Modeling Semantic Influence for Biomedical Research Topics using MeSH Hierarchy

Dan He
IBM T.J. Watson Research
Yorktown Heights, NY
dhe@us.ibm.com

***Abstract*—In this work, we model how biomedical topics influence one another, given they are organized in a topic hierarchy, MeSH, in which the edges capture a parent-child/subsumption relationship among topics. This information enables studying influence of topics from a semantic perspective, which might be very important in analyzing topic evolution and is missing from the current literature. We first define a burst-based action for topics, which models upward momentum in popularity (or "elevated occurrences" of the topics), and use it to define two types of influence: accumulation influence and propagation influence. We then propose a model of influence between topics, and develop an efficient algorithm (*TIPS*) to identify influential topics. Experiments show that our model is successful at identifying influential topics and the algorithm is very efficient.**

Keywords: Bursts, Social Influence, Topic Hierarchies, MeSH

I. Introduction

In recent years, the problem of mining influence in networks, especially social networks, has attracted tremendous levels of interest. One of the key questions is how ideas, information, or influence spread (cascade) through the network. Lots of research [1], [6], [8], [9] has been aimed at problems related to this question — such as *viral marketing*, which targets the most influential users in the network, or *influence maximization*, which selects an initial set of users who ultimately influence the largest number of users in the network.

In this work, we want to model the evolution of research trends, focusing on biomedical research. This is a social process, and can be naturally abstracted into models of interactions between research topics. It is a very important problem to identify how attention spent on topics spreads from one area to others, which may shed light on how these topics will evolve in the future. Identification of influential research topics can be also used to guide investment of resources, such as in getting grants.

It is well known that biomedical topics are organized in a topic hierarchy, MeSH (*Medical Subject Headings*) hierarchy (http://www.nlm.nih.gov/mesh/). Therefore it is natural to model biomedical topic influence using MeSH hierarchy, which provides relationships among topics. The problem of mining influence in a topic hierarchy appears in some ways similar to the problem in social networks. Many models [1], [6], [8], [9] have been shown to be effective in modeling influence among people in social networks.

In traditional social networks, the nodes are usually users, and the users can influence each other through certain events or actions. For example, a user might undertake an action to buy a product, and this action then might affect their friends' decisions on whether or not to buy the product. The actions are temporal in that different actions may occur at different times and we can order the occurrence times of the actions. An action can only affect actions that occur later. A user is also associated with a binary status 'active' or 'inactive', corresponding to whether or not the user takes the action. This status changes at the moment the user takes the action. All users are initially inactive. Some users then take action themselves without being influenced by anyone else; they can be considered as initiators. Other users can be activated by their neighbors at a later time.

Unlike people in social networks, topics in a hierarchy cannot 'take action' to change their status. In our model, we define actions for a topic as 'rapid' changes in its popularity. Since the popularity of a research topic is usually measured by its frequency or the number of occurrences of the topic in the literature (of a related field), the actions can be modeled as 'elevated occurrences' of the topic. Therefore, we are interested in learning the process by which, when a topic has a rapid increase in popularity, this increase in popularity spreads to the topic's ancestors, descendants and siblings.

In order to handle the unique properties of this problem, we present a novel protocol. We first define a 'burst-based action'. Since the action is to model 'elevated occurrences' of topics, which are treated as 'bursts' in the frequency of the topic, we apply the topic-dynamic model of He and Parker [3] to identify such bursts. There are a few works on modeling bursts such as [3], [5], [7], [10]. We adopted the model in [3] because it is suited to burst detection in a hierarchy in which occurrences of the topics are accumulated, and this model of bursts permits extension to represent *burst-based action* of a topic as the complete process of birth, growth, and death of topic bursts. More details are given later.

Next we define two types of influence. When the influence is from child to parent in the hierarchy, we refer to it as *accumulation influence*. When the influence is from parent to child, we refer to it as *propagation influence*. More detailed definitions are given later. Therefore, a node can influence its neighbors through accumulation or propagation.

