# Learning Influence in Complex Social Networks

Henry Franks
Department of
Computer Science
University of Warwick, UK
hpwfranks@gmail.com

Nathan Griffiths
Department of
Computer Science
University of Warwick, UK
N.E.Griffiths@warwick.ac.uk

Sarabjot Singh Anand
Algorithmic Insight, India
sarabjot.singh@gmail.com

## ABSTRACT

In open Multi-Agent Systems, where there is no centralised control and individuals have equal authority, ensuring cooperative and coordinated behaviour is challenging. Norms and conventions are a useful means of supporting cooperation in an emergent decentralised manner, however it takes time for effective norms and conventions to emerge. Identifying influential individuals enables the targeted seeding of desirable norms and conventions, which can reduce the establishment time and increase efficacy. Existing research is limited with respect to considering (i) how to identify influential agents, (ii) the extent to which network location imbues influence on an agent, and (iii) the extent to which different network structures affect influence. In this paper, we propose a general methodology for learning the network value of a node in terms of influence, and evaluate it using sampled real-world networks with a model of convention emergence that has realistic assumptions about the size of the convention space. We show that (i) the models resulting from our methodology are effective in predicting influential network locations, (ii) there are very few locations that can be classified as influential in typical networks, (iii) that four single metrics are robustly indicative of influence across a range of network structures, and (iv) our methodology learns which single metric or combined measure is the best predictor of influence in a given network.

## Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence — *Multiagent systems*; I.2.6 [**Artificial Intelligence**]: Learning

## Keywords

Influence, Networks, Conventions, Learning, Network value

## 1. INTRODUCTION

The interactions between agents in many open Multi-Agent Systems (MAS) domains are constrained by an underlying network connecting individuals. A wide body of research has shown that such networks often display a rich structure that significantly influences the dynamics of agent interactions and the flow of information (e.g. [6, 8, 25, 31, 33, 37]). Determining how these structures affect such processes is important in a variety of fields including computer science, biology, chemistry, sociology and economics. A key question is determining the *network value* of an agent, namely, some measure of its influence. This is often formulated as an *influence maximisation problem* where we aim to pick a (minimal) set of $k$ agents that would maximize the spread of information/behaviour through the population. In this paper, we propose a methodology for learning the network value of agents requiring only (i) a way of estimating the effective influence an agent exerts on a population, and (ii) the ability to sample a portion of a network. Our methodology is domain independent and can be applied 'offline' where the structure of a network is known or 'online' in networks where the structure is unknown but information can be obtained through some API. As such, it is more generally applicable than typical influence maximisation mechanisms.

In this paper, we determine the extent to which various topological metrics predict influence, and use Principal Components Analysis (PCA) and supervised learning to build models that predict influential locations. We evaluate our methodology on a selection of samples from real-world networks, and identify four metrics that effectively predict influence across a range of networks.

In MAS conventions can emerge through the "gradual accretion of precedence" [35], due to feedback effects in which an agent's choice in an interaction influences the choices of agents in the future. Influence is therefore important for convention emergence, since the actions of a highly influential agent are more likely to be reproduced in the rest of the population than a less influential agent. To evaluate our approach, we use a representative model of convention emergence, namely Salazar *et al.*'s [30] language coordination scenario. We also discuss possible sampling techniques, and the overheads inherent in applying our methodology and calculating each metric.

## 2. BACKGROUND

### 2.1 Influence propagation

Domingos *et al.* [4] were amongst the first to determine the *network value* of an individual, using Markov random fields to model markets of individuals. More recently, influence has been investigated in the context of Linear Threshold and Independent Cascade models [17], in which nodes are considered to be either *active* or *inactive*, where active rep-

**Appears in:** *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013),* Ito, Jonker, Gini, and Shehory (eds.), May 6–10, 2013, Saint Paul, Minnesota, USA.

resents believing an idea or adopting a convention. Nodes switch from inactive to active either based on the proportion of neighbours that are active, or by active nodes being given a chance to activate their neighbours. In general, solving the influence maximisation problem itself is NP-hard, however we can obtain approximate solutions. Watts [34] has shown that high degree nodes are more likely to cause cascade effects. Kempe *et al.* [17] showed that a simple hill-climbing algorithm can find a set of $k$ agents whose influence is within a known bound (slightly better than 63%) of the optimal, although their approach remains computationally expensive. More recently, Chen *et al.* [3] proposed a computationally cheap degree-discount heuristic, in which the node degree is discounted when its neighbours are already selected.

Influence maximisation has been extensively investigated (e.g. [12, 14]), but the extent to which these results generalise to open MAS is unknown. For example, a recent study concluded that node in-degree does not correlate with influence in the Twitter social network [20], placing empirical data at odds with the success of node degree in the influence maximisation literature. Although this is an isolated disagreement, we hypothesise that there are various facets of influence propagation that are missed by the influence maximisation problem formulation. Our methodology mitigates this by learning a model of influence from empirical data.

We are aware of relatively few investigations of influence in realistic open MAS. Sen and Airiau [31] investigated convention emergence with private interactions, showing that a small proportion of agents can exert significant influence, with a small convention space with no underlying connecting topology. Other investigations have also shown that a small number of individuals can influence a population [9, 28, 36]. Due to space constraints we omit a discussion of these investigations, however none have focused on the influence maximisation problem in the context of MAS.

## 2.2 Conventions

Conventions are considered to be social rules or standards of behaviour agreed upon by a set of individuals [18, 35]. Typically, a convention is considered to be established if 90% of a population adheres to it around 90% of the time [18]. Conventions can increase levels of coordination in MAS [15], and they are a powerful abstraction tool for modelling the aggregate interactions of agents.

