# A Formal Analysis of Human Information Networks

Lakshminarasimhan Govindrajan, Vaarnan Drolia and Shubham Goyal
*Department of Computer Science,*
*School of Computing,*
*National University of Singapore*

## ABSTRACT

In this paper, we have analysed the many facets and properties of human information networks and explored many of the common paradoxes which relate to the knowledge distribution in these networks. To be very precise, we basically pondered over how the knowledge content and communication skills of the people in these networks affect the relative knowledge of the people within the network as well as the overall knowledge distribution of the network.

## 1. INTRODUCTION

The World Wide Web has become an integral part of our lives and it has even begun to influence our opinions and decisions in a large way. Our society has reached the state where most of the information in shared freely online through an increasing number of mediums of information exchange such as blogs and social networks in addition to the more traditional personal websites.

In this light, we have modelled these phenomena under the aegis of information exchange networks where we take nodes with different properties in terms of knowledge content, probability of information transmission and distance from each other. We generate graphs of varying knowledge distributions and densities which help us to make some very significaant claims.

### 1.1. Claims

Through this study, we seek to validate the non-intuitive paradox which states that the more the information there is, the lesser we tend to actually know. Nodes that are not directly linked can still have strong interactions between them through pathways. Such interactions are termed long-range. We seek to introduce and establish the 'Network Effect', which downplays the significance of the initial configuration of the system and tends to lay more importance on the various long range interactions involving the different components of the system. Moreover, this study introduces a novel method to identify "important" vertices in a graph via the calculation of their "networks" as described in greater detail in the later sections of this report.

### 1.2. Paper Organization

The rest of this paper is organized as follows. In Section 2 we discuss the problem and how we chose to model it in our experiments. In the following section 3, we describe the algorithms for generation of the model and obtaining the results.

Then in section 4, we share the results and their analysis which is finally followed by the Conclusion.

## 2. PROBLEM MODELLING

We propose a mathematical model for a blogging network where we assume each person to be a node. We then consider two domains of knowledge say, A and B and assign each node two weights representing their absolute knowledge in the aforementioned domains. Information flow can take place from node i to node j in both the domains A and B with probabilities of $P_A$ and $P_B$ respectively. We represent this in the form of directed edges. An edge from node i to node j represents information flow from j to i with probability equal to its corresponding edge weights ($P_A$ and $P_B$). The network of a node x is defined as a set of all paths terminating at x with the product of edge weights of each such path greater than a particular threshold value.

The weight of a path is defined as $\sum \frac{w_i p_{i-x}}{2^d}$ where for all nodes i in the path, $w_i$ is the weight of node i, $p_{i-x}$ is the product of probabilities from node i to the source node x and d refers to the number of edges in the path between i and x.

The knowledge content of x is defined as the sum of weights of all paths in x's network together with its own weight. This value is then normalized with the sum of the knowledge content of all nodes of the network to obtain the relative information content.

## 3. ALGORITHM

Our first task was to generate a graph which could represent a human information network. We decided to divide peoples knowledge into two topics, say A and B. In our implementation, we could create graphs having different values for the overall knowledge in subjects A and B. We can also decide how good people are in communicating information on subject A and subject B and generate a graph based on this.

After generating the graph, our next task was to find the influence networks of different nodes for both the topics A and B and then to calculate the relative knowledge of a person which not only includes his inherent knowledge but also the information which was transmitted from the other nodes in its network to it. Finding which nodes are in the network of node X involves doing a modified breadth first search starting from node X. Let us suppose there is a node Y which transmits information to node X via the path Y→C→A→G→X. Now, we know from our model that the edge weight of an edge connecting any two nodes is the probability of information transfer between the two nodes. Therefore, in going from node Y to node C, the amount of information transferred from node Y to node C is actually weight(Y)*P(Y, C). Extending this to information transfer from node Y to node X, we can easily agree that the amount of information transferred from node Y to node X is actually the product of weight of node Y and the edge weights corresponding to all edges which lie in the path.