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

The influence will decrease as it spreads from a node. We can calculate the influence of a node to all other nodes in the hierarchy as its *influence power*, reflecting how influential this node is. To efficiently compute the influence power of all nodes in the hierarchy, we develop an algorithm — *TIPS* — which we show is much more efficient than a traditional breadth-first search algorithm. Our experimental results show that this method not only yields meaningful results, but also efficient and accurate.

II. Models

A. Burst-based Action

To model topic influence, we need to define actions for the topics, which can model elevated occurrences of the topics. The actions are based on the bursts defined in [3], based on technical analysis indicators such as EMA, MACD and MACD histogram. A *burst* is defined as a time interval in which the MACD Histogram value is greater than 0. The *strength* of the burst at a specific timestamp is the MACD Histogram value of that timestamp. We then define the *burst-based action* as the whole process of the occurrence, growth, and death of the burst. And from now on, 'year' will be the only time unit we use. We also say a topic is *activated* when a burst of the topic occurs. The readers can refer to [3] for the details of the topic dynamic model.

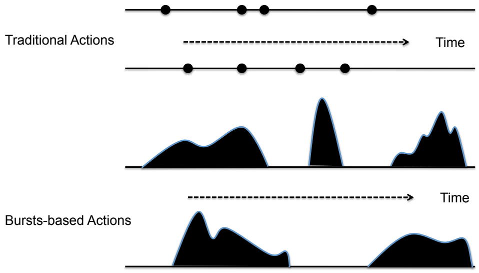

Figure 1. Visualized comparison of the traditional action and our burst-based action.

B. Accumulation Influence

Accumulation influence is defined as frequency accumulation from child nodes to their parent. The reason we define this influence is because in the hierarchy, any occurrence of a descendant is also considered as an occurrence of its ancestors. Each child node contributes to bursts of parent nodes to a certain extent, which is used to quantify the accumulation influence. For each topic c, there may be multiple bursts occurring at different times. We create a node c_i for the i-th burst of topic c. If c's parent is topic p in the hierarchy, all c_i's are child nodes of all p_j's where p_j is the node for the j-th burst of topic p. Then the *accumulation influence* of a child node c_i to a parent node p_j is defined as:

$$AI_{c_i,p_j} = \frac{\sum_{k \in B_{c_i} \cap B_{p_j}} F^k_{c_i}}{\sum_{k \in B_{c_i} \cap B_{p_j}} F^k_{p_j}}$$

where AI_{c_i,p_j} denotes the strength of the accumulation influence from c_i to p_j, B_{c_i} and B_{p_j} denote the set of consecutive burst years for c_i and p_j, respectively, $F^k_{c_i}$ and $F^k_{p_j}$ denote topic occurrence frequency at year k for c_i and p_j, respectively. AI_{c_i,p_j} is always less than 1 since $F^k_{c_i} \leq F^k_{p_j}$ due to the accumulation of frequency. Notice we cannot accumulate bursts instead because the sum of the burst strength at a specific time from all children nodes is usually not equal to the burst strength of the parent node. Therefore, we calculate the accumulation influence probability AI_{c_i,p_j} by accumulating the occurrence frequency during the overlapping years between the burst of p_j and the burst of c_i. Since there is no delay for accumulation, we require the start year of the burst for c_i be before the start year of the burst for p_j. Otherwise the accumulation influence from c_i to p_j is 0.

To show the accumulation is indeed a reasonable operation, we pair the start time of every burst of a node with the start time of every burst of the node's direct parent in the hierarchy, which is the closest burst right after the node's burst. We do this for every node such that we obtain two vectors, one for the burst-based action start time of children nodes, one for the action start time of parent nodes. The actions of parent nodes are the closest actions right after their children's actions. We then compute the correlation of the two vectors. We obtain a strong correlation of 0.87 with p-value 2.2e-16, indicating the occurrences of the actions of children nodes have a strong correlation to the occurrences of the actions of parent nodes. Although the correlation doesn't necessary mean causality, it suggests accumulation is a reasonable operation in the hierarchy which naturally captures the possible causes for the action occurrences of parent nodes. Again, in this work, we assume causalities do exist and we do not discuss how to validate them.