Conventions can be generated offline by system designers or dynamically emerge through interactions. Generating conventions offline is difficult, due to limited knowledge of society characteristics, time variance, and computational expense. Such conventions also lack robustness to evolving populations and environments. As such, much research has concentrated on generating conventions online (e.g. [24, 30]), and this remains an open research problem. Conventions in open MAS are characterised by uniform levels of agent authority, lack of centralised institutions, and complex network structures restricting the interactions between individuals. Such domains are particularly suited to techniques in which specific agents are targeted or inserted to influence convention emergence.

Most models of convention emergence consider small convention spaces (i.e. a small number of possible conventions). Sen and Airiau [31] and Pujol *et al.* [29], for example, both use models with only two potential conventions. An exception is the work of Salazar *et al.* who explore convention emergence in a language coordination domain [30], containing $10^{10}$ possible conventions. Many real-world contexts have a large convention space, and so in this paper we adopt the language coordination domain. Franks *et al.* [9] recently introduced the *Influencer Agent* (IA) concept, in which a small proportion of agents are inserted with goals and strategies chosen specifically to influence the population to adopt particular conventions They demonstrate that (i) IAs can significantly manipulate which convention emerges in a population, even with a large convention space, and (ii) positioning IAs using topological information can increase their efficacy. While there has been some work investigating the role of network topology in convention emergence (e.g. [29]), to our knowledge previous work has not attempted to exploit intrinsic network influence in MAS.

# 3. METHODOLOGY FOR INFLUENCE ESTIMATION

In this section we present a general methodology for predicting the influence of an agent at a given location within a network. We assume the existence of a measure of influence, to be chosen depending on the domain. We also assume a network $G < V, E >$ where V is a set of agents and E is a set of edges that constrain the permitted communications between agents, the ability to sample locations within the network, and either global knowledge of the network or (more practically) the ability to sample smaller sub-networks around the nodes in question.

The offline instantiation of our methodology is as follows:

1. If necessary, sample a sub-graph $G_s \subset G$ from the network around selected locations to obtain a portion of the network of interest. In cases where the domain involves very large populations, this may be required to allow practical application of the methodology. For clarity, in the following steps we use $V$ and $E$ to refer to agents and edges in $G$ or $G_s$ as appropriate.

2. Sample a representative set $S \subset V$ of $n$ locations within the network, where $n << |V|$, where representative implies sampling nodes of both high and low influence.

3. Choose a measure of influence, and a model of influence propagation.

4. Compute the influence of an agent located at each of the locations in S, on the rest of the agent population (e.g. by running multiple simulations of the influence model).

5. Calculate topological location metrics for all nodes in $V$.

6. Build a model using the topological metrics and the estimated influence of agents placed at locations in S to *predict* which network locations are highly influential, and use it to predict the influence of all nodes in $V$.

To perform this methodology online, we modify step 3 as follows. Rather than running multiple simulations for each sampled node, select a measure of influence that can be measured online (e.g. if investigating Twitter, one might choose the number of re-tweets) and measure sufficient data for building the prediction model.

Once the influence of each agent in $S$ has been calculated (step 4), the agents in $S$ can be ranked according to influence. We cannot, however, rank the entire population, since we do not have a measure of influence for all agents. Instead, by building the prediction model (step 6) we are able to *estimate* the influence for any agent in the population based on its topological metrics.

The main computational expense in our methodology is the calculation of topological metrics for all nodes in $V$ and, if used, the influence model simulations for each location in $S$, which is $O(|S||E|k)$, where $k$ is the number of simulation cycles. Depending on the size of $S$ this can be significantly less than using the full network, which is $O(|V||E|k)$. Additionally, step 1 allows us to use a sample of the network to estimate influence, reducing the computational expense of the methodology. We recognise that the expense of computing the location metrics is varied and might be high, and while we do not explicitly account for this within the methodology, we discuss local and approximation algorithms for our selected metrics in Section 4.

In order to effectively predict influence our methodology requires a sample of nodes that reflect the range of influence and topological metrics in the network. If the influence distribution is highly skewed then a random sample will not be representative (we discuss this further in Section 7). Therefore, we propose selecting a sample by stratifying vertices using degree. Since degree is known to be indicative of influence [3], our hypothesis is that this approach will give a more representative sample in terms of influence.

To obtain a stratified sample, we divide the network into bins, and sort the nodes by degree. In this paper we use 10 bins, with a threshold of $N/10$ nodes per bin, where $N$ is the number of nodes in the network. We add all nodes of each degree into the current bin until the threshold is reached. If adding all nodes of a given degree pushes a bin over the threshold, we do not split the remainder but add all nodes of that degree. We then sample an equal number of nodes from each bin, until we reach our required sample size.

## 4. TOPOLOGICAL LOCATION METRICS

A wide range of metrics can be used for quantifying a node's position in a given network. However, based on the literature we hypothesise that the following metrics may be implicated in determining influence.

**Degree:** Node degree is commonly used as a proxy for influence [17], as intuitively the more individuals a node can communicate with the more it is able to influence.

**Local Clustering Coefficient (LCC):** LCC is a measure of the extent to which a node is embedded within a local cluster of nodes with high internal connectivity. It is defined as the proportion of neighbours of a given node that are also neighbours of each other.

**Lowest, Highest and Average Edge Embeddedness:** Edge embeddedness is the number of neighbours two endpoints of an edge have in common. A highly embedded edge is indicative of high levels of community structure in the local area. We define three metrics based on embeddedness: Lowest (LEE), average (AEE) and highest (HEE) edge embeddedness, taken as the lowest, average and highest value of embeddedness a node exhibits over all its edges respectively.

**Lowest, Average and Highest Edge Overlap:** Overlap is the proportion of nodes that are neighbours of both endpoint nodes to the total number of neighbours of both nodes (i.e. a normalised form of edge embeddedness). Again, we define three measures: lowest (LEO), average (AEO) and highest (HEO) edge overlap.