However, the above result though seemingly logical, has a slight flaw. We have not accounted for the number of intermediaries through which the information from node Y to node X passed. This is because we know that information gets attenuated as it passes around as every intermediary has its own version to add. Thus, we assume that whenever an information flow between two adjacent nodes happens, the information that passes though is exponentially decreases. It is obvious that if there are two paths by which information can get transferred from a node to another node, we choose the path through which more

information gets transferred or probability of correct information transfer is more.

We then proceed to compute the network of a node. The network of a node basically includes all the nodes from which information is transmitted to the node. In our experiments, we also constrained the network of a node by means of a threshold value which necessitates that for a group of vertices that are included in the network of a particular node, the probability of information transfer from each of them to the node is greater than the threshold value.

Finally, after calculating the network of a node, we calculate the total knowledge of the node in the particular subject (either A or B) by adding the knowledge transmitted by other nodes in the network to the original weight of the node. Then we sum the total knowledge of the networks of all the nodes in the graph and normalize the knowledge of this nodes network by dividing it with that value. This normalized value actually represents the knowledge of that node in a particular subject relative to the total knowledge of the entire information network on that subject. Then we randomly add different nodes to the graph and connect them with the existing nodes. These new nodes also have some inherent knowledge and thus this phenomenon actually represents the growing information on the web or on human information networks in general. However, when we run the algorithm repeatedly on these new graphs obtained by the addition of new nodes, we try to see whether the total cumulative knowledge of all the people increases or decreases. We also try to see how some people actually increase their knowledge by getting connected to these new nodes or vice versa and how a lot of people actually dont benefit from the increasing total knowledge of the network because their relative knowledge decreases. Not only this, we also model how new nodes form connections with other nodes step by step depending upon their communication skills and analyse the progression of their knowledge content.

## 4. RESULTS AND DISCUSSION

We analyzed two different networks. During the initial analysis (Run 1), both the networks consisted of 100 nodes. 25 fresh nodes were then added to the networks and connections established while moderately improving connections amongst the already existing nodes. The networks were then subjected to another analysis (Run 2). This whole process was repeated to do the final analysis (Run 3). The threshold and density for both the networks were set to 0.3 and 0.6 respectively. The information content of each node on topics A and B were set so that the knowledge about topic A was greater than B in most cases (on an average ratio of 2:1 in favor of A). Nodes in Network 1 were effective communicators in topic A and relatively poorer communicators in topic B. The opposite was true of Network 2 where the nodes were better communicators in topic B.

When the relative knowledge of each node was plotted as a scatter plot after each run, it was generally observed that for a majority of nodes, their relative information content decreased in both domains A and B with increase in the number of nodes i.e. increase in information (Figure 1).

Further, the plots of similar stages for both domains were analyzed on the same scale to observe trends (Figure 2). For Network 1, it was noticed that relative knowledge of nodes was generally greater for domain B than for A for a majority of the nodes in Run 1. This is contrary to intuition as initially, the knowledge values of nodes for domain A were known to be higher and communication between nodes on topic A was better than that of B due to biasing of $P_A$ and $P_B$ values in favor of A. We call this the network effect. As nodes communicate more on topic A, their networks with respect to A are larger as compared to their networks with respect to topic B. Thus, information acquired by the nodes in the various networks for topic A is greater and a lot of nodes gain a lot of new knowledge related to topic A from these networks. Therefore, the relative knowledge of nodes for topic A becomes lesser than that for topic B after Run 1 because the information becomes more evenly distributed for topic A (Figure 2). This scenario starts to change when new nodes are added. This is observed in Run 2 and Run 3. When the new nodes attach to the graph, their relative knowledge in topic A becomes greater than their relative knowledge in topic B as they tend to gain more on topic A from their networks than B (since the nodes of the graph are better communicators for topic A). The dominance of relative knowledge of topic A exhibited by the new nodes is further consolidated when more nodes are added to the network.

