A comparative study of two models for celebrity identification on Twitter

MS Srinivasan *
IBM India, Embassy Golf Links
Koramangala Intermediate
Ring Road, Bangalore
smuthusw@in.ibm.com

Srinath Srinivasa
IIIT-Bangalore,26/C,
Electronics City, Hosur Road,
Bangalore
sri@iiitb.ac.in

Sunil Thulasidasan
Los Alamos National
Laboratory
NM, USA
sunil@lanl.gov

ABSTRACT

The concept of *celebrities* has shaped societies throughout history. This work addresses the problem of celebrity identification from social media interactions. "Celebritiness" is a characteristic assigned to persons that are initially based on specific achievements or lineage. However, celebritiness often transcends achievements and gets attached to the person itself, causing them to capture popular imagination and create a public image that is bigger than life. The celebrity identification problem is argued to be distinct from similar problems of identifying influencers or of identification of experts. We develop two models for celebrity identification. In this paper, we compare the two models on twitter data and highlight the characteristics of each of the models.
Keywords: Celebrity, Social media, Twitter, Influence

1. INTRODUCTION

All societies have celebrity figures that are admired, respected or idolized. Celebrities are well-known personalities, and their actions attract a lot of attention. They are therefore a subject of interest for marketing teams, policy makers, social workers, preachers, teachers, etc. Understanding the dynamics of celebrity formation and identification of potential celebrities is an important problem of interest.

Web based social media adds an extra dimension to celebrity dynamics. Usually, celebrities in the real world are also well-known [27] and admired on social media, and are also employed to promote causes and interests. However, the participatory nature of social media often breeds its own celebrities. There are several cases of personalities who go on to become well-known in the outside world due to their celebrity status on social media. Similarly, several well-known people are not savvy enough on social media to elicit the same level of adulation as they receive in the real world.

A celebrity is not the same as an influencer. Influence of a person is typically specific to a topic [11] and the person's position in the social network. Often, lesser known persons may end up wielding enormous influence on a topic because of their network centrality.

Celebrities, on the other hand, are well-known personalities. They are usually idolized by a subset of the population, and their actions receive more than average attention from the population. While, celebrities are typically influencers, not all influencers are celebrities. Also, when influence is defined as *persuasive power*, celebrities typically do not fare better than influencers in our immediate social network. People are far more likely to be persuaded by a close friend or a family member, than by a celebrity.

"Celebritiness" is characterized more by the *attention* and *adulation* that it elicits, and not by its persuasive or influencing abilities. Thus, celebrities are of interest more as "fronts" or the face of a particular product or idea, that primarily aims to catch attention or create awareness about the product or the idea[1].

While influence is a characteristic associated with a topic or event, "celebritiness" is a characteristic associated with a *person*. A characteristic definition of a celebrity, attributed to Boorstin [4], says: *A celebrity is the person who is well-known for their well-knownness*. Once a person acquires a celebrity status, they become the object of attention themselves.

Also, while celebrities are well-known personalities, not all well-known persons are celebrities. People could also be well-known for notorious reasons. Celebritiness is also not about fame for a specific reason. People usually become famous based on their specific achievements in some domain (winning the Wimbledon) or because of their lineage (son/daughter/spouse of a celebrity). But a celebrity is one, whose fame *transcends* the specific reasons why they became famous in the first place, and becomes associated with the person itself, giving them a larger than life persona in popular imagination.

With this perspective, we have found that celebrity identification per se, has not received much interest from researchers in social media. In this paper, we address the space of celebrity dynamics and explore two computational models for identifying celebrities from social media data,

*This author is a part-time research scholar at IIIT-Bangalore and this research work is carried out as part of his MS thesis

[1]http://mediakix.com/2013/10/fashion-panel-celebrities-vs-influencers-wins/. Last accessed: 08 Aug 2014.

The 20th International Conference on Management of Data (COMAD), 17th-19th Dec 2014 at Hyderabad, India.

based on Twitter. A preliminary form of the first model called Acquaintance-Affinity-Identification (AAI) was proposed by the authors in [26]. This model is based on the notion of source credibility and source attractiveness. In this paper, we propose another model called Action-Reaction (AR), that is based on loyalty and attention. Our AAI and AR models are derived from two different theories in social psychology that look at orthogonal factors towards celebritiness. In this paper, we also provide a comparative study on the two proposed models.

2. RELATED LITERATURE

2.1 Models of influence and expertise

A lot of research effort has gone into identifying "influencers" and computational models for "influence maximization." Motivating applications include viral marketing, sales prediction and counter terrorism. Some examples are [3, 28, 20, 1, 15, 7, 9, 8]. Though several models are proposed for influence maximization in networks, most of them are found to be intractable. In response, several heuristic approximations have been proposed [28, 3, 9, 8] based on degree, centrality-based measures, greedy heuristics and data mining approaches. Such approaches apply diffusion principles to mimic the "word of mouth" behavior in human social environments for spread of information.