**Average Shortest Path Length (ASPL):** The average shortest path length from a given node to any other node in the network.

**Average Neighbour Degree (AND):** The average degree of the nodes that are connected to a given node.

**Closeness Centrality (CC):** Closeness centrality is the reciprocal of the average shortest path length from a node to every other node in the network. By taking the reciprocal, higher closeness centrality indicates that the node is more 'central', in the sense that it has a lower average path length to other nodes.

**Betweenness Centrality (BC):** Betweenness centrality is defined as the fraction of shortest paths between all node pairs in the network that a given node lies on. As such, a node with high betweenness is intuitively influential by virtue of being a conduit for more information flows [27].

**Eigenvector Centrality (EC):** Eigenvector centrality is the amount of time a random walk across the network spends at a given node. It is often interpreted as a measure of influence or importance of a node, since it weights connections to other highly valued nodes more highly than low value nodes. Google's PageRank algorithm is a variant of this measure [27].

**Hyperlink-Induced Topic Search (HITS):** Initially introduced by Kleinberg [19] in analysis of link structure on the world wide web, HITS attempts to determine *hubs* and *authorities* in a network, where a hub is a node that links to many authorities, and an authority is a node that is linked to by many hubs. Each node is assigned an authority score which is used as a topological feature.

In total we evaluate 14 metrics for the extent to which they predict node influence. Broadly, each metric can be linked to influence as follows. HITS, ASPL, and the centrality measures (EC, BC, and CC) measure the ability of a node to manipulate information flow in a network. LCC, embeddedness, and overlap measure the extent to which a node is part of a cluster of nodes. Clusters have efficient internal information propagation, and nodes in clusters are likely to be able to influence that cluster effectively. Degree is a measure of how many nodes a given individual is able to directly influence, while AND is a measure of how many nodes a given individual can indirectly influence to a depth of 2.

While several of these metrics are highly tractable, some require (i) global knowledge of the network, (ii) significant computational resources, or both. We cannot typically expect to provide both, but approximations exist for the less tractable metrics. The most significant difficulty is with the centrality measures, which typically require both full knowledge of the network and significant computational resources.

Gregory [13] has proposed *h-betweenness*, a local measure that considers paths of maximum length $h$. Computation

| Sampling mech. | BC | | EC | | CC | |
|---|---|---|---|---|---|---|
| | $h = 2$ | $h = 3$ | $h = 2$ | $h = 3$ | $h = 2$ | $h = 3$ |
| BFS | 0.90 | 0.96 | 0.36 | 0.42 | -0.37 | 0.41 |
| SNS | 0.75 | 0.93 | 0.14 | 0.36 | -0.49 | 0.62 |
| MHRW | 0.02 | 0.02 | 0.50 | 0.61 | -0.72 | -0.74 |
| MHRWDA | 0.02 | 0.02 | 0.51 | 0.61 | -0.73 | -0.74 |

**Table 1: Correlation between estimated centrality using h-betweenness and actual centrality, for Betweenness Centrality (BC), Eigenvector Centrality (EC), and Closeness Centrality (CC).**

| | Local Data | | Global Data |
|---|---|---|---|
| Metric | Computable | Approximatable | Fast approx. |
| Degree | ✓ | | |
| LCC | ✓ | | |
| EE | ✓ | | |
| EO | ✓ | | |
| AND | ✓ | | |
| BC | x | ✓ | |
| CC | x | x | ✓ |
| EC | x | ✓ | |
| HITS | x | x | ✓ |

**Table 2: Data requirements and computational tractability for the metrics we consider.**

involves calculating betweenness on a Breadth-First Search (BFS) induced sub-graph of depth $h$ around the target node. Table 1 shows the correlation between estimated centrality measures (for completeness, we include CC and EC) and the actual value over 15 networks of size 1000 in each of our datasets (see Section 5) for $h = \{2, 3\}$. The sampling method has a significant impact on estimation accuracy, showing that (i) estimation of metrics is sensitive to local structure, and (ii) each sampling technique produces very different network structures.

PageRank, a variant of EC, can be calculated using only local information and $O(e^{-1})$ nodes for a given error bound $e$ [2]. Given global knowledge, CC is computable in either $O(n^3)$ or $O(nm + n^2 logn)$. There exists a fast approximation algorithm but this still requires global knowledge [7]. While there are no known local algorithms for HITS, an approximation algorithm for HITS-like ranking algorithms gives considerable efficiency gains [11]. Calculating HITS does not require global knowledge, but requires a snowball sample around an initial seed set of nodes (around 200 for the WWW) [19].

Table 4 summarises this discussion — for clarity, we have categorised HITS as requiring global information, but note that it requires sampling of a portion of the global network rather than the entire network itself.

## 5. NETWORK SAMPLING ALGORITHMS

A wide variety of synthetic network generators have been proposed, but tend to be poor models of real-world networks [22, 26]. Use of real-world networks is typically constrained by (i) impractically large node counts, and (ii) limited knowledge of the global structure. Consequently, sampling part of the network is often necessary.

Ideally, the sampled structure should display similar properties to the full network, including clustering coefficient, average degree, degree distribution [10], and edge embeddedness distribution [32]. A wide variety of sampling techniques have been proposed (e.g. [10, 16, 21]). To evaluate our approach, we use Snowball Sampling (SNS), Metropolis-Hastings Random Walk (MHRW) [10], and MHRW with Delayed Acceptance (MHRWDA) [21]. Each starts with a randomly chosen node in a seed set. SNS iteratively adds neighbours to the sample at random from the neighbours of sampled nodes using a breadth-first search, until the threshold is reached. MHRW and MHRWDA perform a random walk with biased transition probabilities, with the aim of producing a uniform sample. MHRWDA uses modified MHRW transition probabilities to reduce the chance of backtracking.