The other situation (Network 2, Figure 3) is the case when knowledge on topic A is greater than the knowledge on topic B for a majority of the nodes (as was the case for Network 1) but communication is biased in favor of domain B. Run 1 for Network 2 was quite similar to that of Network 1 in the sense that the network effect came into play. We observed that relative knowledge of most of the nodes was greater for domain A than for domain B since the nodes were poorer communicators in topic A. This happened regardless of the initial knowledge content of these nodes for the two domains. However, when newer nodes were added to the network, the difference in their relative knowledge for the two domains was not as evident as it was for Network 1. This is because the advantage that the nodes had of having greater initial knowledge in domain A was cancelled out by the fact that they were poor communicators in topic A.

Thus, we see that this model helps us to observe and explains some non-intuitive trends. This can have many interesting implementations. These findings can prove to be crucial in explaining some important phenomena in social networks such as opinion leadership. Identifying maximum weight paths can also help in easy and efficient propagation of information through the network.

FIGURE LEGENDS

**Figure 1:** Variation in the value of relative knowledge of each node at each stage of the three step simulation for both Networks 1 and 2 as specified.

**Figure 2:** Comparing relative knowledge contents of each node in topics A and B after each step of the simulation for Network 1.

**Figure 3:** Comparing relative knowledge contents of each node in topics A and B after each step of the simulation for Network 2.

Figure 1:

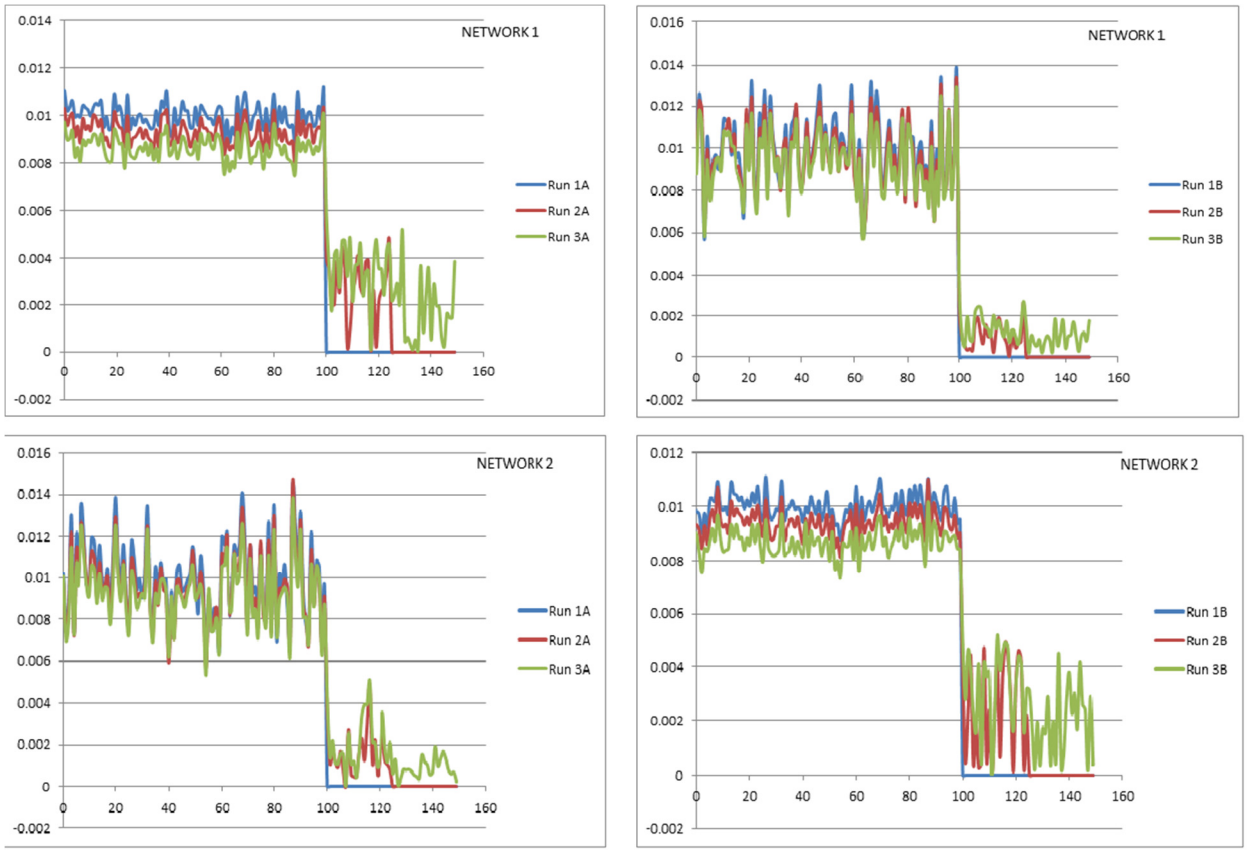
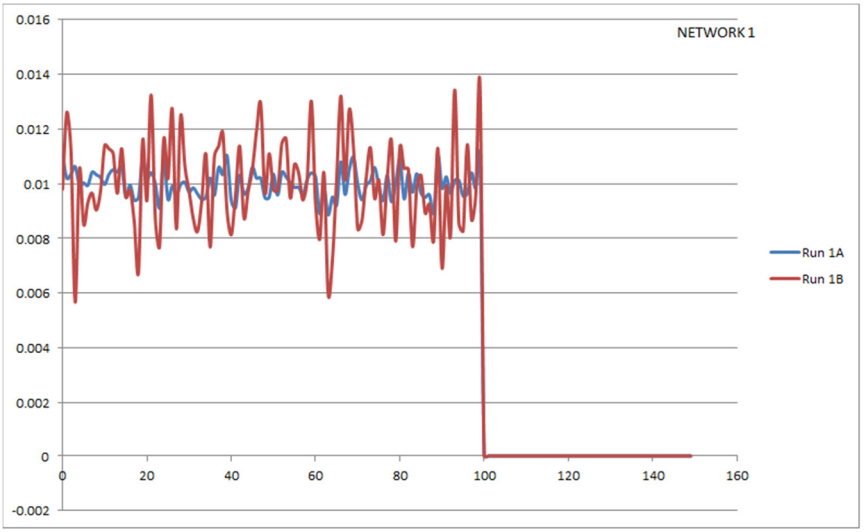
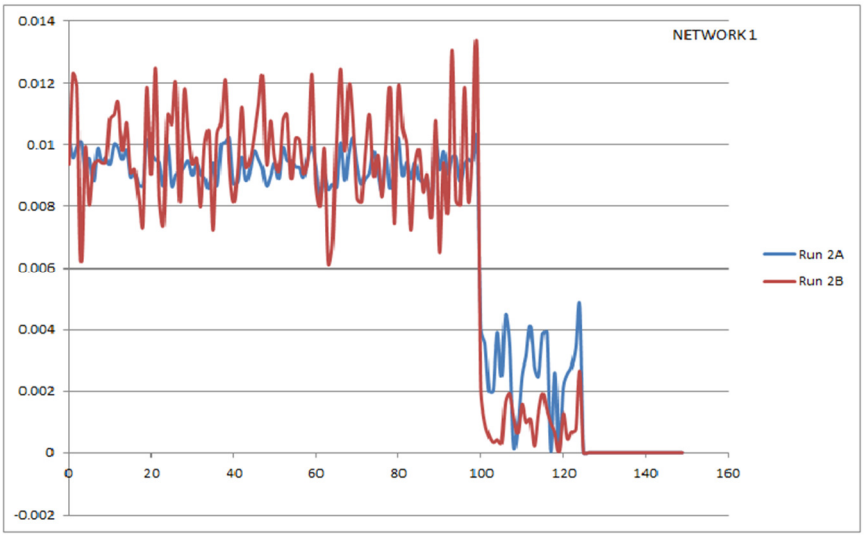