C. Propagation Influence

The accumulation of influence models influence spread from child to parent. Conversely, however, the parent can influence the child — by propagation influence. Propagation influence is similar to the traditional social influence in that the actions are temporal. In our problem, the burst of the parent can propagate to its children. The occurrence time of the bursts of parent and child can be used to compute the probability of influence from the parent to the child. Goyal et al. [2] proposed several probability models for the influence of one node to another node based on temporal data. However, in our problem setting, these models cannot be applied directly. This is because in traditional social influence problem, each action is assumed to occur at a specific timestamp and each node has only binary status: 'active' and 'inactive'. Therefore these actions can be considered as single timestamp actions. In our problem setting, the burst-based action occurs at a specific timestamp and lasts for a certain amount of time (and then vanishes). During the time period the bursts occur, they have continuous values, or strengths. The actions in our problem are thus two-dimensional regions defined by the burst periods. We show a visual comparison of the traditional social network action and our burst-based action in Figure 1. As one can see, the traditional actions are all discrete points along the time axis. Our burst-based actions correspond to two-dimensional (strength and time) burst regions.

To handle the burst-based actions in our problem, we need to design a new probability model. We follow the Independent Cascade model to define the propagation influence probability between any pair of nodes. The propagation influence probability depends on the occurrence time of the action, the strength of the action and also the length of the period of the action. Similar to the traditional social influence, the more distant the occurrence times of the two actions, the smaller the influence between them. On the other hand, we assume that the stronger the actions, the stronger the influence, since a small burst is likely to be caused by the random fluctuation of the frequency. Also because the actions span a time period, influence needs to be accumulated over the whole period. Based on this intuition, we define *propagation influence* from a parent node p_j to a child node c_i as:

$$\begin{aligned} PI_{p_j,c_i} &= \frac{\sum_{m\in B_{c_i}} \sum_{n\in B_{p_j}, n<m} I_{n,m} H(\frac{S^n_{p_j}+S^m_{c_i}}{2})}{|B_{c_i}| \times |B_{p_j}|} \\ H(s) &= \begin{cases} 1-\frac{1}{s^\alpha} & 1-\frac{1}{s^\alpha}>0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

where PI_{p_j,c_i} denotes the propagation influence from p_j to c_i, B_{c_i} and B_{p_j} denote the set of consecutive burst years for c_i and p_j, respectively, $S^m_{c_i}$ and $S^n_{p_j}$ denote the strength of the burst at year m for c_i and at year n for p_j, and $H(s)$ is a monotone function with respect to s within the range of [0,1], where s is the average burst strength of the two actions. We set $H(s) = 1 - \frac{1}{s^\alpha}$, which is a power-law model that is often used in social network mining problems and α is a decay parameter. We simply set α to be 1. If $1 - \frac{1}{s^\alpha}$ is negative, indicating the averaged burst strength is less than 1, which is very small, we simply set it to be 0 such that $H(s)$ is guaranteed to be within the range of [0,1]. $I_{n,m}$ denotes the traditional social influence between two actions occurring at timestamp n and m, where n needs to be less than m.

Popular models for traditional social influence between two actions include an exponential model and a power-law model:

$$\begin{aligned} exponential\ \ model: &\quad P(\delta) \propto e^{-\frac{\delta}{\alpha}} \\ power-law\ \ model: &\quad P(\delta) \propto \frac{1}{\delta^\alpha} \end{aligned}$$

Here $P(\delta)$ is the traditional influence probability between the two actions, δ is the difference between the occurrence times of the two actions, α is a decay parameter. In our experiments, we tested different values for α. The results do not vary much. Therefore, we simply set α to be 1.

Therefore, propagation influence is accumulated for each year m in the burst period of c_i, and this can be influenced by all years $n < m$ in the burst period of p_j, weighted by the average strength of the two bursts. Just as in models for traditional social influence, the start year of p_j must precede the start year of c_i for the propagation to be meaningful.