SNS (and others, including BFS and vanilla random walks) are biased towards high node degrees [10], but SNS can produce good coverage of the local area around the start node. It is subject to greater variation between samples but may be useful for ensuring that a wide variety of structural properties are tested. MHRW and MHRWDA converge towards the node degree distribution exhibited in the full network, but there are no guarantees about the reproduction of any other metrics or structural properties.

In this paper, we use three networks: (i) a peer connection network from Gnutella (a P2P file-sharing platform), (ii) the Enron email dataset, and (iii) the arXiv general relativity section collaboration network[1]. The Enron and arXiv networks are both based on human interactions, but are generated by very different processes: the Enron dataset represents email communications, while arXiv is based on more formal links made through research collaborations. Conversely, Gnutella is a computational network of links in a P2P system. Since these networks are generated by very different processes they display varied structural properties, allowing us to evaluate our methodology on a range of structures. MHRW and MHRWDA sampling explicitly consider only undirected networks, and so we treat each network as undirected.

The high-level metrics are summarised in Table 3. Data for each sampling technique is averaged over 15 networks of 1000 nodes per sampling technique per dataset (for a total of 135 networks). The global clustering coefficient (GCC) is the average of the clustering coefficients for each node. Diameter describes the longest shortest path-length between any pair of nodes. Centralization measures how much heterogeneity exists in a graph [5], defined as:

$$Centralization = \frac{max(k)}{N} - \frac{mean(k)}{N-1}$$

where $k$ denotes node degree and $N$ the number of nodes. Centralization indicates the variation of node degree in the network — low centralization indicates that most nodes have a similar connectivity, whereas high centralization implies a higher degree of structural variation. Centralization is a useful indication of the extent to which a mechanism has generated a uniform sample.

No single technique produces an ideal sample. The standard deviation between samples is highest using SNS, indicating a large variation in structural properties between samples. Centralization is high using SNS, indicating a higher level of internal heterogeneity. Both MHRW and MHRWDA produce networks with metric values closer to

[1] All taken from the Stanford large network dataset collection, http://snap.stanford.edu/data/

| Graph | Nodes | Edges | Avg.Degree | GCC | Diameter | Centralization |
|---|---|---|---|---|---|---|
| Gnutella-FULL | 62586 | 147892 | 4.726 | 0.005 | - | 0.001 |
| Gnutella-SNS | 1000 (0) | 1197.8 (54.5) | 2.40 (0.11) | 0.02 (0.008) | 7.33 (0.98) | 0.034 (0.008) |
| Gnutella-MHRW | 1000 (0) | 1122.3 (13.1) | 2.24 (0.03) | 0.008 (0.004) | 36.4 (4.4) | 0.005 (0.001) |
| Gnutella-MHRWDA | 1000 (0) | 1120.1 (10.8) | 2.24 (0.02) | 0.007 (0.003) | 38.7 (3.22) | 0.005 (0.001) |
| Enron-full | 36692 | 183831 | 10.02 | 0.497 | 13 | 0.037 |
| Enron-SNS | 1000 (0) | 7751 (3760) | 15.5 (7.5) | 0.44 (0.13) | 4.5 (0.92) | 0.51 (0.31) |
| Enron-MHRW | 1000 (0) | 4480 (443) | 8.96 (0.89) | 0.52 (0.03) | 11 (1.41) | 0.10 (0.02) |
| Enron-MHRWDA | 1000 (0) | 4495 (255) | 9.00 (0.51) | 0.52 (0.02) | 10.6 (1.04) | 0.10 (0.02) |
| arXiv-full | 5242 | 14496 | 5.526 | 0.530 | 17 | 0.014 |
| arXiv-SNS | 1000 (0) | 3663 (405) | 7.32 (0.81) | 0.53 (0.04) | 8.67 (1.18) | 0.06 (0.01) |
| arXiv-MHRW | 1000 (0) | 3561 (413) | 7.12 (0.83) | 0.57 (0.02) | 14.3 (1.04) | 0.05 (0.01) |
| arXiv-MHRWDA | 1000 (0) | 3190 (394) | 6.38 (0.79) | 0.58 (0.02) | 15.5 (1.41) | 0.04 (0.01) |

**Table 3: Summary of structural metrics, averaged over 15 networks for each sampling technique and dataset.**

the full network than SNS. However, the diameter of MHRW and MHRWDA is approaching that which we see in the full graph, due to the random walks covering large areas of the network. Since these networks no longer display the small-world property, we cannot assert that many of the structural properties of the full network are reproduced, beyond the node degree distribution (as discussed above).

To effectively evaluate our methodology we consider a range of network structures, and use a portfolio of network samples derived using a variety of sampling techniques. Using SNS allows us to run our model on samples which are more representative of localised areas of the full network, and the high variance between samples indicates that a wide variety of structural properties will be included. Using MHRW and MHRWDA allows analysis of samples which more accurately reproduce the full degree distribution, but we cannot make any assertions about other properties. Given the large diameters, there are likely to be other as-yet undocumented biases. In this paper, therefore, we sample from each of the Gnutella, Enron, and arXiv datasets, taking 5 samples using each of the SNS, MHRW, MHRWDA sampling methods, giving a total of 45 network samples.

# 6. EXPERIMENTAL SETUP

To evaluate our methodology, we adopt Salazar *et al.*'s model of language coordination [30] (introduced in Section 2). In each experiment, we insert a single fixed-strategy Influencer Agent (IA) [9] at a randomly chosen location and measure the extent to which the population converges on the strategy of the IA. This model exhibits a natural measure of influence, as described below.

