., Travitzki, R. & Puschmann, C. What sticks with whom? twitter follower-followee networks and news classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 6, 6–13 (2012).

31. Mohammadikian, R. & Bigdeli, P. Retweet graphs of political hashtags in 2021 iranian presidential election. Dryad [Dataset], DOI: 10.5061/dryad.34tmpg4wq (2025). Forthcoming.

32. Mohammadi, S., Moradi, P., Firouzabadi, S. M. & Jafari, G. The footprint of campaign strategies in farsi twitter: A case for 2021 iranian presidential election. *Plos one* **17**, e0270822 (2022).

33. Bigdeli, P., Moradi, P. & Jafari, R. Bots election: Unveiling the complex network of social botnets. 10.21203/rs.3.rs-4773878/v2 (2024).

34. Nichols, L. Social desire paths: an applied sociology of interests. *Soc. Curr.* **1**, 166–172 (2014).

35. Evkoski, B., Mozetič, I., Ljubešić, N. & Kralj Novak, P. Community evolution in retweet networks. *Plos one* **16**, e0256175 (2021).

36. JafariAsbagh, M., Ferrara, E., Varol, O., Menczer, F. & Flammini, A. Clustering memes in social media streams. *Soc. Netw. Analysis Min.* **4**, 1–13 (2014).

37. Costa, L. d. F. Further generalizations of the jaccard index. *arXiv preprint arXiv:2110.09619* (2021).

38. Newman, M. E. Detecting community structure in networks. *The Eur. physical journal B* **38**, 321–330 (2004).

39. Newman, M. E. Communities, modules and large-scale structure in networks. *Nat. physics* **8**, 25–31 (2012).

40. Almaas, E., Kovács, B., Vicsek, T., Oltvai, Z. & Barabási, A.-L. Global organization of metabolic fluxes in the bacterium escherichia coli. *Nature* **427**, 839–843 (2004).

41. Glasgow, K. & Fink, C. Hashtag lifespan and social networks during the london riots. In *Social Computing, Behavioral-Cultural Modeling and Prediction: 6th International Conference, SBP 2013, Washington, DC, USA, April 2-5, 2013. Proceedings 6*, 311–320 (Springer, 2013).

42. Rosenstiel, T., Sonderman, J., Loker, K., Ivancin, M. & Kjarval, N. Twitter and life. https://americanpressinstitute.org/twitter-and-life/. Accessed: 2025-04-19.

43. McClain, C., Widjaya, R., Rivero, G. & Smith, A. The behaviors and attitudes of us adults on twitter. (2021).

44. Newman, M. E. Modularity and community structure in networks. *Proc. national academy sciences* **103**, 8577–8582 (2006).

45. Ying, Q. F., Chiu, D. M., Venkatramanan, S. & Zhang, X. User modeling and usage profiling based on temporal posting behavior in osns. *Online Soc. Networks Media* **8**, 32–41 (2018).

46. Vaca Ruiz, C., Aiello, L. M. & Jaimes, A. Modeling dynamics of attention in social media with user efficiency. *EPJ Data Sci.* **3**, 1–15 (2014).

Author contributions statement

R.M. contributed to the conception of the project, conducted data analysis and simulations, and wrote the manuscript. B.A. contributed to project conception and manuscript writing. P.B. contributed to project conception and data handling. R.J. contributed to project conception, manuscript writing, and supervised the overall project.

Additional information

Competing interests
The authors declare no competing interests.

Data availability
The simulation code used in this study is available at the project's GitHub repository. The datasets supporting this article have been uploaded on Dryad[31].