D. *Influence Graph*

Once we define accumulation influence and propagation influence, we can build an influence graph $G = (V, E)$, where V is the set of nodes for all bursts of the set of topics in the hierarchy, and E is the set of edges between pairs of bursts whose influence probability is not 0. These edges need to be consistent with the edges in the hierarchy, namely the two corresponding topics c, p of an edge $e = (c_i, p_j) \in E$ must be connected by an edge in the hierarchy. The edges in E have directions, indicating who influences whom, which may not be consistent with the direction of the 'is-a' relationship in the hierarchy. Therefore children and parents in the hierarchy can affect each other through different types of influence, and the edges in E are also labeled accordingly with either accumulation influence or propagation influence, as well as the probability of influence.

We define the following *burst occurrence constraints* for the two types of influence:

1) *Accumulation Influence Constraint* from a_i to b_j: $ts(a_i) < ts(b_j)$ and $te(a_i) > ts(b_j)$.
2) *Propagation Influence Constraint* from a_i to b_j: $ts(a_i) < ts(b_j)$.

Here $ts(a_i)$ and $te(a_i)$ are the start year and end year of the burst for node a_i, respectively. The *Accumulation Influence Constraint* thus requires the burst from a child node a_i to overlap with the burst of the parent node b_j. The *Propagation Influence Constraint* requires the start year of the burst from a parent node a_i to be no later than the start year of the burst of the child node b_j. Therefore between a child node and a parent node, due to these constraints, we can only have one type of influence, not both.

Since the edges in the graph are directed, the influence graph tells us how the bursts propagate through the hierarchy, such as how likely the burst of a child activates the burst of its parent and how a child influences its siblings through the parent. Also given the directed edges, it is possible that there is no path between a pair of nodes. We say node a *has a path to* node b when a and b are connected by at least one path and the direction of the path is from a to b. The burst occurrence constraints guarantee that there is no loop in the graph.

Given the graph, we can compute the influence probability of a node a to a node b. Consider a simple case where a and b are siblings and their parent is c. a can influence b through their co-parent c. Assume we have $AI_{a,c}$ and $PI_{c,b}$, then the influence probability for the burst of a to the burst of node b is $AI_{a,c} \times PI_{c,b}$. Let $Ancestor(a_i)$ and $Ancestor(b_j)$ be the set of ancestors of a_i and b_j, respectively, $Path_k(a_i, b_j)$ is the set of nodes on the path from a_i to b_j. If there is no path from a to b, the influence probability between a and b is 0. For a more general case, we define the influence probability of the i-th burst of a to the j-th burst of b as:

$$\begin{aligned} IP_k(a_i, b_j) &= AI_k(a_i) \times PI_k(b_j) \\ AI_k(a_i) &= \prod_{c \in \{Ancestor(a_i) \cap Path_k(a_i, b_j)\}} AI_{a_i, c} \\ PI_k(b_j) &= \prod_{d \in \{Ancestor(b_j) \cap Path_k(a_i, b_j)\}} PI_{d, b_j} \\ IP(a_i, b_j) &= 1 - \prod_k (1 - IP_k(a_i, b_j)) \end{aligned}$$

where $IP(a_i, b_j)$ is the influence probability between a_i and b_j, $IP_k(a_i, b_j)$ is the influence probability between a_i and b_j following the k-th path between a_i and b_j. All bursts of the nodes involved must follow the burst occurrence constraints. We define $IP(a_i, b_j)$ as $1 - \prod_k (1 - IP_k(a_i, b_j))$ since there can be multiple paths between a_i and b_j, and b_j can be activated multiple times. Therefore the probability is 1 minus the probability that none of these attempts activate b_j.