From the three networks introduced above, we sample 45 sub-networks of 1000 vertices. We demonstrate our methodology on two sets of data: (i) 50 locations sampled at random from each network, and (ii) 50 locations sampled using a stratified approach. We run our simulation 20 times for each location, for a total of 1000 simulation runs per network sample. We measure the extent to which the agent at each location influenced the rest of the population and calculate the 14 metrics of location. We use Principal Components Analysis (PCA) for unsupervised learning and fit Linear Regression (LR) models for supervised learning. We run new simulations using the location predicted as most influential by each model and determine the extent to which influence has increased against random placement.

We use the Java Universal Network/Graph Framework[2] in our simulations and Cytoscape[3] for offline structural analysis of networks. Statistical analyses are performed using R[4] and Weka[5].

## 6.1 Language coordination domain

In the language coordination domain agents attempt to establish a social convention in the form of a shared vocabulary. We adopt the formulation of the domain described by Salazar *et al.* [30], in which agents are associated with a *lexicon* that maps *words* to *concepts*. We use their parameter settings of 10 concepts and 10 words, with 10 mappings per lexicon (giving a convention space of size $10^{10}$). Each timestep, three phases are executed. First, each agent in turn communicates a single mapping from its lexicon to a single randomly chosen neighbour. It is assumed that agents can determine whether the recipient's lexicon contains the same mapping, in which case the communication is *successful*, otherwise it is *unsuccessful*. Second, each agent has a chance to propagate part of its lexicon to all of its neighbours, along with the *communicative efficacy* of the lexicon, defined as the proportion of successful communications in the last 20 communications. Third, each agent has a chance to update their internal lexicon based on the partial lexicons received from their neighbours, using a two-point crossover. Agents use an *elitist* strategy, such that they update their lexicon with the received mappings that have the highest communicative efficacy.

Over time, a shared lexicon (or set of lexicons) emerges. We define the *dominant lexicon* as the one that is shared by the highest number of agents. Each simulation is run for 50000 timesteps, and each agent propagates their lexicon with a probability of 0.01 and updates their lexicon with a probability of 0.01. By the end of a typical simulation run 600–800 agents have adopted the dominant lexicon.

In this model, we define an agent's *influence* as the similarity between its lexicon ($L$) and final dominant lexicon in the population ($L'$) using Jaccard's similarity coefficient: $J(L, L') = |L \cap L'|/|L \cup L'|$, where a similarity of 1 implies that agents use an identical lexicon, and 0 implies that there are no mappings in common.

# 7. RESULTS

In this section, we analyse the predictive power of individual metrics, and apply our methodology by constructing a number of models to predict influence.

[2] http://jung.sourceforge.net/

[3] http://www.cytoscape.org/

[4] http://www.r-project.org/

[5] http://www.cs.waikato.ac.nz/ml/weka/

| Network | Average lexicon similarity | | | | | Number of wins (normalised) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Degree | HE | EC | HITS | Random | Degree | HE | EC | HITS | Random |
| arXiv-SNS | 0.58 | 0.54 | **0.6** | 0.54 | 0.16 | 0.47 | 0.41 | **0.5** | 0.40 | 0.03 |
| arXiv-MHRW | 0.54 | **0.6** | 0.56 | 0.58 | 0.18 | 0.41 | **0.50** | 0.47 | 0.42 | 0.02 |
| arXiv-MHRWDA | **0.6** | 0.58 | 0.56 | 0.48 | 0.16 | **0.5** | **0.5** | 0.47 | 0.37 | 0.02 |
| Enron-SNS | **0.58** | 0.56 | 0.48 | 0.54 | 0.16 | 0.45 | **0.47** | 0.34 | 0.42 | 0.02 |
| Enron-MHRW | 0.62 | 0.48 | 0.55 | **0.63** | 0.16 | 0.55 | 0.41 | 0.44 | **0.56** | 0.02 |
| Enron-MHRWDA | 0.56 | 0.58 | **0.6** | **0.6** | 0.16 | 0.47 | 0.47 | **0.6** | **0.6** | 0.02 |
| Gnutella-SNS | 0.4 | 0.3 | 0.38 | **0.5** | 0.18 | 0.30 | 0.18 | 0.25 | **0.41** | 0.06 |
| Gnutella-MHRW | 0.22 | 0.18 | 0.2 | **0.38** | 0.16 | 0.08 | 0.05 | 0.06 | **0.21** | 0.02 |
| Gnutella-MHRWDA | **0.3** | 0.2 | **0.3** | 0.34 | 0.18 | **0.2** | 0.06 | 0.17 | 0.18 | 0.04 |

**Table 4: Results for placing a single IA at a location maximising a chosen topological metric. The best performing metrics in each row are shown in bold.**

Inspecting the extent to which individual metrics predict influence may allow us to refine our models, and analysis of the correlations between each metric and influence reveals that Degree, EC, HEE, and HITS all robustly correlate with influence over all networks. These metrics are statistically significantly correlated in over 90% of the networks (with correlations ranging from 0.68 in the arXiv networks to 0.27 in the Enron networks), whereas the other metrics statistically significantly correlate only in isolated networks (on average, in 48% of networks). Correlating with influence in isolated networks is likely to be due to unique network structures, and these metrics are less likely to indicate influential nodes in the general case. This corroborates previous research on the link between node degree and influence (e.g. [3]), but to our knowledge this is the first time that EC, HEE and HITS have been shown to predict influence.

Ranking nodes by each of the four identified metrics results in significant overlap over the top 5 vertices — with 7.8 unique nodes over the top 5 for each metric (a 0.39 proportion, standard deviation 0.15), where disjoint sets would give 20 unique nodes. While each metric selects roughly similar sets as most influential, their relative rankings are unique. Figure 1 plots normalised EC, HEE and HITS against degree, from which we can see the correlations. Interestingly, HEE and HITS clearly bisect the population, which may be useful for splitting a population into influential and non-influential nodes, while EC has an approximately linear relationship with degree.