Hajian, et al. [16] propose a formal model for measuring influence on *FriendFeed* social network[2]. This model computes "Magnitude of Influence" (MOI) for each user based on the number of hits generated for a user posting. The model then computes the "Influence Rank" using MOI. Further, Gosh, et al. [14] define influence as the number of in-network votes a user's post generates and applied it on *Digg* social network[3].

Forestier, et al. [12] propose a framework for extracting celebrities from the online discussions[4]. This framework uses three different meta-criteria. The first meta-criteria identifies the potential celebrities who have more than average number of in-degree, out-degree, posts compared to other people in the community. The second meta-criteria is based on the participation of the user in different forums. The third meta-criteria is based on the citations of names and the quoted texts.

Our problem of finding celebrities is slightly different from the influence maximization problem. Celebritiness is a characteristic attributed to people and have to do with who they are, more than their position in the network.

A related problem is of identification of "experts" in a community and recommending experts based on need [10, 30, 31]. While people with superlative expertise tend to become celebrities, expert identification is not the same as celebrity identification. Expertise is defined within some context or topic, while a celebrity figure may be context-free. Celebrities need not be experts and not all experts are likely to be celebrities.

2.2 Models of Celebrity

The concept of celebrities has been a topic of interest in social science and social psychology for several decades. A definition of celebrity by Mills [24] says:

> Celebrities are names that need no further identification. *Those who know them so far exceed those of whom they know as to require no exact computation.* Wherever they go, they are recognized, and moreover, *recognized with some excitement and awe.* Whatever they do has publicity value. More or less continuously, over a period of time, they are the material for the media of communication and entertainment.

The emphasized parts of the definition above provide vital inputs into understanding and building computational models for celebrity identification.

Another popular definition for a celebrity from Boorstin [4] that says: *A celebrity is the person who is well-known for their well-knownness*, suggesting a recursive or self-fulfilling nature of the phenomenon of celebrity formation.

In social psychology, research on the topic of celebrity endorsement rests on two general models:

1. The source-credibility model
2. The source-attractiveness model

The source-credibility model [18] defines celebritiness of a communicator to be a function of expertise and in turn, credibility. In this model, "expertness" is defined as the perceived ability of the celebrity to make valid assertions and "trustworthiness" is defined as the perceived willingness of the celebrity to make value assertions. Celebrities exhibiting expertness and trustworthiness are credible and to this extent, persuasive.

The source-attractiveness model [23], considered to be a component of an earlier model called the "source valence" model and draws on the research in social psychology [22], considers factors like: *familiarity*, *likeability*, and *similarity* between the source of communication and its recipient. In this model, "familiarity" is defined as knowledge of the celebrity through exposure. The second factor, "likeability" is defined as affection for celebrity as a result of celebrity's physical appearance and behavior. The last factor, "similarity" is defined as a supposed resemblance between the celebrity and consumers who are mostly celebrity fans. This model holds that celebrities who are known to, liked by, and/or similar to the consumer are attractive and, to this extent persuasive.

One of our models, called the AAI model [26] discussed in this paper is based on a combination of the source-credibility and the source-attractiveness model. This model is simplified to three factors and adapted to fit into Twitter communications.

Another model by Friedman and Friedman [13] determines celebritiness based on three other factors: *attention*, *recall* and *loyalty* from the population. The *attention* factor measures the amount of attention the celebrity gets from their fans. The *recall* factor defines the ability of the fan to re-collect the celebrity names. The *loyalty* factor measures the fans' loyalty towards the celebrity by providing support for the given celebrity.

The source-credibility and source-attractiveness models focus on *familiarity*, *likeability* and *similarity* between the celebrity and the fans. But the Friedman and Friedman

[2]http://friendfeed.com
[3]http://digg.com
[4]http://huffingtonpost.com

model [13] focus on *attention*, *loyalty* and *recall* factors. As the two models of social psychology look at orthogonal factors, we develop our second computational model called the AR model, inspired from the Friedman and Friedman model. Further, the properties of the celebrities identified using the two models are compared.

2.3 Influence Models on Twitter

Twitter being a popular social media in recent days, lots of research has been performed on its data sets to analyze influence. Cha, et al. [6] measure influence in Twitter using in-degree, replies and mentions. They find that the most followed users do not necessarily score highest on the other measures. PageRank-like scores have been proposed to measure influence on Twitter [29]. The *Social Network Potential (SNP)* score [2] is based on the average of two measures *Retweet and mention ratio* and *Interactor Ratio*. *Retweet and Mention ratio* is calculated as the amount of tweets that are amplified or lead to a communicative action between the communicator and another user divided by total amount of tweets of the communicator. The Interactor Ratio is measured as the ratio between the number of users who retweet or mention the communicator and the total amount of followers of the communicator.

