Hidden Markov Models Models for identification of influence in Social Media

D. Simmie, M. G. Vigliotti, C. Hankin
Imperial College London, South Kensington Campus,
SW7 2AZ, London

Abstract

Influential agents in networks play a pivotal role in diffusing information. Generally well connected hubs they have the ability to spread ideas and shape the beliefs of others. Influential leaders may rise to their position quite quickly. Capturing this evolution of influence is of benefit to a varied number of application domains such as: counter-terrorism, policing or marketing. We propose a new model for capturing both absolute influence and also temporal influence. The unified model is a combination of network topological methods and observation of influence relevant effects on the network. We provide an application of Hidden Markov Models for capturing this effect on the network.

1 Introduction

We live in a digital world, and we are able to get information about nearly everything. We are surrounded by information all the time. One of the biggest challenges of our time is to make sense all the information around us. The aim of our work is to provide a sensible framework to model, analyse a large part of the information available around us. To achieve our goal, we follow the scientific method which is based on the principle that we observe a phenomenon, and we assume that the events related to the original phenomenon are generated acceding to a certain laws. We assume that it is possible to infer some, or all, laws that are generated by the observations. In this work, we assume that the information available to us has been generated by some laws, and our goal is to uncover the laws that govern the phenomenon.

Our general aim is to make sense of human behaviour, if at all possible. In particular we are interested in understanding how social interaction arise, and what are the dynamics that fuel such interaction. Our point of view is that we can observe certain interaction among humans, and collect data about that, and we can then understand, from a mathematical point of view, if there exist laws that govern human relationships. If such laws are found, then we can predict certain events, at least as far as large group of people is concerned. Prediction of the behaviour of a single person remains in the field of psychology and it is outside the scope of this work.

The job of collecting accurate data regarding social interaction has been greatly facilitated by the advent of *Online Social Media/Networks* . Online Social Networks such as *Twitter* and Facebook have provided a medium to rapidly organize and communicate to decentralized groups of people at a scale and velocity as yet unseen. Recent years have seen several high profile protests and demonstrations which have used social media to diffuse messages throughout the protest network. It has been speculated, with respect to some social movements, that an acute observer of social medias, could have predicted them. In this paper, we deal with the matter of whether this hypothesis could be justified. We consider a particular form of human expression *influence*. Influence has been described [1] as:

- *the capacity or power of persons or things to be a compelling force on or produce effects on the actions, behavior, opinions, etc., of others*

We also define the two key roles in an influence network: *influencer* and *influncee*. The *influencer* has a measurable effect on the behaviour of the *influencee*.

Influential users are important in shaping the actions and behaviour of the peers that they have influence over. As contemporary social networks like *Facebook* and *Twitter* grow to hundreds of millions of users it is possible for a single influential entity to reach a vast number of individuals and have a large effect on the actions and behaviour of their peers and hence by extension, society at large.

Identifying influential agents in a network has many potential applications. For covert networks this may focus the effort of surveillance on those suspected on being leaders within the organisation [12]. In networks that are characterised by influence relationships, such as *Twitter* identifying those *influencers* may aid in the discovery of missing edges in the network.

[1] Source: http://dictionary.reference.com/browse/influence Access Date: 09/10/2012

To robustly model the evolution of influence problem we have used real data from a sampled *Twitter* network. A discussion of the sampling technique and data collection can be found in section 3.1. We examine other research into influence in *Twitter* in section 1.2 and outline our experimental approach in section 3.2. We conclude by evaluating our results in section **??** and summarising our contribution in 4.

1.1 The problem

In our society we regard a leader as a person who can influence a large number of people. A leader is not born instantaneously, but his influence is built over time.

If we look at social media present in our society, we can consider that as users modify their status over time, we wish to predict some of these modifications. Mathematically we can think of each modification that affects social media as a random variable. The changes over time of a family of random variables would form a stochastic process, and the aim is to look at correlation among various stochastic processes to make predictions about how people might behave.

We define a leader, or a person of great influence somebody who has a *great number* followers(*influencees*). Followers(*influencees*) are persons that change their behaviour in reaction to something that the leader has done. A leader, generally, gains his/her influence over a period of time, and *Twitter* offers a great platform to identify the characteristics, or the signs of individuals that will become leaders. Furthermore, influence is not a stationary property, it can be lost if not exercised. This lead to the conclusion that we need a temporal model to accurate capture the behaviour of influence. What our model should be able to capture is the temporal evolution of the influences, its rising, decline. Ideally we also aim to have some sort of predictive model, such that it could be possible to identify a group of people that are arising to become influentials.

1.2 Related work

The majority of work quantifying influence in Social Networks is either: network topology based [1, 9, 18] or from direct observations (e.g. *retweet* counts) [5, 7, 13]. Some research examines both [7, 13] however currently no model has combined both the structural properties of the network and the influence relevant observables is not a unified model, or has captured the temporal evolution of influence.

Cha *et al.* [7] examine three different types of influence in the *Twitter* network: in-degree, *retweets* and mentions. There was little overlap in the top ranked users by each type of influence metric, only *retweets* and mentions displayed any correlation. One of the key finding of this research on near-complete *Twitter* data is that the most popular users (via in-degree) are not always the most influential at diffusing content or stimulating an engaged audience. Hence we posit that the current metric used by *Twitter* to suggest people to follow, in-degree, has it's shortcomings as a unified measure of interest. A model combined from network structure and relevant observables could capture those users who are both popular (of interest) and are visibly affecting the network.

The PageRank algorithm [14] provides a means for ranking Web pages according to how interesting they are in a network. PageRank uses the sum of the ranks of the in-degree of a node to determine the importance of a node. One of the main differences between in-degree as a measure of importance and PageRank is that a node with a small number of very important backlinks can be just as interesting as a node with a large amount of relatively unimportant in-links. The principal idea behind PageRank also holds for Social Networks, for example a hypothetical user with only one important follower such as Stephen Fry [2] has the ability to reach a user-base several orders of magnitude higher than a user with 500 followers who each have no followers.

Derivatives of PageRank are popular for quantifying influence in *Twitter* [18, 9, 1]. The Topic-Sensitive PageRank algorithm captures importance with respect to a particular topic. TSPR differs from PageRank by maintaining a set of vectors, one for each topic in place of PageRank's single rank vector. TSPR ranks pages according to a biased set of 16 topic vectors (taken from ODP [2]), maximum likelihood estimation is used to discover the topic of a query given the query terms.

Similarly to PageRank, TSPR was developed for Web page ranking for query engines, *TwitterRank* [18] was devised for ranking influential users in the *Twitter* network. It provides a topic based ranking of influentials. Weng *et al.* observe high link reciprocity in their sample and postulate that this would also be true of the population - a claim disputed by the near-complete data presented in [7] which records low link reciprocity (10%). Cha states that this low link reciprocity means that the *Twitter* network displays more of influence relationship than that of homophily - suggested by Weng. *TwitterRank* performs topic distillation (Latent Dirichlet Allocation) on a user's

[2] c. 5.5 M followers as of 07-Mar-2013

tweets to determine the topics they are interested in. It then uses a topical similarity function to alter the transition probability model of the random surfer from follower to friend. This provides the topical influence for a user, general influence is the summation of the user's individual *TwitterRanks* by the amount of weight associated with that rank.