Table 4 shows the results of placing an IA at the location that maximises each heuristic, where a *win* is defined as a simulation run in which the dominant lexicon in the population has at most 2 different mappings from the IA lexicon. Results are averaged over each class of network. We see significant gains across all four metrics, particularly in the arXiv and Enron networks. With random placement, an agent is only able to successfully influence the population 2 times in 100, but placing by heuristic can increase this to 60 times in 100. There is no consistency in which metric performs best, and this is likely due to unique network structures in each class of network.

We subsequently apply our methodology by learning three models: (i) the Principal Component (PC) that most correlates with influence, (ii) a Linear Regression (LR) model on all 14 metrics, and (iii) a linear regression model on Degree, EC, HEE and HITS (4LR), which are the best 4 heuristics as discussed above. We consider two sampling approaches for selecting a representative set of nodes: random and stratified (as described in Section 3).

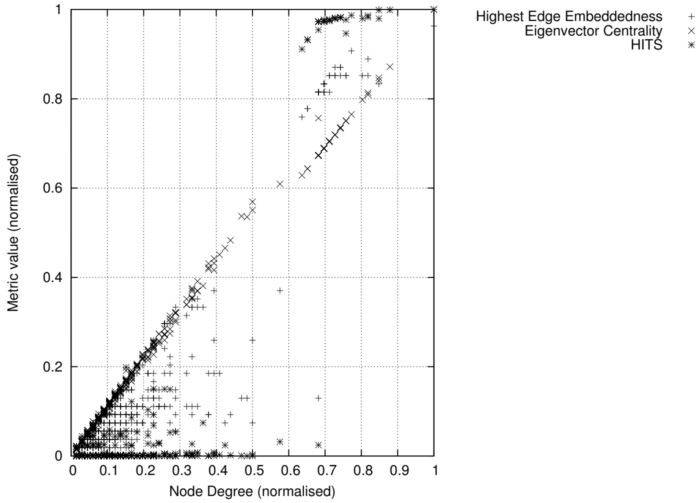

**Figure 1: Correlation of HEE, EC, and HITS with node degree in an example arXiv-SNS sample.**

Table 5, shows the correlations between predicted influence and actual influence, demonstrating that models learnt on randomly sampled vertices are particularly poor. Conversely, learning on stratified data shows high correlations, indicating higher quality models. This corroborates our hypothesis that there are relatively few nodes of influence in a network, with the majority having similar and low influence. This indicates that a random sample is not representative in terms of influence, and so the stratified approach should be used. Figure 2 plots the predicted influence in an arXiv-SNS sample using the LR model on a stratified sample, and we can clearly see that less than 10% of the vertices account for almost all the influence. The Gnutella-MHRW samples do not fit a model using randomly sampled data, since the majority of nodes are zero-valued for many of the metrics. This occurs for nodes of very low degree, and is an extreme example of the effect discussed above.

To evaluate the efficacy of each prediction model, we place an IA at the location predicted as most influential by each model, and repeat the simulations. Table 6 shows the results using models learnt on stratified sampling. We have omitted results for models learnt using random sampling, since they are less effective: across all networks, average lexicon distance is 0.35 (standard deviation 0.1) and the average proportion of wins is 0.2 (standard deviation 0.12). Nodes selected as influential by these models exhibit less than half the influence of those selected by either individual heuristics

|  | Random |  |  | Stratified |  |  |
|---|---|---|---|---|---|---|
| Network | PC | LR | 4LR | PC | LR | 4LR |
| arXiv-SNS | -0.08 | 0.164 | 0.19 | 0.71 | **0.91** | 0.90 |
| arXiv-MHRW | 0.018 | 0.10 | 0.11 | 0.67 | **0.93** | 0.92 |
| arXiv-MRHWDA | -0.03 | 0.34 | 0.08 | 0.75 | **0.88** | 0.86 |
| Enron-SNS | -0.03 | 0.16 | 0.13 | 0.69 | 0.80 | **0.85** |
| Enron-MHRW | -0.10 | 0.17 | 0.21 | 0.71 | **0.88** | 0.87 |
| Enron-MHRWDA | 0.08 | 0.06 | -0.03 | 0.73 | **0.90** | 0.89 |
| Gnutella-SNS | -0.01 | 0.03 | -0.10 | 0.33 | **0.67** | 0.58 |
| Gnutella-MHRW | 0.02 | - | - | 0.44 | **0.75** | 0.65 |
| Gnutella-MHRWDA | 0.09 | -0.15 | 0.06 | 0.36 | **0.73** | 0.52 |

**Table 5: Correlation of each model with influence, using separate training and test data.**

| Network | Average lexicon similarity |  |  |  | Number of wins (normalised) |  |  |  |
|---|---|---|---|---|---|---|---|---|
|  | PC | LR | 4LR | Ran. | PC | LR | 4LR | Ran. |
| arXiv-SNS | 0.44 | 0.42 | **0.58** | 0.16 | 0.34 | 0.30 | **0.50** | 0.03 |
| arXiv-MHRW | 0.5 | 0.32 | **0.62** | 0.18 | 0.42 | 0.20 | **0.55** | 0.02 |
| arXiv-MHRWDA | 0.34 | 0.38 | **0.6** | 0.16 | 0.22 | 0.27 | **0.50** | 0.02 |
| Enron-SNS | 0.62 | 0.32 | **0.68** | 0.16 | 0.56 | 0 | **0.62** | 0.02 |
| Enron-MHRW | 0.2 | 0.5 | **0.58** | 0.16 | 0.30 | 0.36 | **0.53** | 0.02 |
| Enron-MHRWDA | 0.34 | 0.16 | **0.52** | 0.16 | 0.21 | 0.06 | **0.43** | 0.02 |
| Gnutella-SNS | 0.18 | **0.46** | 0.36 | 0.18 | 0.03 | **0.37** | 0.24 | 0.06 |
| Gnutella-MHRW | **0.4** | **0.4** | 0.24 | 0.16 | 0.27 | **0.29** | 0.10 | 0.02 |
| Gnutella-MHRWDA | **0.38** | 0.36 | 0.36 | 0.18 | **0.25** | 0.24 | 0.22 | 0.04 |

**Table 6: Results for placing an IA at a location chosen by the predictive models. The best performing placement strategies are shown in bold.**

or the models learnt from stratified sampling, indicating that random sampling of nodes does not give a sufficient range of influential nodes to generate accurate models.