Weng, et al. [29] propose *Twitter Rank*, an extension of the Page Rank algorithm, to measure influence by not only taking followers and interactions into account, but also by analyzing topical similarities with the help of a ranking method similar to PageRank. An interesting aspect of this work is that in the analyzed sample of Singapore-based users a high reciprocity (e.g.,mutual following relationship) was found.

Kwak, et al. [19] compared three different measures of influence: number of followers, Page Rank, and number of retweets, finding that the ranking of the most influential users differed depending on the measure.

Hatcher, et al. [17] develop an influence metric on twitter based on both "content" and "conversation" aspect. The "content" aspect is measured based on the number of tweets posted by the user and also based on the number of tweets posted by the members in the user's network. The "conversation" aspect is measured based on the number of replies, retweets and mentions received from the members of the user's network and also based on the number of replies, retweets and mentions received by the members of the user's network.

3. IDENTIFICATION OF CELEBRITIES

In our work, we use Twitter as the social media of interest, for celebrity identification. The participatory nature of online social media adds a new dimension to the celebrity problem. Usually celebrities of the outside world are also treated as celebrities in social media, attracting a lot of followers. So just looking at follower count may appear to be a good measure of celebritiness. However, this is not always the case [6].

There is a thriving ecosystem of "buying" followers[5][6][7] on Twitter, discrediting the follow score. In addition, celebrities from the outside world, often are not as active or elicit the same amount of attention for their activities, inside Twitter.

We hence develop and explore two computational models of celebrity scoring. The first, called the AAI model (for *acquaintance*, *affinity* and *identification*) is derived from the source-credibility and source-attractiveness models used in social psychology. We provide interpretations for acquaintance, affinity and identification on Twitter. The second model called the *Action-Reaction* (AR) model is a more direct measure of one's fame, based on the reactions they elicit.

Before introducing the specific models, we first formally describe the Twitter dataset on which the models were built.

3.1 Twitter data model

A Twitter dataset is formally modeled as follows:

$$\mathcal{D} = (T, N, \alpha, \mu, \rho, \tau, \gamma) \quad (1)$$

Here T is the set of tweets in the sample and N is the set of twitter accounts in the sample. The terms α, μ, ρ, τ and γ refer to ensembles of functions representing different elements of the Twitter universe. They are described in detail below.

α refers to an ensemble of functions around authorship. The function $author : T \rightarrow N$ maps a tweet to its author. The function $tweets : N \rightarrow 2^T$, gives the set of all tweets authored by a given twitter account. Given a set of tweets W, we will also be using a function $authors : 2^T \rightarrow 2^N$, defined as

$$authors(W) = \bigcup_{t \in W} author(t)$$

that gives the set of authors, given a set of tweets.

μ, ρ and τ refers to an ensemble of functions pertaining to mentions, replies and retweets respectively. The functions $mentionsof : N \rightarrow 2^T$, $repliesof : N \rightarrow 2^T$ and $retweetsof : N \rightarrow 2^T$ define the set of tweets that represent mentions, replies or retweets of a given twitter account.

Analogously, the functions *mentionsby*, *repliesby* and *retweetsby* are defined, which are all of the form $N \rightarrow 2^T$ and define the set of mentions, replies and retweets performed by an author.

γ refers to an ensemble of functions pertaining to followship. The function $follows : N \rightarrow 2^N$ depicts the set of accounts followed by a twitter account. Analogously, the function $followers : N \rightarrow 2^N$ gives the set of followers for a given account.

3.2 The AAI Model

The first of the two models, called the AAI model, is presented here. Based on the various studies in the psychology of celebrity [23, 18, 22] around source-credibility and source-attractiveness, we arrived at three attributes defining a celebrity:

1. Well-knownness
2. Likeability
3. Identification

A celebrity is someone who is well-known. But they are not just well-known, they are also *liked* by the population. Finally, idolization of celebrities happen because a

[5]http://pinchlikes.com/buy-twitter-followers#. Last accessed: 21 June 2013.

[6]http://www.wordstream.com/blog/ws/2013/05/16/buying-twitter-followers-pros-and-cons#. Last accessed: 21 June 2013.

[7]http://buytwitterpro.com/buy-twitter-followers#. Last accessed: 21 June 2013.

large number of people in the population *identify* with the celebrity in some form.

It can be seen that the three measures can be pipelined to form three hierarchical layers. We can only like someone whom we know, and we can only identify ourselves with someone whom we like. This gives us the 3-layer AAI model.