Kwak *et al.* [13] examined three different influence metrics: in-degree, PageRank and *retweets* . They found in-degree and PageRank to be related, by applying a generalized version of Kendall's tau rank correlation. Similarly to [7] they found the influence ranking produced from *retweets* was sufficiently different form either of the two topology based approaches. They also analysed the active period of trends on *Twitter*. An active period is a period of trend activity whereby the trend is discussed at least once every 24 hours. Most (73%) topics were found to have a single active period and the majority of the active periods were a week or shorter (˜70%).

Bakshy *et al.* [5] use shortened URL diffusion cascades as their quantification of influence on *Twitter* . The source of a shortened URL is the seed of that referral within the network. They captured data that contained bit.ly [3] URLs posted in *tweets*. The number of nodes of the URL diffusion tree signifies influence, those users producing the largest cascades are the most influential. Regression trees are used to provide a prediction of future *influencers* (or future influential URLs). The prediction is found to be unreliable and the authors suggest targeting a number of "ordinary *influencers*" so as to benefit from averaging effects and to provide the best cost benefit for social marketing strategies.

Two additional influence ranking mechanisms, which are available on the web are: TunkRank [1] and Trst.Rank and Trst.Quotient [3]. TunkRank is an interesting extension to PageRank for the *Twitter* domain. It takes the probability of *retweet* into account and the attention span of a user and is summarised by the following equation:

- $Influence(X) = \sum_{Y \in Followers(X)} \frac{1+p*Influence(Y)}{|Friends(Y)|}$

The Trst.Rank is a logarithmically scaled influence rank, the derivation details of which are not revealed publicly. The Trst.Quotient is used to differentiate between users of similar rank. It measures the trustworthiness of a user and a low quotient figure connotes a spam or abusive account. Our model also uses a combination of two influence measures but *Buzz* differs from to the Trst.Quotient as it measures the actual effect a user is having on the network not the perceived quality of their account.

[3]Now bitly.com

2 Approach

In the context of the on the analysis of *Twitter* , we propose the HMM with the aim of identify the temporal evolution of people who have high influence on a (social) group. In simple words, we wish to understand how a leader, or a person with massive influence develops over time. By contrast, a Bayesian classification technique, will allow us to establish, at a given time, whether a person is influential. In this sense it make sense to model influence as a *stochastic process*. The main reason to use probabilistic methods is that it can hardly determine with full certainty the behaviour of a large number of individuals, but there is hope that we determine some trend in social interactions with a certain probability at different points in time.

2.1 Potential models

Temporal probabilistic models, such as Markov Chains, describe the temporal evolution of a family of random variables. With respect to this simple definition, Hidden Markov Models (HHM), are also temporal models.

Static models, by contrast, specify probability of certain properties, without taking into consideration, from a probabilistic point of view, the time evolution of such property. For example, classification, which is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known, can be seen as a static model.

In the context of the analysis of *Twitter* , we have used HMM with the aim of identify the temporal evolution of people who have high influence on a (social) group. In simple words, we wish to understand how a leader, or a person with massive influence develops over time. By contrast, classification technique, will allow us to establish, at a given time, whether a person is influential.

The majority of current influence approaches, which are also non-temporal, for Social Networks involve examining only the structure of the network. The in-degree of a node is an effective approach for ascertaining popularity within a social network. However it has been shown that popularity is limited [7] in describing the influence capability of a node. Users may inherit this popularity from their real world persona and may or may not also be influential in the online space. There are at least four different classes of influence in Social Networks as summarised from [7]:

- Network Importance

- Content Value
- Name Value
- Engagement

In networks such as *Twitter* the in-degree of a node reliably captures their popularity as it is the count of users who have chosen to follow them. The other three measures of influence cannot be discovered by topological methods alone. For these influence relevant variables it is necessary to construct a model based on observations of the data.

We split these four metrics into two distinct properties of network influence:

- Reach
- Buzz

Reach, ρ, is an individual's capacity to affect other users in a network. For simplicity in this study we use PageRank [14] for reach, however we hope to examine alternative algorithms such as the *Twitter* specific TunkRank [1] in the future.

Buzz, β, is the actual observance of influence relevant behaviour in the target network. It is temporal.

There are two core challenges to capturing temporal influence. The first problem concerns the identification or classification of influence at any time point. Probabilistic latent variable models are an effective method for identifying an abstract hidden state from related observables [19]. However they lack inherent temporality and would need to be calculated again at each time point for all observables.

The second challenge concerns the evolution of those latent states from one time point to the next. Time-Series prediction techniques such as exponential smoothing would be able to provide a future value for our influence related observations but do not relate the observed value to the hidden state (level of influence) which is of interest to us.

We now explain how use HMMto capture influence, and buzz.

2.2 Hidden Markov Models

We assume the reader to be familiar with elementary notions of probability theory [11, 8, 10]. We use the capital letters X, Y, Z to indicate *random variables* and F, G for probability distribution functions. We write $\mathbb{P}(E)$ to indicate the probability of an event E and $\mathbb{P}(E|E')$ for the conditional

probability. A *stochastic process* $\{X(t) : t \in T\}$ is a family of random variables where $X(t)$ takes values in (i.e. has as range) a set $S \subset \mathbb{R}^d$ called its *state space*. The state space can be discrete (e.g. integers) or continuous. The set T is usually interpreted as the time. In this setting we consider only stochastic process with discrete state space and continuous time. A Markov Process is a stochastic process in which the probabilistic future behaviour, the 'evolution' of the system, depends only on the current state. In other words, stochastically, the past history of the process does not influence its future behaviour.

Definition 2.1 *Consider the state space* $S \stackrel{df}{=} \{s_i : i \in \mathbb{N}\}$. *The family of random variables* $\{X(t) : t \in \mathbb{N}\}$ *is a Discrete Time Markov Chain (*DTMC*) with state space* S *if :*

$$\mathbb{P}(X(t) = s_t | X(1) = s_1 \ldots, X(t-1) = s_{t-1}) = \mathbb{P}(X(t) = s_t | X(t-1) = s_{t-1}) \quad (1)$$

where $s_t \in S$ *is the state the chain at time* t.

Property (1) is the *Markov property* or *memoryless* property. The Markov property says that the probability of the chain to be in a given state at time step t depends only on the previous time step $t-1$. It does not matter what happened before the time step $t-1$. The chain is called *time-homogenous* if:

$$\mathbb{P}(X(m+1) = s' \mid X(m) = s) = \mathbb{P}(X(t+1) = s' \mid X(t) = s) = p_{s,s'} \quad (2)$$

for all $t, m \in \mathbb{N}$.

In other words, to find out the probability of moving from state s to state s' after m number of time units is independent form the time the chain was in state s. In what follows we consider only time-homogenous DTMC. Any time homogenous DTMC with $|S| = n$ number of states there exists a unique square matrix, $\mathbb{P}$ $(n \times n)$ called the *transition probability matrix*

$$\mathbb{P} = \begin{pmatrix} p_{s_1 s_1} & p_{s_1 s_2} & \cdots & p_{s_1 s_n} \\ p_{s_2 s_1} & p_{s_2 s_2} & \cdots & p_{s_2, s_n} \\ \vdots & \vdots & \vdots & \vdots \\ p_{s_n s_1} & p_{s_n s_2} & \cdots & p_{s_n s_n} \end{pmatrix}$$

such that $\mathbb{P}$ is stochastic i.e. $\sum_{j=1}^{n} p_{s_i s_j} = 1$, for all $1 \leq i \leq n$.

The Chapman-Kolgomorov equation, allow us to compute the probability of a chain moving from state to state in a finite number of steps.