E. Influence Power

We define the *influence power* of a node a_i, $IP(a_i)$ as the sum of the influence probability from a_i to all other nodes b_j's in the influence graph $G = (V, E)$. Therefore, influence power quantifies how important a topic is in the biomedical research with respect to the semantic influence.

$$IP(a_i) = \sum_{b_j \in V, b_j \neq a_i} IP(a_i, b_j)$$

Our goal is to compute the influence power for all topics. The two intuitive key factors that affect the influence power of a topic are the *frequency* of the topic and the *connectivity* of the topic. Since the influence is burst-based, the strength of the bursts depends on the frequency of the topics. The higher the frequency, the more likely the burst is strong, and therefore the higher the influence probability from the topic to other topics. The connectivity of a topic indicates how many other topics this topic can influence. The more topics the current topic can influence, the higher its influence power.

However, it is not necessarily the case that a higher frequency or connectivity lead to a higher influence power. As our model indicates, propagation influence depends not only on the strength of the burst, but also on the difference of the times of the bursts. Even though the bursts may be strong, a significant time difference in the exponential model or power-law model may decrease the influence probability dramatically. On the other hand, even though a topic may influence many other topics, if the influence probability of the edges in the influence path is low, the influence power of the current topic would not be high.

With the definition of influence power for a node in the influence graph, we can compute the influence power for each node and rank nodes accordingly. A naive algorithm applies breadth-first search from each node to compute its influence power. The complexity of the naive algorithm is $O(|V|^3)$, where $|V|$ is total number of bursts for all nodes in the hierarchy, or equivalently the number of nodes in the influence graph. In our problem setting, $|V|$ is around 160,000. The complexity can be then very high. The naive algorithm is not efficient in the following two respects: (1) If there are multiple paths from one node to another, and the paths share some edges, the naive algorithm may explore nodes on these edges multiple times. (2) If paths from two different nodes share edges, the naive algorithm explores nodes on these edges multiple times.

Next we develop a more efficient algorithm **TIPS**(**t**wo-l**i**sts de**p**th-fir**s**t) search algorithm. In our algorithm, we start from all the source nodes in the influence graph, which only have out-going edges but no incoming edges. We do depth-first search from each source node and we compute the influence probability along the way. However, to avoid exploring the same node multiple times, we maintain two lists for each node a searched, a *To* list and a *From* list. The *To* list stores all the nodes which have paths to the node a. The *From* list stores all the nodes to which a has paths. During the depth-first search, if a node a is reached and a has not been explored before, we update the *To* list for a and the *From* lists for all nodes that have paths to a. However, when a has been already explored (meaning the depth-first search finds a new path to a), we don't need to explore a any further. Since we have maintained two lists for a, for all the nodes on the new path to a, we update their *From* list as well as the influence probability from them to a. Meanwhile we update the *To* list of a by adding all nodes on the new path to a. Since the new path to a will affect a's influence probability as well as the influence probability of all the nodes in a's *From* list, we update the influence probability from a to all the nodes in a's *From* list. Once we have finished searching the whole graph in this way, the influence power of a node a will be the sum of the influence probabilities from a to each of the nodes in its *From* list.

Although theoretically the complexity of the algorithm *TIPS* is still $O(|V|^3)$, in reality *TIPS* is much more efficient: when we are trying to explore a node that has been explored before, *TIPS* only needs to update the *To* and *From* lists, and the nodes that need to be updated can be obtained directly from the two lists. By contrast, breadth-first search needs to explore nodes on common edges repeatedly, which requires searching for neighbors of the nodes, and this search turns out to be time consuming. However, *TIPS* consumes more memory to save the two lists. Therefore, *TIPS* trades time efficiency for memory efficiency.

F. *Baseline Models*

According to our model, two criteria affect the influence probability between a pair of topics: (1) distance of the two topics in the hierarchy (the greater their distance, the lower the influence probability); (2) the influence probability of the edges in the path (the higher the probability of the edges, the higher the overall influence probability). Distance in the hierarchy measures the semantic relatedness of the two topics; more related topics are more likely to influence each other. On the other hand, influence probability of the edges measures the likelihood of influence between the two topics. The occurrences of topics not only have a strength, but also generally span certain time period. Therefore, in our influence probability definition, we accumulates the influence over all time units of the period and also weight the influence with the strength of the bursts. The two criteria are both important for the model.