The proposed 3-layer model is called the AAI model and has three separate scores:

1. Acquaintance score (A)
2. Acquaintance-Affinity score (AA)
3. Acquaintance-Affinity-Identification score (AAI)

Acquaintance is a measure that determines how well a person is known in the community. Affinity measure determines how much the person is liked by the community and the Identification measure determines the extent to which others identify themselves with the person being studied. Their interpretations on Twitter are provided below.

3.2.1 Acquaintance Score

Acquaintance Score $A(i)$ for twitter account i is the measure of the number of people who knows the person i as a proportion of the population of the sample. In Twitter, acquaintance is established by *any* evidence that depicts knowledge of one account by another. This includes a follow or reply or retweet or mention.

The acquaintance score $A(i)$ for twitter account i thus given by:

$$A(i) = \frac{|followers(i) \cup authors(mentionsof(i) \cup repliesof(i) \cup retweetsof(i))|}{|N|} \tag{2}$$

3.2.2 Acquaintance-Affinity Score

Acquaintance-Affinity Score $AA(i)$ is a measure of how well a person is liked by the community and by whom. The affinity score is weighted by the acquaintance score so that being liked by well-known people increases the affinity content, as compared to being liked by lesser known persons.

Affinity is measured as a function of how much others respond to the activities of the person in question. It might be argued that reaction to one's action may also be due to animosity. While it may well be the case, it is unlikely that animosity will elicit sustained reactions over time. In addition, to "love to hate" someone can also be viewed as some form of affinity. Person j who "loves to hate" i is perhaps displaying envy which in turn is a form of affinity for what constitutes the characteristics of person i.

To measure affinity, we calculate three scores. Reply score (R), Mention Score (M) and Retweet Score (RT).

Reply Score $R(j|i)$ is defined as a conditional probability measure. Given that the person i replied to a tweet, the probability that the tweet was created by person j is represented as the Reply Score $R(j|i)$

$$R(j|i) = \frac{|repliesby(i) \cap repliesof(j)|}{|repliesby(i)|} \tag{3}$$

Mention Score $M(j|i)$ and Retweet Score $RT(j|i)$ are calculated in an analogous fashion.

There have been observations in the literature that retweets and replies are dependent on various factors like content influence, commonality in interests and network influence [25, 21]. However, since celebritiness is associated with people rather than with content or issues, we find these nuances irrelevant for our problem.

The Acquaintance-Affinity score AA(j) is then calculated as;

$$\begin{aligned} AA(j) = &\sum_{i \in E_r} A(i) * R(j|i) + \\ &\sum_{i \in E_m} A(i) * M(j|i) + \\ &\sum_{i \in E_{rt}} A(i) * RT(j|i) \end{aligned} \tag{4}$$

where

$$E_r = authors(repliesof(j))$$

$$E_m = authors(mentionsof(j))$$

$$E_{rt} = authors(retweetsof(j))$$

3.2.3 Acquaintance-Affinity-Identification Score

The final layer of scoring is the AAI score, which is a measure of how well a person is identified in the community and how likeable are the people in the community who identified the person in question.

To measure how well the person is identified by the people in the community, we use the follower measure. The rationale here is that following someone is a decision for the long term – indicating that we value their tweets in our timeline.

Then the Acquaintance-Affinity-Identification Score $AAI(j)$ is calculated as;

$$AAI(j) = \sum_{i \in followers(j)} \frac{AA(i)}{|follows(i)|} \tag{5}$$

Thus, the AA score of a person is divided among all the people followed by them, to contribute to their AAI score. The AAI score is tagged as the celebrity score.

3.3 The Action-Reaction Model

While the AAI model is based on scores assigned by users to other users, there is no focus on the amount of activity around a celebrity. This prompted us to develop another model based on observing how much of activity does a person elicit and how well it reflects his/her celebrity status. This model is called the Action-Reaction (AR) model. The AR model is based on the Friedman and Friedman model around attention, recall and loyalty. We arrived at two attributes of defining a celebrity:

- Attention
- Loyalty

The Action-Reaction model is based on two measures: Action measure and Reaction measure. The Action measure attributes the amount of *loyalty* fans pay to the celebrity. The attention is measured in terms of replies, mentions, retweets in Twitter. And the Reaction measure attributes to the amount of *attention* the celebrity elicits for every action.