For simplicity we write $p^{(m)}_{s_j, s_m}$ for $=\mathbb{P}(X(m) = s_i \mid X(0) = s_j)$.

Definition 2.2 (Chapman-Kolmogorov)

$$\mathbb{P}(X(m+r) = s_i \mid X(0) = s_j) \quad = \quad \sum_k p^{(m)}_{s_k s_i} p^{(r)}_{s_j s_k}. \tag{3}$$

This can also be rewritten as $\mathbb{P}^{(m+r)} = \mathbb{P}^{(m)}\mathbb{P}^{(r)}$.

We now define the probability of a chain to be in a state s_i a time step m as $\pi^{(m)}_i = \mathbb{P}(X(m) = s_i)$.

By the law of total probability we have that:

$$\begin{aligned} \pi^{(m)}_i &= \sum_k \mathbb{P}(X(m) = s_i \mid X(0) = s_k)\pi_k(0) \\ &= \sum_k p^{(m)}_{s_i,s_k}\pi_k(0) \end{aligned} \tag{4}$$

Let's now consider the row vector $\pi(m) = (\pi_1(m), \pi_2(m), \ldots \pi_n(m))$. We can compute such vector as follows

$$\pi^{(m)}\mathbb{P} = \pi^{(0)}\mathbb{P}^{(m)}$$

where $\pi^{(0)}$ such that $\sum_k \pi^{(0)}_k = 1$ is the initial vector distribution.

The vector $\pi^{(m)}$ is called the *transient distribution*. The transient distribution represent the probability that we find the chain in a certain state a specific time.

We write simply π when we refer to a distribution that it is independent of the time or the initial state. Such a distribution is defined as

$$\pi = \lim_{m\to\infty} \pi^{(0)}\mathbb{P}^{(m)}$$

when such a limit exists. π is called the *steady state distribution* and we stress out that each element of the vector has to be strictly positive. The steady state distribution, represent the probability that we find the chain in a certain state independently of the time. It means that the chain has reach an equilibrium, which was not reached in the the transient phase.

A DTMC has a steady state distribution only if enjoy specific properties. For the purpose of this work we shall say that if the state space of the chain is finite and irreducible, then the DTMC has a steady state distribution. The state space of a DTMC is irreducible if every state is reachable from every other state, there exists an m for which $p^{(m)}_{s_i,s_j} > 0$ for every pair of states s_i, s_j.

We can now proceed to the definition of Hidden Markov Model (HMM) [17, 15, 6]. Hidden Markov Models can be used to establish whether a

given DTMC is a suitable probabilistic model that explains a number of observations directly correlated to the behaviour of the chain.

Therefore, together with a DTMC the HMM introduces another probabilistic relation between the states of the DTMC and some possible observations. The basic idea is that given a finite set of observations, the HMM allows to infer the best model DTMC that fits the set of observations.

We assume the existence of a countable set of symbols $\mathsf{V} = \{v_i : i \in \mathbb{N}\}$ and of a family of random variables $\{V(t) : t \in \mathbb{N}\}$ associated to the observations.

Definition 2.3 *A Hidden Markov Model (*HMM*) is a tuple* $\mathbf{M} \stackrel{df}{=} \langle \mathbb{P}, \mathbb{Q}, \pi^0 \rangle$ *where:*

1. $\mathbb{P}$ *is the transition matrix of the underlying* DTMC.
2. $\mathbb{Q}$ *is the probability matrix of the observation symbols such that the entry of such matrix are defined as*

$$q_{s_j \to v_k} = \mathbb{P}(V(t) = v_k | X(t) = s_j)$$

3. $\pi(0)$ *is the initial distribution associated to the transition matrix.*

The definition above allows us to define the *observation sequence* $\mathcal{O} \stackrel{df}{=} \{o_i : i \in \mathbb{N}\}$ where the variable o ranges over the set of symbols V. For convenience we define a *finite observation sequence* of length n as

$$\mathcal{O}_n = o_0 o_1 o_2 \ldots o_{n-1}.$$

Now with the definition of HMM there are a number of calculations that we can perform

The other computation we can perform is given a sequence $\mathcal{O}$ and a model $\mathbf{M}$ what is the probability of $\mathbb{P}(\mathcal{O}_k|\mathbf{M})$?

As we assume that all observations are independent then we can write:

$$\mathbb{P}(\mathcal{O}|\mathbf{M}) \quad = \quad \mathbb{P}(o_1|\mathbf{M})\mathbb{P}(o_2|\mathbf{M})\ldots$$

where

$$\mathbb{P}(o_j|\mathbf{M}) \quad = \quad \mathbb{P}(o_j|\langle \mathbb{P}, \mathbb{Q}, \pi^0 \rangle)$$

for all j. Now under the assumption that there we know a finite sequence of states, $\mathsf{Seq} = w_0 w_1, w_2, w_3, \ldots, w_T$ of one realisation of the DTMC we define

$$\mathbb{P}(\mathsf{Seq}|\langle \mathbb{P}, \mathbb{Q}, \pi^0 \rangle) = \pi^0_{w_0} p_{w_0 w_1} p_{w_1 w_2} \ldots p_{w_{T-1} w_T}$$

where w_i ranges over S. We consider a finite sequence of observations, $\mathcal{O}_{T+1}$, of length $T+1$, generated by the Seq and we have that

$$\mathbb{P}(\mathcal{O}_{T+1}|\mathsf{Seq}, \langle \mathbb{P}, \mathbb{Q}, \pi^0 \rangle) = q_{o_0 \to o_1} q_{o_1 \to o_2} \cdots q_{o_{T-1} \to o_T}$$

Thus we can see (using the a variation of the law of total probability $\mathbb{P}(A|C) = \sum_B \mathbb{P}(A|B,C)\mathbb{P}(B|C)$) that

$$\begin{aligned} \mathbb{P}(\mathcal{O}_{T+1}|\langle \mathbb{P}, \mathbb{Q}, \pi^0 \rangle) &= \sum_{\forall \mathsf{Seq}} \mathbb{P}(\mathcal{O}_{T+1}|\mathsf{Seq}, \langle \mathbb{P}, \mathbb{Q}, \pi^0 \rangle)\mathbb{P}(\mathsf{Seq}|\langle \mathbb{P}, \mathbb{Q}, \pi^0 \rangle) \\ &= \sum_{w_0, w_1, w_2, \ldots w_T} \pi^0_{w_0} p_{w_0 w_1} q_{o_0 \to o_1} \cdots p_{w_{T-1} w_T} q_{o_{T-1} \to o_T} \end{aligned}$$

To compute equation 5, for an finite or infinite sequence of states, we can use an inductive procedure on the observations.