To evaluate the importance of the above two criteria in defining a model for our problem setting, we further develop two baseline models to compute the influence power of the topics:

1) Probability-only Model: This model ignores the distance information as well as the hierarchy and compute the influence of the two topics directly based on their burst-based actions, namely $IP(a_i, b_j) = PI_{a_i,b_j}$, where PI_{a_i,b_j} is the same formula as the one for propagation influence. We use the same formula for propagation influence in this model such that both the strength and the time period of the bursts are considered. This model basically assumes "No hierarchy".
2) Distance-only Model: This model assumes that the influence probability of the edges in the influence path depends only on the start time of the bursts and ignores the strength of the bursts as well as the time period the bursts span. Thus the influence probability between two adjacent topics a_i, b_j is I_{a_i,b_j}, where I_{a_i,b_j} is the traditional social influence between the time-stamps for the burst-based actions of topics a_i and b_j. This model assumes "one-dimension definition", where we do not accumulate the influence for all the time units in the burst period.

Since the frequency and the connectivity of the topics have significant impact on their influence power, we can use the rank of the topics with respect to their frequency and connectivity to evaluate the performance of a model. The most influential topics by a good model should thus reward topic frequency and connectivity with relatively high rank. Our experiments later show that our model performs much better than the two alternative models.

III. Experimental Results

We applied *TIPS* on the MeSH hierarchy. MeSH terms (*Medical Subject Headings*) are a very widely-used hierarchy of topics in biomedical research. Each article in PubMed is annotated with descriptive MeSH terms. (Here we use 'topic' and 'term' interchangeably.) We crawled these articles for the years 1950 through 2008. For each term in the MeSH hierarchy, we counted its frequency of occurrence in each year, and accumulated these frequencies through the hierarchy (so that the frequency of a node is the sum of the frequencies of all descents of the node as well as of itself).

Depth	Number	Total MeSH term Number
1	109	109
2	1087	1495
3	2013	6527
4	1836	12446
5	1011	12702
6	525	7960
7	179	4304
8	83	1820
9	31	800
10	6	242

Table I

The number of topics with different depth in the top-10% most influential MeSH terms, where small depths mean higher generality.

To find the burst-based actions, we used the parameters (4,8,5) for the MACD histogram, as suggested by [3]. We first constructed the influence graph for the hierarchy according to our definition of accumulation influence and propagation influence. The total number of nodes in the graph was 160,075 and the total number of edges was 163,603. As the ratio of edge number to vertex number was almost 1, indicating that the chance of seeing multiple paths between two vertices was relatively low, *TIPS* finished in just 12 seconds.

We wanted to validate the ranks of these terms, ensuring that our method did indeed quantify the influence power of the terms in the hierarchy meaningfully. When a term has high frequency, it is more likely to have high influence power. Meanwhile, when the number of vertices reachable from a node is high (we can call this number the graph *connectivity*), or alternatively speaking, a term is able to influence many other terms, the term is more likely to have high influence power.

However, as we discussed before, influence power does not depend entirely on frequency and connectivity, as shown in Table I. For example, not all topics with depth 2 and 3 are of high influence power. On the contrary, there are many highly influential terms with higher depth, or lower generality, especially for terms of depth 4 and 5. This is consistent with the observation from the work of He and Wu [4] that researchers tend to use neither too general terms nor too specific terms. For illustration purpose, we also show the top-3 most influential terms of depth 4 and 5, respectively, in Table II.

In Figure 2, we show the frequency and the connectivity ranks of a term, normalizing all rank values to lie in [0,1] so that the higher the value, the higher the rank. In this way we can visualize the relationship between frequency and connectivity rank for the top-100 most influential topics.

Model	*top-1*	*top-2*	*top-3*	*top-4*	*top-5*	*top-6*	*top-7*	*top-8*	*top-9*	*top-10*
Dis-only model	7	2	2	2	2	2	1	1	1	1
Prob-only model	77	21	19	19	18	16	10	7	4	4
Our model	77	63	45	39	37	36	32	31	30	29

Table III
THE NUMBER OF NODES ACTIVATED BY THE TOP-10 MOST INFLUENTIAL NODES FOR THREE DIFFERENT MODELS.