The Action measure is a conditional probability measure. For a given person j, the Action measure of person i towards

j computes the probability that if i has acted on someone else's tweet, what is the probability that the tweet was from j.

$$A(j|i) = \frac{|A_m(i \to j) \cup A_r(i \to j) \cup A_{rt}(i \to j)|}{|mentionsby(i) \cup repliesby(i) \cup retweetsby(i)|} \quad (6)$$

where

$$A_m(i \to j) = mentionsof(j) \cap mentionsby(i)$$

$$A_r(i \to j) = repliesof(j) \cap repliesby(i)$$

and

$$A_{rt}(i \to j) = retweets(j) \cap retweetsby(i)$$

The Reaction measure is also a conditional probability measure. Given that the tweet from person j has elicited a response, the reaction measure for a target person i measures the probability that the response was from i. It is given by:

$$R(i|j) = \frac{|A_m(i \to j) \cup A_r(i \to j) \cup A_{rt}(i \to j)|}{|tweets(j)|} \quad (7)$$

The Action score $A(j)$ of person j is measured as the sum of all the action measures of the set of people S who acted upon j:

$$A(j) = \sum_{i \in S} A(j|i) \quad (8)$$

We normalize the Action score between 0 and 1. The normalized Action score $\|A(j)\|$ is calculated as follows:

$$\|A(j)\| = \frac{A(j)}{A_{max}} \quad (9)$$

where A_{max} is the maximum Action score across the sample.

Similarly, the Reaction score $R(j)$ of the person j is measured as the sum of all the reaction measures of the set of people S who reacted up on j.

$$R(j) = \sum_{i \in S} R(i|j) \quad (10)$$

We normalize the Reaction score between 0 and 1. The normalized Reaction score $\|R(j)\|$ is calculated as follows:

$$\|R(j)\| = \frac{R(j)}{R_{max}} \quad (11)$$

where R_{max} is the maximum Reaction score across the sample.

The action score $\|A(j)\|$ and reaction score $\|R(j)\|$ represents two dimensions of the celebrity j. The action score measures the *loyalty* of the person's fans and the reaction score measures the *attention* celebrity j elicits for every action. We represent the action-reaction measure as a vector $\vec{AR_j}$.

$$\vec{AR_j} = (\|A(j)\|, \|R(j)\|) \quad (12)$$

Since the scores are normalized, the maximum values they take is 1. We represent the "ideal celebrity measure" as $\vec{I} = (1, 1)$. We then measure the cosine similarity θ_j between the Action-Reaction vector $\vec{AR_j}$ and the ideal celebrity measure $\vec{I}$.

$$\theta_j = \frac{\vec{AR_j} \cdot \vec{I}}{\|AR_j\| \|I\|} \quad (13)$$

The celebrity score $C(j)$ is then represented as the product of the magnitude of $\vec{AR_j}$ and the cosine similarity θ between action-reaction $\vec{AR_j}$ and "ideal measure" $\vec{I}$.

$$C(j) = \theta_j * \vec{AR_j} \cdot \vec{I} \quad (14)$$

We can see that the celebrity score $C(j)$ can be simplified as:

$$C(j) = \frac{\|A(j)\| + \|R(j)\|}{\sqrt{2}} \quad (15)$$

4. EVALUATION RESULTS

Dataset

We have used Twitter APIs to collect tweets, user details and relationship between users (follow network). Twitter APIs impose limitations [8] on the number of requests per application for a time duration. Considering these limitations, We started of with a seed of 10 users, identified randomly from the tweet streams.

We collected all the tweets from the users' timeline. We also collected user details and the people who they followed. The new users identified in this step formed the level 1 users. This was repeated till a depth of three. The table 1 shows statistics about our evaluation dataset

Table 1: Tweet Data set Statistics

Number of Users	66 thousand
Number of Tweets	99 million
Number of Replies	6.9 million
Number of Mentions	21.9 million
Number of Re-tweets	4.8 million

Figure 1 shows the follower distribution in our data set. Figure 2 shows tweet distribution. The follower distribution and the tweet distribution of our data set resembles much of the large twitter data set [29]. We use this as an indicator of the representativeness of our sample.

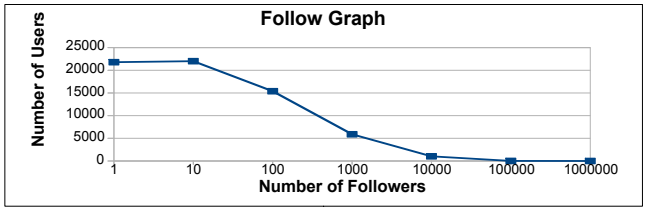

Figure 1: Followers Distribution

Celebrity and Influencers

For purposes of comparison, we identified two existing algorithms to find influencers: PageRank [5] and SNP (Social

[8] https://dev.twitter.com/docs/rate-limiting/1.1/limits

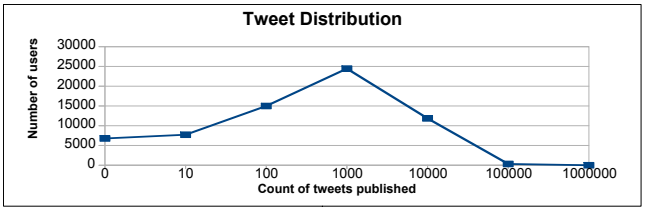

Figure 2: Tweet Distribution

networking potential) [2]. A more recent measure called the Twitter rank [29] was rejected in favor of PageRank. This is because, Twitter rank measures *influence* based on topical similarities, which is quite different from the celebritiness problem. "Celebritiness" is associated with a person and can be independent of topic. Hence we used the more straightforward PageRank measure for comparison. We used JUNG (Java Universal Network/Graph Framework) for computing PageRank.