Figure 2:


NETWORK 1
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
-0.002
0
20
40
60
80
100
120
140
160
Run 1A
Run 1B


NETWORK 1
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
-0.002
0
20
40
60
80
100
120
140
160
Run 2A
Run 2B

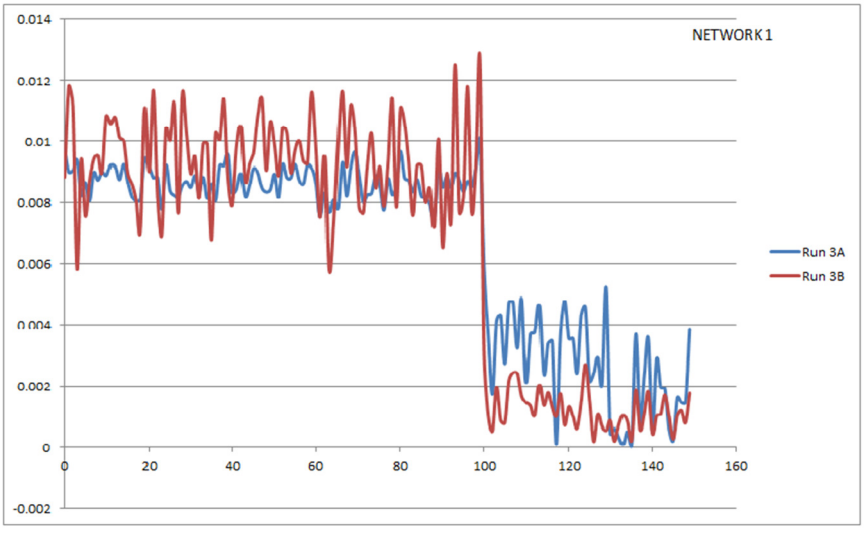
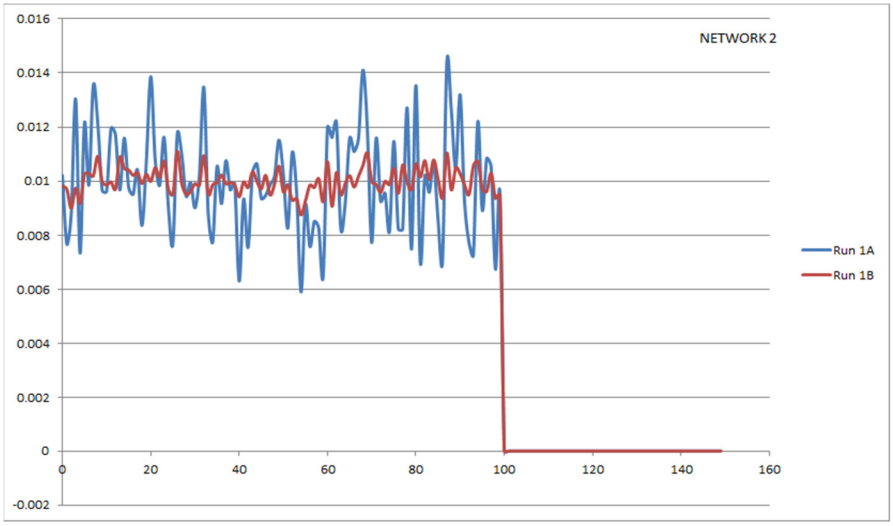
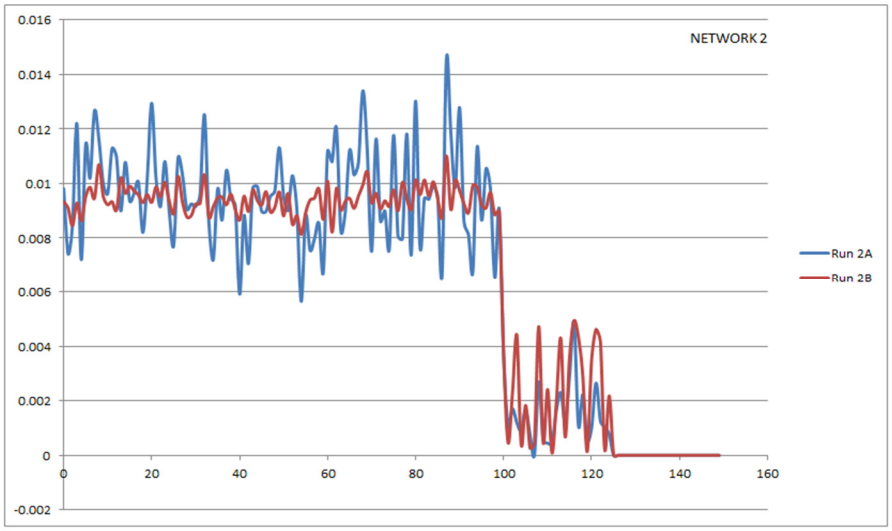
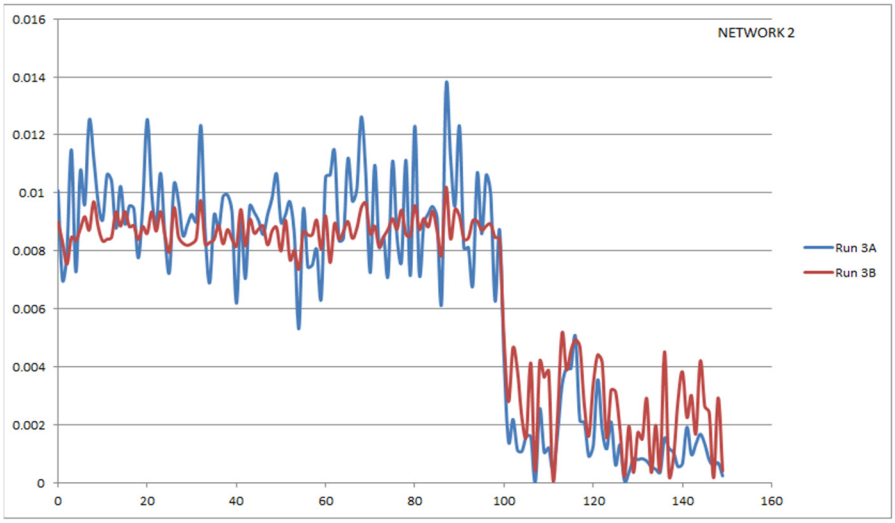