Definition 2.4 *Let* $\mathbf{M} = \langle \mathbb{P}, \mathbb{Q}, \pi^0 \rangle$ *and* HMM *and assume that the there is an infinite sequence of states* $\mathsf{Seq} = w_0 w_1, w_2, w_3, \ldots, w_n, \ldots$ *that generate the observation* $\mathcal{O}$*. Computation of* $\mathbb{P}(\mathcal{O}|\langle \mathbb{P}, \mathbb{Q}, \pi^0 \rangle)$ *is inductively defined as follows:*

$$\begin{aligned} \mathbb{P}(o_0|\langle \mathbb{P}, \mathbb{Q}, \pi^0 \rangle) &= q_{w_0 \to o_0} \pi^{(0)}_{w_0} \\ \mathbb{P}(o_0, o_1, \ldots, o_{n+1}|\langle \mathbb{P}, \mathbb{Q}, \pi^0 \rangle) &= \\ & \sum_{w_n \in \boldsymbol{S}} \mathbb{P}(o_0, o_1, \ldots, o_n|\langle \mathbb{P}, \mathbb{Q}, \pi^0 \rangle) p_{w_n w_{n+1}} q_{w_{n+1} \to o_{n+1}} \end{aligned}$$

Of course, if our sequence of observations is finite, we consider $\mathcal{O} = \mathcal{O}_{T+1}$ then the final step gives us that

$$\mathbb{P}(\mathcal{O}_{T+1}|\langle \mathbb{P}, \mathbb{Q}, \pi^0 \rangle) = \sum_{w_{T+1} \in \boldsymbol{S}} \mathbb{P}(o_0, o_1, \ldots, o_T|\langle \mathbb{P}, \mathbb{Q}, \pi^0 \rangle)$$

2.3 *Buzz* HMM

We build a HMM to model the temporal evolution of influence. As this is temporal, the buzz created by a user will be modelled as a DTMC. We have chosen four hidden states to represent influence or *Buzz* in HMM, and the DTMC is outlined in figure 1.

The observations used, at each time point, to determine the hidden buzz state are as follows:

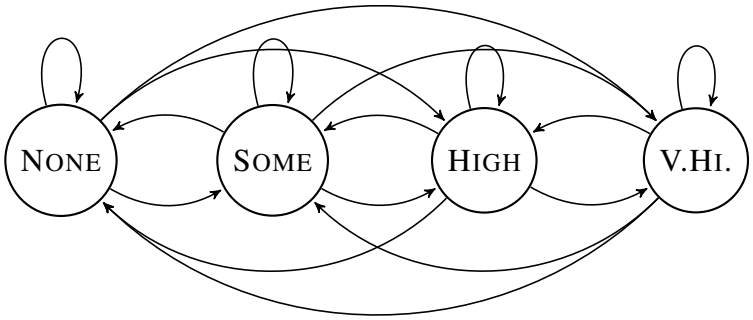

Figure 1: Buzz HMM Latent States

- Content: *retweet* count, v_1
- Identity: mention count, v_2
- Engagement: unique reply count, v_3

Each of these observations is a feature of the data, and they have been extracted can be found in the Section 3.1. The *retweet* count is a measure of the content diffusion ability of a user. It has been shown that more interesting content tends to produce longer cascades [5], hence users with high *retweets* are generally producing interesting content. The *retweet* is a powerful mechanism for reaching a large audience easily on *Twitter* , *retweeted* tweets have been shown to reach an average of 1,000 users regardless of the follower count of the tweet source. This unique property makes a *retweet* the most interesting of all the influence relevant metrics.

On *Twitter* a reply is a response to another user's tweet. A mention is addressed to another user but not in reply to a tweet. Our definition differs from the *Twitter* definition, in that we do not consider a mention to be a reply. Being mentioned represents name value of a user, they generally contain less linked content (URLs) and the most mentioned users have been shown to be celebrities [7].

To measure the engagement level of a user we use the number of unique replies recorded at each time point. Unlike the other two metrics a user can directly affect this metric by initiating more conversations, hence as we see in sec. **??**. *Retweet* and mention ranks have been shown to exhibit a power-law distribution with the majority of "ordinary users" commanding very little or no influence. Hence we expect the NONE state to be the most

common and that that other three will be in the long tail of the distribution. These three states seek to identify users creating a measurable effect on the network by displaying any or all of the following influence class behaviours: content value, name value or engagement.

For a discrete time Hidden Markov Model their is only one emission at each time point. We have chosen to combine the multivariate data streams $(v_1^{\{1:T\}}, v_2^{\{1:T\}}, v_3^{\{1:T\}})$ into a single observation, an alternative would have been to have used an HMM for each observation stream, however the cost of training and analysing three separate models would not be offset by a gain in expressiveness.

The observables, $V = v_1, v_2, v_3$, are combined into a single variable by combining the different observations stream into intervals and then combining these intervals into a single categorical variable.

We have chosen to represent the ranges by four symbols for *retweets*, $R = \{None, Some, High, VeryHigh\}$, two symbols each for mentions and unique interactions $M \wedge U = \{Low, High\}$. As *retweets* are considered the most important observable we wanted additional ability to differentiate between levels of *retweet* . We then categorise observations from each of the streams by assigning values into an interval for each symbol, the total symbol set $S = R \times M \times U$ as shown in table 2.

Each interval variable is given it's ranges by examining the top quantiles for each influence metric and choosing a quantity which differentiates users accordingly. The ranges, which have assigned from observed data are given in table 1.

Retweet	Mention	Uniq. Int.
$N = 0$	$L < 6$	$L < 30$
$0 < S < 6$	$H \geq 6$	$H \geq 30$
$6 \leq H < 15$		
$V \geq 15$		

Table 1: Observation Symbol Ranges

The procedure for generating these observations for model consumption is as follows:

1. Calculate quantiles for each influence metric.
2. For each observation assign to influence interval.

Value	*Retweet*	Mention	Uniq. Int.
1	N	L	L
2	N	L	H
3	N	H	L
4	N	H	H
5	S	L	L
6	S	L	H
7	S	H	L
8	S	H	H
9	H	L	L
10	H	L	H
11	H	H	L
12	H	H	H
13	V	L	L
14	V	L	H
15	V	H	L
16	V	H	H

Table 2: Observation Symbol Key

When completed each user will have their own observation set $V^{\{1:t\}}$ for all time periods in sample.

The initial transition matrix is produced from a Dirichlet prior, $\mathbb{P} \sim \text{Dir}(\alpha_1, \ldots \alpha_k), k = 4$. The emission matrix though is more complicated because of the amount of observation symbols and the many possible different assignments to each state.

The approach we have taken was to first capture the user observation distributions. We then convert a user based emission matrix into an observation based emission matrix. The first step is to find the number of different user distributions in the data. This can be performed by analysing the graph of means for each user per influence metric (*retweet*, mention and unique reply).

Once determined we sample the the users by splitting each influence observation into n sets. The mean is used to order user's observations for each influence metric. The top user quantiles $(q_1, \ldots, q_k)$ are selected for each metric and added to the set for that quantile.

$$Q^{q_j} = R^{q_j} \cup M^{q_j} \cup U^{q_1}, j = \{1, \ldots, k\} \tag{5}$$

There is potential for overlap between the *influencer* sets. The top *influencer* set is left as is but the sets below it must be the complement of the one before it:

$$I^{q_j} = \begin{cases} Q^{q_j}, & \text{if} j \ = 1, \\ Q^{q_j} \setminus I^{q}_{j-1}, & \text{if } j > 1. \end{cases} \tag{6}$$

Users in the top quantile set have the highest *retweets* , mention and unique interactions in the data set, users in the bottom have the lowest. We flatten the different users sets of observations out into a single observation vector.

$$\mathbf{o} = obs_1 \cup obs_2 \cup \ldots obs_k \tag{7}$$

One we have the vector for each level of user we can create a frequency vector for each observation symbol:

$$\mathbf{f}_k = |s \subseteq \mathbf{o}|| \, / \, |\mathbf{o}| \tag{8}$$

The user emission matrix is composed, row-wise, of these individuals frequency vectors.

$$U_{m,n} = \begin{pmatrix} f^1_{11} & f^1_{12} & \cdots & f^1_{1n} \\ f^2_{11} & f^2_{12} & \cdots & f^2_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ f^m_{11} & f^m_{12} & \cdots & f^m_{1n} \end{pmatrix}$$

To create the buzz emission matrix we must first assign the 16 symbols to their most likely state. Table 3 displays this assignment.