MeSH Term	Depth	Rank
Antigens, CD	4	23
Gram-Negative Aerobic Rods and Cocci	4	25
Phosphotransferases	4	36
Adenocarcinoma	5	108
Insurance, Health	5	146
Proto-Oncogene Proteins	5	149

Table II
THE TOP-3 MOST INFLUENTIAL TERMS OF DEPTH 4 AND DEPTH 5, RESPECTIVELY.

As a comparison, we also plot the top-100 most influential topics by the distance-only model and the probability-only model. The terms ranked high by our method also have high rank in both frequency and connectivity, indicating that they indeed tend to have high influence power. On the contrary, the most influential topics by the other two alternative models quite often do not have high rank of connectivity and frequency; their ranks appear relatively random. This indicates our model can better capture topics with high influence power.

To further validate the effectiveness of our model in finding the most influential nodes in the hierarchy, we do propagation in the hierarchy from the top-10 most influential nodes identified by our model, the distant-only model and the probability-only model. The probability of one node activating another node is defined based on the accumulation and the propagation probabilities, according to the structure of the hierarchy. We show the number of nodes activated by each set of most influential nodes. For illustration purpose, we sort the number of activated nodes in decreasing order. As we can see in Table III, the most influential nodes from our model activates the most number of nodes in the hierarchy, suggesting our model truly captures the structure of the hierarchy and learns the influence based on the structure.

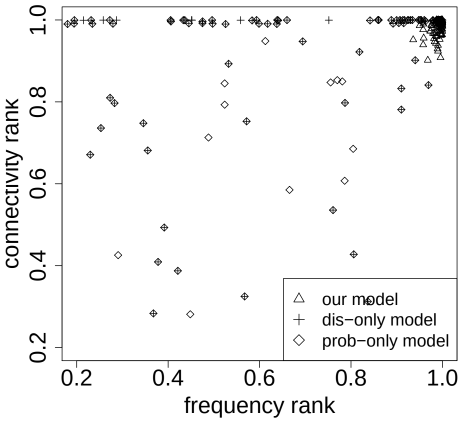

Figure 2. The frequency rank versus the connectivity rank for the top-100 most influential topics by our model, the distance-only model and the probability-only model.

IV. Conclusions

In this work, we have modeled semantic influence probabilities among topics in a hierarchy with burst-based accumulation influence and propagation influence. We have shown that similar models in social network influence mining can be naturally adapted here. We also developed an efficient algorithm to rank topics according to their influence power, defined as the sum of the influence probability of a topic to all other topics. Our experimental analysis shows that our burst-based topic influence model is meaningful and our algorithm to compute the influence of topics is efficient.

References

[1] P. Domingos and M. Richardson. Mining the network value of customers. In *Proceedings of SIGKDD*, pages 57–66. ACM, 2001.

[2] A. Goyal, F. Bonchi, and L.V.S. Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 241–250. ACM, 2010.

[3] D. He and Douglas S. Parker. Topic Dynamics: an alternative model of 'Bursts' in Streams of Topics. In *Proceedings of SIGKDD*, July 25-28, 2010.

[4] Dan He and Xindong Wu. Ontology-based feature weighting for biomedical literature classification. In *IRI*, pages 280–285, 2006.

[5] J. Kleinberg J. Leskovec, L. Backstrom. Meme-tracking and the dynamics of the news cycle. In *Proceedings of SIGKDD*, 2009.

[6] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of SIGKDD*, pages 137–146. ACM, 2003.

[7] Jon M. Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of SIGKDD*, pages 91–101, 2002.

[8] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *Proceedings of SIGKDD*, page 429. ACM, 2007.

[9] M.G. Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. *KDD10*, 2010.

[10] Yunyue Zhu and Dennis Shasha. Efficient elastic burst detection in data streams. In *Proceedings of SIGKDD*, pages 336–345, 2003.