We picked the top 25 celebrities identified from each of the algorithms (AAI model, AR model, PageRank and SNP) and we prepared a single list of celebrities for user evaluation by removing duplicates. We then presented the celebrity list of 65 celebrities to the 200 volunteers for user evaluation. We requested volunteers to identify or vote for celebrities from the given list. The user evaluation was "blindfolded" – meaning, the volunteers were unaware of the algorithms behind the celebrity list used for the evaluation.

We picked up the top 20 celebrity candidates based on user votes. Then we compared the top 20 celebrities voted by the volunteers with the top 20 celebrities identified by each of the algorithms.

Figure 3 shows that the AAI model performed well in terms of identifying celebrities. Celebrities identified by the AR model also agreed with user votes. But in addition to people, it also identified popular twitter accounts like YouTube, Überfacts, Funny Tweets and others, which were not identified as celebrities by users. We discuss the characteristics of the celebrities identified by AAI model and AR model in the later section.

We also computed the *average agreement* among the users based on the user votes in the top 20 results. Both AAI model and AR model showed an average agreement of 65% among users voting for these candidates.

Most celebrities identified in our algorithm have a good influence scores as well, going by their SNP measures. This result indicates that most celebrities are influencers where as not all influencers are celebrities. Table 2 shows the top 20 celebrities identified by various algorithms used in the experiment.

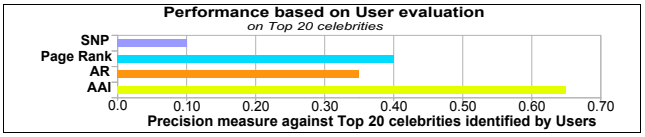

Figure 3: User evaluation results

We also compared the performance of the algorithms with Forbes top celebrities list and Klout Score [9]. We took top 25 people from the Forbes celebrity list Celebrities[10] and top 25 people from the Forbes India celebrity list (as the initial seed users considered in the sample were from India) [11] and merged them as 50 Forbes celebrities. We removed names from this list that did not feature in the sample at all.

[9] http://www.klout.com

Table 2: Top 20 Celebrities

AAI Model	AR Model	Page Rank Model	SNP Model
Barack Obama	Justin Beiber	OMG Facts	The Wanted
Kim Kardashian	Youtube	I Do That Too	Delta Goodrem
Amitabh Bachchan	Niall Horan	yfrog Social	Follow @WizKhalifa
Rihanna	Liam Payne	yfrog	ZoomTV
Justin Bieber	zaynmalik1D	Techmeme	Saj
Ellen DeGeneres	Shah Rukh Khan	Mediagazer	Beyonce Knowles
Oprah Winfrey	Jay Sean	Barack Obama	Star Plus
Shah Rukh Khan	Harry Styles	Kim Kardashian	iTunes
Lady Gaga	Yuvraj Singh	The White House	Annie Mac
Dalai Lama	Xstrology	Oprah Winfrey	Pope Francis
Kanye West	Amitabh Bachan	Ellen DeGeneres	LMAO
Priyanka	Uber Facts	jimmy fallon	RDB
OMG Facts	Ariana Grande	Office of VP Biden	New York Post
jimmy fallon	Lady Gaga	Conan O'Brien	DJ Khaled
Conan O'Brien	Funny Tweets	LMAO	Blake Griffin
Drizzy	Nicki Minaj	Michelle Obama	darkchild
Karan Johar	Jai J.D. Brooks	Kanye West	Keri Hilson
Ryan Seacrest	Mr.Carter	Rihanna	Sohanny
Abhishek Bachchan	Bruno Mars	Lady Gaga	Shriya Saran
Salman Khan	Rihanna	Ashton Kutcher	Kanye West

We computed the precision measure and Figure 4 shows the comparative performance of different algorithms. We observed that our AAI model performed well compared to other algorithms. This result shows the celebrities identified by the AAI model as also globally recognized celebrities.