Now we have four states $\{N, S, H, V\}$ and 16 possible assignments to those states we need to know the probability of the states emitting those symbols. To determine this we use the earlier user-based emission matrix. Firstly we find the relative frequency of each observation by summing the row values and normalising this vector to sum to 1:

$$US = \sum_{j=1}^{n} \sum_{i=1}^{m} U_{i,j} \tag{9}$$

$$UN = \sum_{j=1}^{n} \sum_{i=1}^{m} U / \, ||US|| \tag{10}$$

Value	State	Value	State
1	N	9	H
2	S	10	H
3	N	11	H
4	S	12	V
5	S	13	V
6	S	14	V
7	S	15	V
8	H	16	V

Table 3: Observation State Assignment

$$SB = UN_i, i = \{2, 4, \ldots, 7\} \tag{11}$$
$$HB = UN_j, j = \{8, \ldots, 11\} \tag{12}$$
$$VB = UN_k, k = \{12, \ldots, 16\} \tag{13}$$

Now we can construct the *Buzz* emission matrix. The first row of this corresponds to the first row of the user matrix. These are the ordinary users who have no influence. All others rows are derived from the normalised *Buzz* state rows.

$$\begin{gathered} BE_1 = UE_1 \\ BE_i = (SB/\,||SB||)^2 \\ BE_j = (HB/\,||HB||)^2 \\ BE_k = (VB/\,||VB||)^2 \end{gathered}$$

Once the initial model has been created form the data it is then trained using Baum-Welch parameter estimation.

2.3.1 Buzz Rank

The Viterbi path determines the highest likelihood for a sequences of hidden states given an observation set [16]. We apply a transform to the output of the Viterbi path that produces a *Buzz* score for each time point and also for the set thereof.

As well as determining which state a user is currently in we wish to associate a numeric value to that state so that we may produce a quantitative

score. In addition to quantifying the value of being in a state at a given time point, we also wish to rank users over multiple time points.

We have assigned a numeric value to each state, $\{S = \{s_1, \ldots, s_k\} : s \in \mathbb{N}, 1 \leq s \geq 3, k = 4\}$ and given each an importance, β which is calculated by the following function:

$$\beta = \begin{cases} 0, & \text{if } s = 0, \\ \exp(s), & \text{if } s > 0. \end{cases} \tag{14}$$

The *Buzz* value assigns an importance that is exponentially increasing for higher states. The NONE state has been set at zero so as to not provide any positive credit in calculating the *Buzz* score. With this formulation users who spend longer in the higher states will be rewarded accordingly with considerably more weight going to the higher states.

State	Value	Buzz
None	0	0
Some	1	2.72
High	2	7.39
V. High	3	20.09

The influence metrics (*retweets* and mentions) follow a negative binomial distribution, hence the reason choosing a similar continuous probability distribution. The *BuzzRank* is the summation of all *Buzz* scores across all time points.

$$\textit{BuzzRank} = \sum_{t=1}^{k} \beta_t \tag{15}$$

2.3.2 InfluenceRank: Combining Reach and Buzz

The combination of the network structure based *Reach* (implemented as PageRank here) and the observation capturing *Buzz* provides the quantification of Influence in this research. The *Buzz* is recorded from observables and related via the HMMto an ordinal representation of the latent buzz state. Should this model have enough information it could conceivably measure all of a user's influence, however there are many influence related observations we cannot measure, examples include:

- Click-through on URLs.

- User reading post and taking subsequent action.
- User reading post and it informing their decision/perspective.

Theses actions highlight why our observables alone are not sufficient to model the influence problem. The role of user in the a network structure can give considerable insight into their effect on the network where observables may be missing or as complementary evidence in all cases.

In seeking a unified model of influence we have combined, the capacity for an individual to affect the network that is the *Reach* and the demonstrable effect recorded from direct observation of network interaction, the *Buzz*.

$$\textit{InfluenceRank} = \sum_{t=1}^{k} \rho_t \times \beta_t \tag{16}$$

3 Case study: *Twitter*

3.1 Data Collection

To evaluate influence on *Twitter* we sampled a specific sub-network centred around the Scala programming language [4]. Being limited to 150 requests per hour using the *Twitter* API we wanted a network with a good level of interaction for the size of the sample. We choose the Scala creator Martin Odersky as the seed node for the network and then crawled the profiles of his followers (11647 as of 16-Nov-2012). The aim was produce a network with higher modularity than a random sample of the same size. This sampling strategy is effective for getting a base level of interaction with a very restricted API limit however it introduces some selection bias, principally all sampled users are to some extent influenced by the seed node (as they all follow him) hence he is ignored from our analysis.

Our sampling started on the 16th of November 2012 and ended on the 8th of December 2012. Similarly to [7] we ignored inactive users, users who had less than 5 tweets in their *timeline* (*tweet* history). Due to an artificial limit on the amount of *tweets* imposed by *Twitter* at most 3200 *tweets* are returned for any user. For our experiment we chose the interval [15-Jan-2012,16-Nov-2012] as this was the 85% percentile of all user start times and gave the best trade-off of tweet volume and sufficient experimental duration. We had to ignore any users who had tweet counts greater than 3200 and their first tweet time after 15-Jan-2012 which was 429 users leaving 8547 users remaining as the sample network. In-line with other research on the *Twitter* network [7], the degree distribution of the sample is highly skewed. The

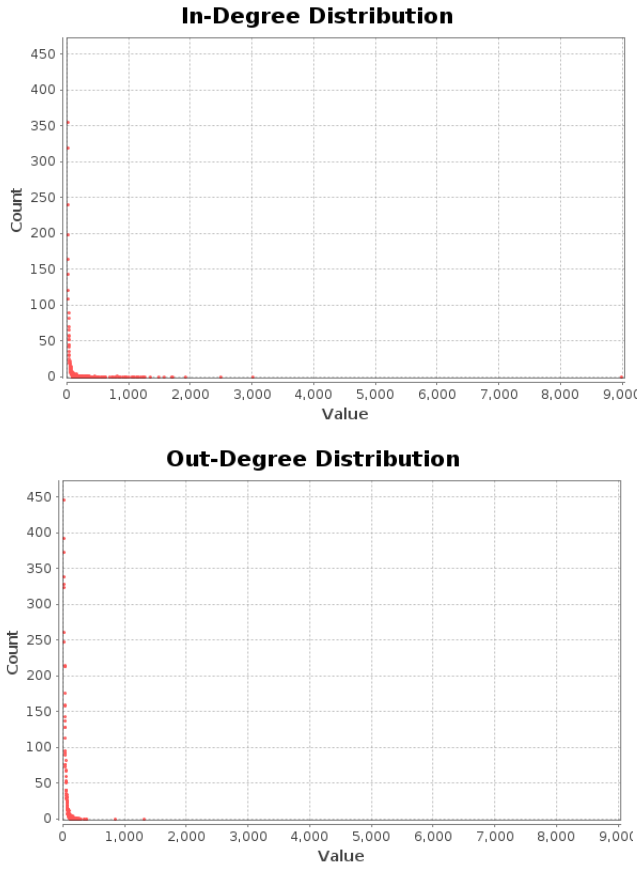

Figure 2: Degree Distributions

notable outlier for in-degree distribution in figure 2, with an in-degree of nearly 9,000 users is the seed node Martin Odersky, this is to be expected as the sample was constructed from users who follow his updates. For clarity we confirm that the graph contains, users as nodes and connections between them as edges. An in-link means that some user follows you and an out-link signifies that you follow that user. Hence the in-degree captures those users who follow a node and out-degree is the users a node follows.