Targeting IAs using locations predicted as influential by models based on stratified data results in significant gains in influence. In the arXiv, Enron and Gnutella-SNS networks, these increases are roughly equal to that gained by placing by single metric compared to random placement. In the arXiv and Enron networks, the best performing model is 4LR, indicating that the other metrics are unlikely to contribute to influence prediction. We believe that 4LR is learning *which* metric is best to place by, given the results in Table 4. In Gnutella, 4LR is always outperformed by PC or LR, indicating that metrics other than Degree, EC, HEE and HITS are indicative of influence in these networks. Moreover, the linear combination of metrics in these networks outperforms placement by single metrics. The Gnutella networks show reduced potential for influence compared to Enron and arXiv, and exhibit lower edge counts, average degree, and clustering coefficients, and higher diameters. All these properties reduce the ability of an agent to exert influence, and may provide an indication of the likely efficacy of our methodology prior to application.

Our results suggest that if computational expense is an issue, targeting by Degree (or EC, HEE or HITS) will yield significant gains in influence, but if computational expense is less important then applying our methodology results in further gains. If our methodology is applied using online measurements of influence (i.e. not requiring repeated simulations), the computational cost is significantly reduced.

# 8. CONCLUSIONS AND FURTHER WORK

In this paper, we describe a methodology for learning the influence of nodes in a network. We evaluate our model using a representative model of convention emergence on networks sampled using a variety of techniques from three base datasets. We corroborate results in the literature that degree is highly indicative of influence, and additionally also show that Eigenvector Centrality, Highest Edge Embeddedness, and HITS are linked to influence. Applying our proposed methodology gives significant gains in influence. In the arXiv and Enron networks, linear regression on these 4 metrics gives the best results and is comparable with the gains in influence from placing using single metrics over random placement, indicating that the model learns which metric best predicts influence in that network. In the Gnutella networks, our models outperform single metric placement. Supervised learning using LR almost always outperforms unsupervised learning using PCA.

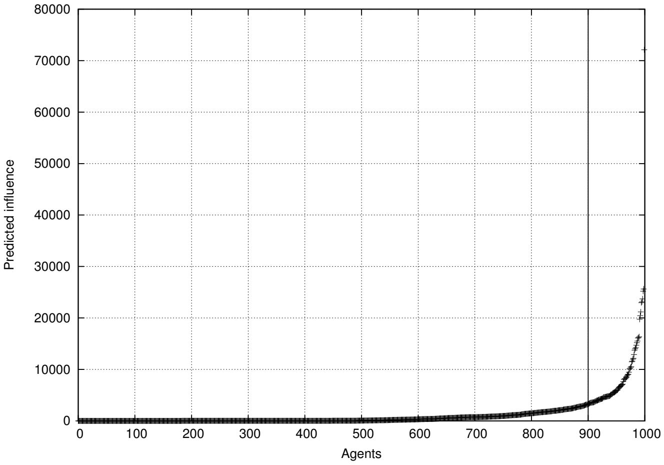

**Figure 2: Predicted node influence in an arXiv-SNS sample.**

The form of influence in the arXiv and Enron samples and the Gnutella samples are significantly different. The Gnutella samples demonstrate (i) that single metric heuristics do not guarantee optimal influence, and (ii) that different network structures result in significantly varied ranges of influence. We believe that the overall network metrics (such as average degree, clustering coefficient, or diameter) may indicate the potential for maximising influence in a given network, and we intend to test this in future work, along with other models of influence propagation to ensure our methodology generalises.

# 9. REFERENCES

[1] R. Albert and A.L. Barabási, 'Statistical mechanics of complex networks', *Reviews of Modern Physics*, **74**:47–97, (2002).

[2] R. Andersen, C. Borgs, and J. Chayes, 'Local computation of PageRank contributions', in *Proc. of the 5th Int. Conf. on algorithms and models for the web-graph*, pp. 150–165, (2007).

[3] W. Chen, Y. Wang, and S. Yang, 'Efficient influence maximization in social networks', in *Proc. of the 15th Int. Conf. on Knowledge Discovery and Data Mining*, pp. 199–208, (2009).

[4] P. Domingos and M. Richardson, 'Mining the Network Value of Customers', *Proc. of the 7th Int. Conf. on Knowledge Discovery and Data Mining*, pp. 57–66, (2001).

[5] J. Dong and S. Horvath, ‘Understanding network concepts in modules.’, *BMC systems biology*, **1**:24, (2007).

[6] D. Easley and J. Kleinberg, *‘Networks, crowds, and markets: Reasoning about a highly connected world’*, Cambridge University Press, (2010).

[7] D. Eppstein and J. Wang, ‘Fast approximation of centrality’, in *Proceedings of the 12th ACM-SIAM symposium on discrete algorithms* , pp. 228–229, (2001).

[8] Z. Fagyal, S. Swarup, A. Maria Escobar, K. Lakkaraju, and L. Gasser, ‘Centers and peripheries: Network roles in language change’, *Lingua*, **120**(8):2061–2079, (2010).

[9] H. Franks, N. Griffiths, and A. Jhumka, ‘Manipulating convention emergence using influencer agents’, *Journal of Autonomous Agents and Multi-Agent Systems*, 26(3):315-353, 2013.