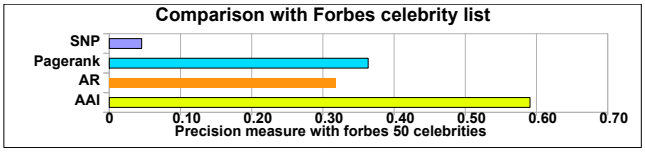

Figure 4: Comparison with Forbes 50 Celebrities

We identified the Klout score[12] manually for each of the top 25 celebrities identified by each of the algorithms. We plotted the Klout score against each celebrity identified from 4 algorithms. Figure 5 shows that the celebrities identified by the AAI model and AR model have consistent Klout score unlike the PageRank and SNP model.

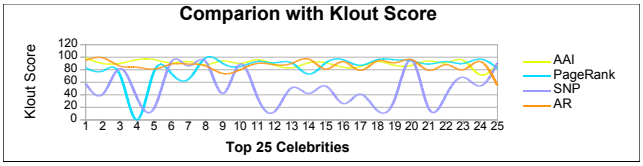

Figure 5: Comparison with Klout Score

In order to match the negatives, we measured the Klout scores of people who featured in the *bottom* 20 of the AAI and AR model as well. The average Klout scores for the top 20 and bottom 20 groups are showed in Table 3. The Klout scores seem to be positively correlated with both AAI and AR scores, but since it is a proprietary measure, we could not explore further than this.

[10] http://www.forbes.com/celebrities/list/

[11] http://forbesindia.com/lists/2012-celebrity-100/1395/1

[12] We used Klout score only as a comparative backdrop and is not intended for benchmarking, as the Klout score is a proprietary measure.

Table 3: Average Klout score Comparision

	AAI model	AR model
Average Klout score for Top 20	90	86
Average Klout score for Bottom 20	83	75

Real world Celebrity versus Twitter Celebrities

We analyzed the results of our experiment to find whether real-world celebrities are also Twitter celebrities. We plotted the AAI score versus number of followers in the graph. We expect Twitter celebrities to have high *AAI score* and real-world celebrities to have large number of followers.

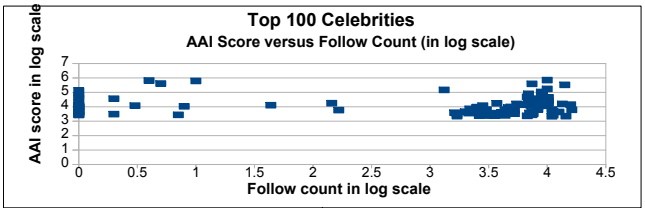

Figure 6: AAI score versus Number of Followers

Figure 6 shows strong co-relation between the *AAI score* and the number of followers. It also shows that a large portion of Twitter celebrities are also real-world celebrities. However, there is also significant portion of Twitter celebrities who do not have a large following. They seemed to have attained celebrity status based on their activities inside Twitter.

Twitter also has another interesting attribute called the "verified" flag. These represent Twitter accounts (typically of celebrities) that are manually verified to belong to the well-known personality. This is by far the most credible benchmark in Twitter for celebrity identification.

We picked the top 100 celebrities from AAI model and AR model and calculated the number of celebrities with the "verified" flag. We found a huge match of 95% of the celebrities in the AAI model as having the "verified" flag. The AR model accounted for a 80% match.

Comparing AAI and AR models

To understand more about the celebrities identified by the AAI model and AR model, we grouped the celebrities into 3 groups

1. Celebrities exclusive to AR model
2. Celebrities common to AR model and AAI model
3. Celebrities exclusive to AAI model

Table 4 shows the celebrities identified under the above mentioned groups. Figure 7 shows the average twitter statistics like number of mentions, number of replies, number of retweets, number of followers for the celebrity groups identified in the Table 4.

Figure 7 shows significant amount of mentions and the followers for the celebrities common to the AAI model and AR model. The celebrities exclusive to the AR model shows more Twitter actions like replies, mentions, retweets on the celebrity where as the celebrities exclusive to AAI model shows more of followers and less of actions when compared

Table 4: AR model and AAI model celebrities

Celebrity group	Celebrity names
Celebrities exclusive to AR Model	Niall Horan, Liam Payne zaynmalik1D, Xstrology UberFacts, Ariana Grande Funny Tweets, Jai J.D Brooks Mr. Carter
Celebrities common to AR and AAI Model	Justin Bieber, Shah Rukh Khan Amitabh Bachchan, Rihanna Lady Gaga
Celebrities exclusive to AAI Model	Oprah Winfrey, OMG Facts jimmy fallon, Conan O'Brien Ryan Seacrest

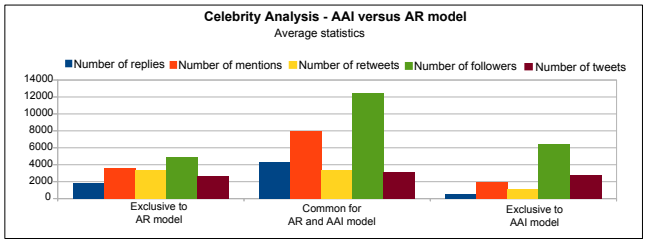

Figure 7: AAI celebrities versus AR celebrities based on Twitter statistics

to the celebrities exclusive to AR model. Figure 7 hints that the celebrities identified from the AAI model are more of real world celebrities and attracts more people making them well-known, well-liked and well-identified in the community. It also hints that the celebrities identified from the AR model are more of "Twitter celebrities" and are more popular within the Twitter community and attracts more actions within the community.