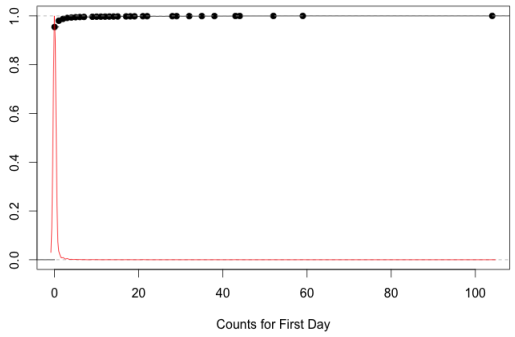

Figure 3: *Retweet* counts exhibit a ZINB distribution

3.2 Experimental Setup

Our experimental evaluation of influence in the Scala centered sub-network of *Twitter* lasted from [15-Jan-2012,16-Nov-2012]. This consists of aggregated counts for 43 weeks of observations. The observations were *retweets* , mentions and unique replies of the users. *Retweets*, mentions or replies of users not present in the dataset were ignored. Users who failed the criteria outlined in sec. 3.1 were also excluded. To properly train the HMM, it was necessary to ensure that all observation sequences were of the same length, hence all users have 43 observations. This has a considerable effect on the skew of the data and produces a Zero-Inflated Negative Binomial Distribution (as shown in figure 3) whereby the data is heavily skewed towards zero value observations.

3.3 Analysis

?? From our sample of 8547 users over 43 observation weeks with aggregated: *retweet*, mention and unique replies counts we have extracted the following statistics.

3.3.1 Users

To train a HMMusing Baum-Welch multiple sequence training it is necessary for those sequences to be of the same length. Most of the users sampled had very few tweets, as can be seen by the tweet distribution (fig. 4). This meant that most sequences contained high levels of inactivity. This miss-

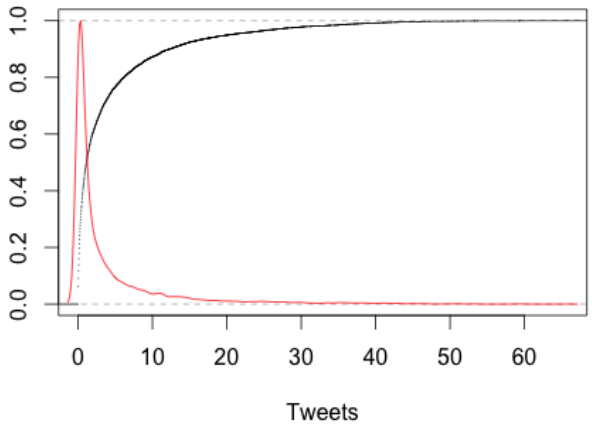

Figure 4: User *Tweet* Distribution and ECDF.

ing data was padded out by representing inactivity the same as having zero *retweets*. These may not be exactly similar but they have a similarly negligible effect on the buzz created by a user at a time point.

Figure 5 displays the highly skewed influence metric distributions. In this figure we see the distribution of mean value for users across all time points in the observation interval. A *retweet* or mention mean greater than 1 is very rare as is a unique reply mean greater than ten. *Retweets* and mentions follow a similar zero-inflated distribution whilst the unique replies have more users with positive means.

3.3.2 Activity

Understanding the activity of users is important in modeling influence as any seasonality or trend in activity levels can reduce the accuracy of the ranges selected by the process outlined in section 2.3. Figure 6 delineates the mean activity levels for each time period (week) of the experiment whilst also showing the mean influence metric values for those weeks. It is clear that there is no trend in the activity levels. If there is seasonality present (the Summer months have slightly lower mean values) it is not very significant.

Originally we had captured data at the day granularity but found this to be too volatile. The day data had two different levels of activity: Saturday, Sunday and Monday were the least active and Tuesday, Wednesday, Thurs-

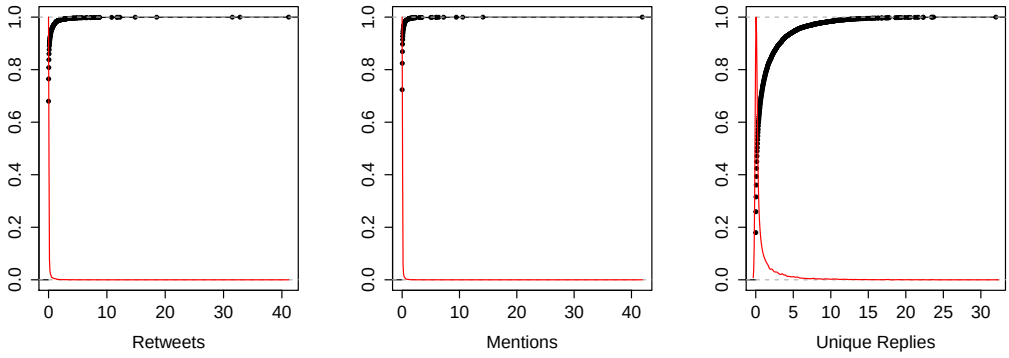

Figure 5: Red line displays the density estimate for the **user mean influence values**. The black line is the Empirical CDF.

day and Friday displayed the most activity. Also by aggregating up to the week level, the percentage activity increased significantly as is highlighted in figure 7.