NETWORK 1
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
-0.002
0
20
40
60
80
100
120
140
160
Run 3A
Run 3B

Figure 3:


NETWORK 2
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
-0.002
0
20
40
60
80
100
120
140
160
Run 1A
Run 1B


NETWORK 2
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
-0.002
0
20
40
60
80
100
120
140
160
Run 2A
Run 2B


NETWORK 2
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
0
20
40
60
80
100
120
140
160
Run 3A
Run 3B

## 5. CONCLUSION

After the experiments and analysis, we have come to conclude that the normal assumption that as the knowledge of a network increases, the people become more knowledgeable is flawed. We witnessed a general trend that the relative knowledge of a person actually decreased in most cases when more people joined the network, thus bringing in more knowledge. Along the way, we also made many other interesting observations or explanations, the most notable of which is the Network Effect. Our experiments have led us to conclude that the relative knowledge of any one person in a society where the communication channels are very good and a free-flow of information is possible can never be significantly higher than others, however knowledgeable that person might be. Thus, we believe that this analysis will prove to be highly beneficial in the future understanding of the dynamics of social networks.

## 6. FUTURE WORK

Classifying knowledge into multiple domains can help us understand the dynamics of human information networks in a more realistic manner. More accurate estimation of the different parameters can be achieved through more exhaustive experimentation on larger and real-life data.