[10] M. Gjoka, M. Kurant, C. Butts, and A. Markopoulou, ‘Walking in facebook: A case study of unbiased sampling of OSNs’, in *Proc. of the 29th Conf. on Information Communications* , pp. 2498–2506, (2010).

[11] S. Gollapudi, M. Najork, and R. Panigrahy, ‘Using bloom filters to speed up HITS-like ranking algorithms’, in *Proceedings of the 5th Int. Conf. on Algorithms and models for the web-graph*, pp. 195–201, (2007).

[12] A. Goyal and F. Bonchi, ‘A data-based approach to social influence maximization’, *Proceedings of the VLDB Endowment*, 5(1):73–84, (2011).

[13] S. Gregory, ‘Local Betweenness for Finding Communities in Networks’, Technical Report, University of Bristol (2008).

[14] B. Hajian and T. White, ’On the Interaction of Influence and Trust in Social Networks’, in *Proceedings of the 1st Workshop on Incentives and Trust in E-Commerce*, pp. 63–75 (2012).

[15] N. Jennings, ‘Commitments and conventions: The foundation of coordination in multi-agent systems’, *The Knowledge Engineering Review*, **8**(03):223–250, (1993).

[16] L. Jin, Y. Chen, P. Hui, C. Ding, T. Wang, A. V. Vasilakos, B. Deng, and X. Li, ‘Albatross Sampling : Robust and Effective Hybrid Vertex Sampling for Social Graphs’, in *Proceedings of the 3rd ACM Int. workshop on MobiArch*, pp. 11–16, (2011).

[17] D. Kempe and J. Kleinberg, ‘Maximizing the spread of influence through a social network’, *Proc. of the 9th Int. Conf. on Knowledge Discovery and Data Mining*, pp. 137–146, (2003).

[18] J. Kittock, ‘Emergent conventions and the structure of multi-agent systems’, in *Artificial Intelligence*, **141**(1-2):171–185, (1993).

[19] J. Kleinberg, ‘Authoritative Sources in a Hyperlinked Environment’, *J. of the ACM*, **46**(5):604–632, (1999).

[20] A. Khrabrov and G. Cybenko, ‘Discovering Influence in Communication Networks Using Dynamic Graph Analysis’, in *Proc. of the IEEE 2nd Int. Conf. on Social Computing*, pp. 288–294, (2010).

[21] C.-h. Lee, X. Xu, and D. Y. Eun, ‘Beyond Random Walk and Metropolis-Hastings Samplers: Why You Should Not Backtrack for Unbiased Graph Sampling’, in *Proceedings of the 12th ACM Koint Int. Conf. on Measurement and Modeling of Computer Systems*, pp. 319–330, (2012).

[22] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, ‘Statistical properties of community structure in large social and information networks’. *Proc of the 17th Int. Conf. on World Wide Web*, pp. 695–704, (2008).

[23] P. V. Marsden, ‘Egocentric and sociocentric measures of network centrality’, *Social Networks*, 24(4):407–422, 2002.

[24] J. Morales, M. López-Sánchez, and M. Esteva, ‘Using Experience to Generate New Regulations’, in *Proc. of the 22th Int. Conf. on Artificial Intelligence*, pp. 307–312, (2011).

[25] E. Mossel and S. Roch. Submodularity of Influence in Social Networks: From Local to Global. *SIAM Journal on Computing*, 39(6):2176–2188, (2010).

[26] M. Newman, ‘The structure and function of complex networks’, *SIAM review*, 45(2):167–256, (2003).

[27] M. Newman, ‘The mathematics of networks’, *The New Palgrave Dictionary of Economics* (2008)

[28] J. Oh and S. Smith, ‘A few good agents: multi-agent social learning’, in *Proc. of the 7th Int. Conf. on Autonomous Agents and Multiagent Systems*, pp. 339–346, (2008).

[29] J. Pujol, J. Delgado, R. Sangüesa, and A. Flache, ‘The role of clustering on the emergence of efficient social conventions’, *Proc. of the 19th Int. Conf. on Artificial intelligence*, pp. 965–970, (2005).

[30] N. Salazar, J. Rodriguez-Aguilar, and J. Arcos, ‘Robust coordination in large convention spaces’, *AI Communications* **23**(4):357–372, (2010).

[31] Q. Sen, ‘Scale-free topology structure in ad hoc networks’, in *Proc. of the 11th Int. Conf. on Communication Technology*, pp. 21–24, (2008).

[32] A. Sridharan and J. Nastos, ‘Statistical Behavior of Embeddedness and Communities of Overlapping Cliques in Online Social Networks’, *Distribution*, in *Proc. of the 30th Int. Conf. on Communication Technology*, pp. 546–550, (2011).

[33] D. Villatoro, N. Malone, and S. Sen. ‘Effects of interaction history and network topology on rate of convention emergence’, in *Proceedings of 3rd Int. Workshop on Emergent Intelligence on Networked Agents*, pp. 13–19 (2009).

[34] D. Watts, ‘A simple model of global cascades on random networks’, *Proc. of the National Academy of Sciences*, **99**(9), pp. 5766–5771, (2002).

[35] H. Young, ‘The economics of convention’, *The J. of Economic Perspectives*, **10**(2):105–122, (1996).

[36] C. Yu, J. Werfel, and R. Nagpal, ‘Collective decision-making in multi-agent systems by implicit leadership’, in *Agents and Multiagent Systems*, volume 3, pp. 1189–1196, (2010).

[37] M. G. Zimmermann and V. M. Eguíluz. ‘Cooperation, social networks, and the emergence of leadership in a prisoner’s dilemma with adaptive local interactions’, *Physical review. E, Statistical, nonlinear, and soft matter physics*, 72(5 Pt 2):056–118, (2005).