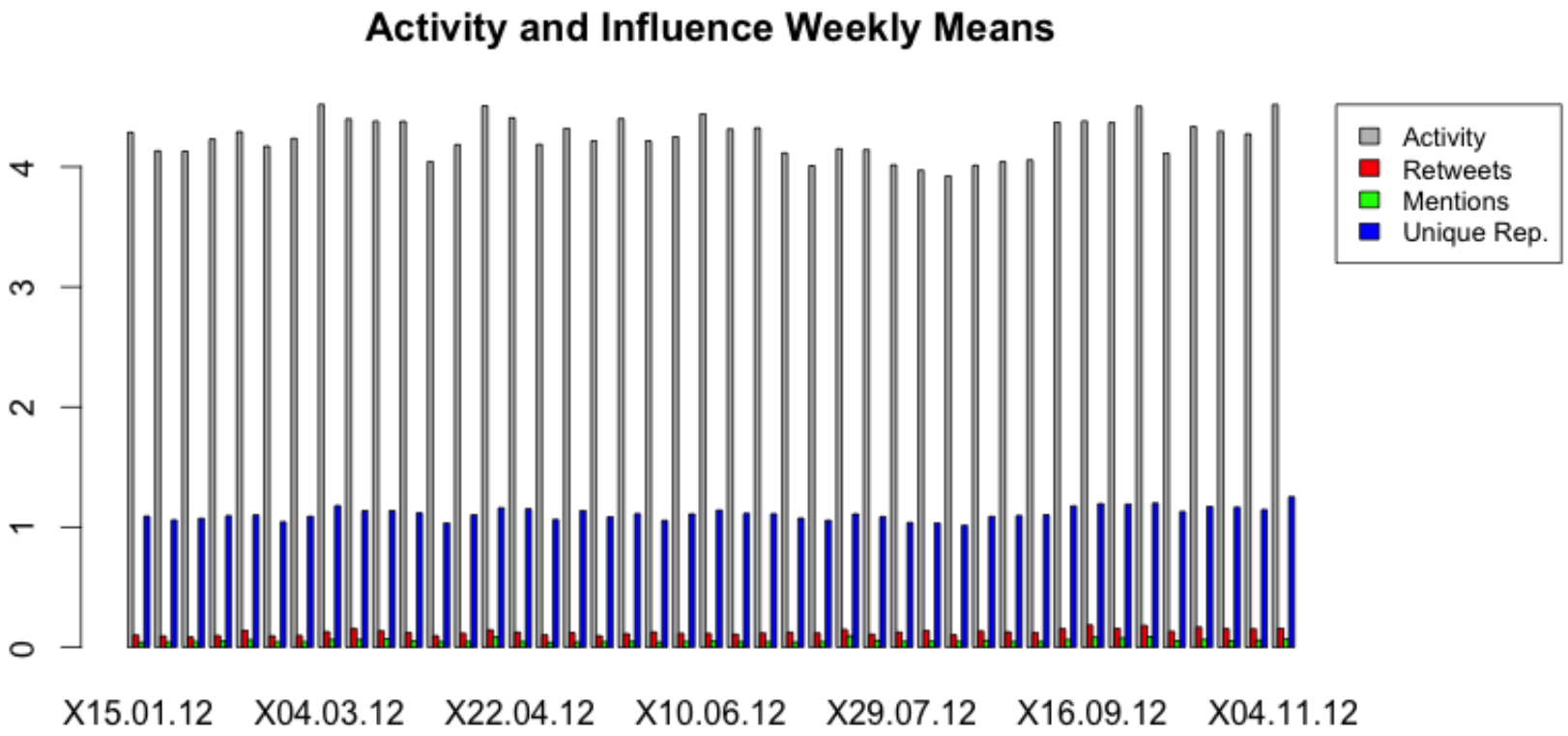
Activity and Influence Weekly Means
Activity
Retweets
Mentions
Unique Rep.
0
1
2
3
4
X15.01.12
X04.03.12
X22.04.12
X10.06.12
X29.07.12
X16.09.12
X04.11.12

Figure 6

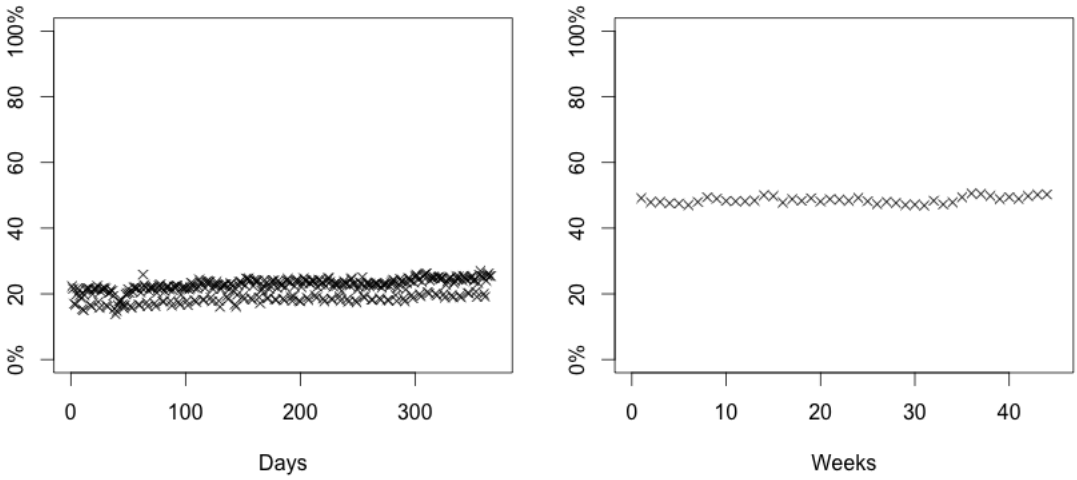

Figure 7: Comparison of Day/Week *Tweet* Activity %

3.3.3 Temporal Data

The weekly aggregated observations means are displayed in figure 9. *Retweets* and unique replies are normally distributed but mentions has a slight skew. From our initial observation of the data we hypothesise that the spikes or outliers in mention means are due to exceptional events that effect the Scala network, such as the launch of the Functional Programming Principles course on Coursera [4] in September 2012. Further analysis of tweet text would be required to investigate that hypothesis fully.

The most important observational aspect of the data from a weekly perspective are the quantile values for the observation at each time point. These form the basis for the intervals described in 2.3. By calculating the quantile of interest we can see what values are truly rare and which observations should be of interest to us. Figure 8 displays the quantiles for each influence metric and also highlights just how skewed the data is with 95% of users not registering any influence relevant observables for most weeks.

[4] https://www.coursera.org/course/progfun

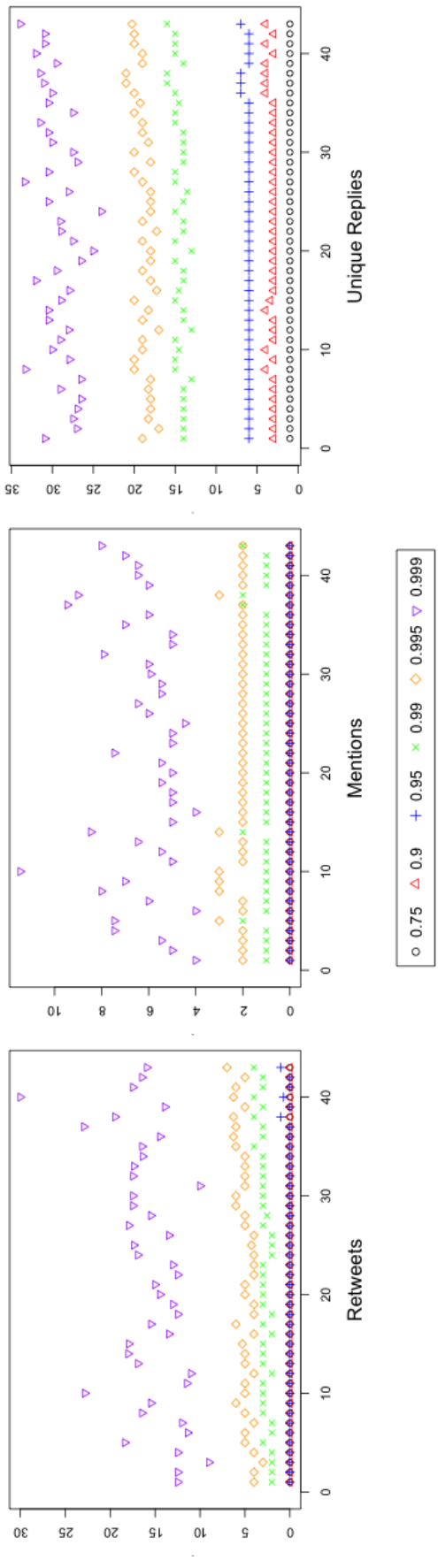

Figure 8: Top Quantiles for Influence Metrics

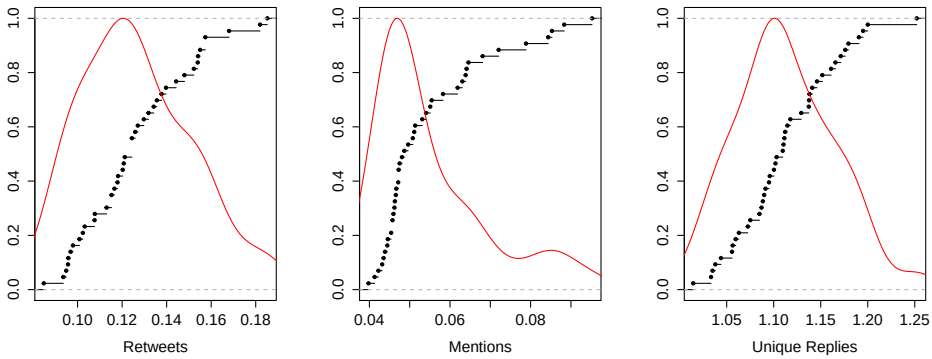

Figure 9: Red line displays the density estimate for the **weekly mean influence values**. The black line is the Empirical CDF.

3.4 P-Value Comparison

The p-value, borrowing the term from [7] (not statistical significance), is the probability of a particular *tweet* being a *retweet* , mention or unique reply of given user.

$U(X)_t$ = count of user influence metric at time t, X = influence metric
$T(X)_t$ = total count of influence metric at time t, X = influence metric

$$\mathbb{P}(R)_t = U(R)_t / T(R)_t$$
$$\mathbb{P}(M)_t = U(M)_t / T(M)_t$$
$$\mathbb{P}(U)_t = U(U)_t / T(U)_t$$

$$p_t = \frac{\mathbb{P}(R)_t + \mathbb{P}(M)_t + \mathbb{P}(U)_t}{3} \tag{17}$$

We compare the p-value rankings for each time point against the temporal influence rankings. Figure 10 displays the rank correlation for the top 1% of users for both our influence value and the p-value calculation. It is apparent that there is little to no correlation between the ranks produced by each method. It is noted that this calculation of the reference probability is simplistic and gives uniform weights to each influence metric. We have yet to determine the reason for the independence of these measures but it warrants further investigation.

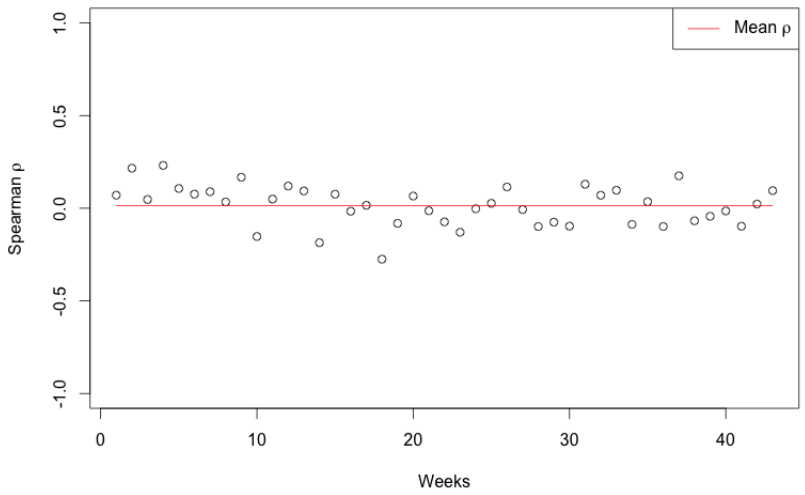

Figure 10: Influence vs P-Value Rank Correlations

3.5 Correlations

Table 4 details the correlations between the influence relevant metrics recorded by this research.

The influence score is obviously correlated with its factors: reach and buzz. It also shows strong correlation with in-degree (0.83 much stronger than the buzz only in-degree correlation of 0.56). It is not correlated to the activity of a user, in fact even the temporal buzz shows very little correlation with activity. This is not entirely surprising as more *tweets* $\neq$ more *retweets* or mentions. However active users do tend to have higher unique replies and this has some effect in increasing the correlation of buzz and activity.

Interestingly mentions are not correlated with any of the other metrics, although closest to in-degree, this highlights that they are a different kind of influence to *retweets* and that they are most closely related to popularity.

4 Conclusion

We have presented a novel model for capturing influence in Social Networks using data sampled from the *Twitter* network. The approach of combining network based topological methods with direct observed data is new and has

potential to overcome the individual limitations of either model.

Using a HMM to model the buzz created by a user in a network allows us to model this quantity temporally and allows for prediction of future buzz states.

One of the key challenges in this area is the lack of validation, it is not currently possible to validate a model as there is no gold standard training set of temporal influence.

References

[1] http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank/.

[2] http://www.dmoz.org/.

[3] http://www.infochimps.com/datasets/twitter-census-trst-rank.

[4] http://www.scala-lang.org/.

[5] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 65–74, New York, NY, USA, 2011. ACM.

[6] Jeff Bilmes. What HMMs can do. *IEICE Transactions in Information and Systems*, 3:869–891, 2006.

[7] Meeyoung Cha, Hamed Haddadi, Fabrcio Benevenuto, and Krishna P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *in ICWSM 10: Proceedings of international AAAI Conference on Weblogs and Social*, 2010.

[8] P.G. Harrison and N.M. Patel. *Perfomance Modelling and Communication Networks and Computer Architectures*. Addison-Wesley, 1992.

[9] Taher H. Haveliwala. Topic-sensitive pagerank. In *WWW*, pages 517–526, 2002.

[10] P.G. Hoel, S.C. Port, and c.J.Stone. *Introduction to Stochastic Processes*. Houghton Mifflin Company, 1972.

[11] John G. Kemeny and J. Laurie Snell. *Finite Markov Chains*, chapter II. D. Van Nostrand Company, inc., 1960.

[12] Valdis E Krebs. Uncloaking terrorist networks. *First Monday*, 7(4-1), 2002.

[13] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[14] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[15] Lawrence R. Rabiner. Tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 7:257286, 1989.

[16] Lawrence R. Rabiner. Readings in speech recognition. chapter A tutorial on hidden Markov models and selected applications in speech recognition, pages 267–296. 1990.

[17] Ramon van Handel. *Hidden Markov Models*. Lecture Notes, Princeton University, 1 edition, 2008.

[18] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 261–270, New York, NY, USA, 2010. ACM.

[19] Rongjing Xiang, Jennifer Neville, and Monica Rogati. Modeling relationship strength in online social networks. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 981–990, New York, NY, USA, 2010. ACM.

	Influence	Reach	Buzz	In-Deg	Activity	Active.Weeks	*Retweets*	Mentions	Uniq. Rep.
Influence	1.00	0.81	0.68	0.83	0.11	0.07	0.82	0.40	0.21
Reach	0.81	1.00	0.61	0.82	0.10	0.08	0.64	0.34	0.21
Buzz	0.68	0.61	1.00	0.56	0.36	0.23	0.85	0.41	0.49
In-Deg	0.83	0.82	0.56	1.00	0.10	0.10	0.73	0.49	0.19
Activity	0.11	0.10	0.36	0.10	1.00	0.64	0.19	0.14	0.83
Active.Weeks	0.07	0.08	0.23	0.10	0.64	1.00	0.15	0.10	0.55
Retweets	0.82	0.64	0.85	0.73	0.19	0.15	1.00	0.44	0.28
Mentions	0.40	0.34	0.41	0.49	0.14	0.10	0.44	1.00	0.16
Uniq. Rep.	0.21	0.21	0.49	0.19	0.83	0.55	0.28	0.16	1.00

Table 4: Influence Relevant Correlations