Figure 8 plots the average action score and reaction score for the celebrity groups identified in Table 4.

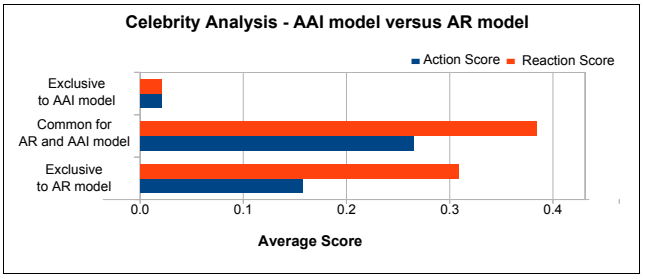

Figure 8: AAI celebrities versus AR celebrities based on AR score

Figure 8 shows that celebrities in general have a higher reaction score. This indicates that popular celebrities get discussed in the Twitter world irrespective of the celebrity tweets in terms of mentions.

Figure 9 plots the average celebrity score for AR model and AAI model for the celebrity groups identified in Table 4.

Figure 9 shows significantly higher AAI score for the celebrities exclusive to AAI model and the celebrities common to AR and AAI model. With respect to AR score, the

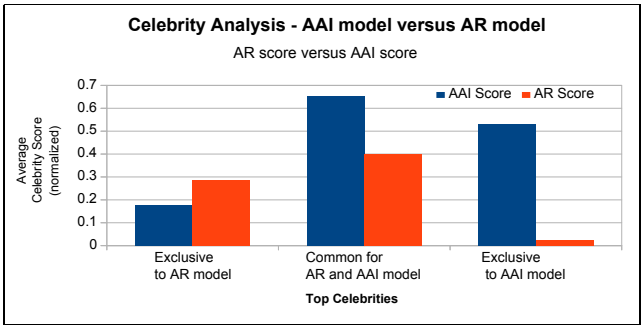

Figure 9: AAI celebrities versus AR celebrities based on celebrity score

celebrities exclusive to AR model and common to both AR and AAI shows significantly higher AR score compared to celebrities exclusive to AAI model.

Based on the observations from Figure 7, Figure 8 and Figure 9, we can understand that though the people identified by the AR model as well as AAI model are celebrities, the characteristics of the celebrities identified by the AR model and AAI model exhibit clear differences.

Celebrities identified by the AAI model show significantly high number of followers and good amount of actions, especially mentions. The celebrities with less action in the AAI model are compensated with their popularity and are well-identified by the people in the community and hence they appear as top celebrities.

Celebrities identified by the AR model show significantly high action (loyalty) scores and good amount of followers. Celebrities with less number of followers in the AR model are compensated with their popularity within the twitter community to generate significant number of actions and hence they appear as top celebrities.

Celebrities identified by the AAI model are more of real-world celebrities and attracts more people making them well-known, well-liked and well-identified in the community. Celebrities identified by the AR model are more of "Twitter celebrities" including non-person accounts like Überfacts that are popular within the Twitter community to generate significant number of actions.

Since both these algorithms seem to measure different signals, in an application setting one can envisage a generic celebrity score:

$$Celebrity(j) = \alpha * AAI(j) + (1 - \alpha) * C(j) \qquad (16)$$

where $0 \leq \alpha \leq 1$. When $\alpha = 0$, the identified celebrities are more of Twitter celebrities and when $\alpha = 1$, the identified celebrities are more of real-world celebrities.

5. CONCLUSIONS AND FUTURE WORK

Celebrity dynamics on social media is an interesting phenomenon, and the proposed algorithms provide promising results to be practically applicable. By providing interpretations for acquaintance, affinity, identification, loyalty and attention from appropriate signals, the proposed algorithms can be easily ported to datasets from other forms of social media. The distinction between AAI and AR distinguishes between celebrities within the social media versus celebrities in the real-world outside.

In the future, we plan to extend this model to identify celebrities in a given domain and apply this model on other forms of user generated content, in addition to social media datasets.

6. REFERENCES

[1] N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the influential bloggers in a community. In *Proceedings of the international conference on Web search and web data mining*, WSDM '08, pages 207–218, New York, NY, USA, 2008. ACM.

[2] I. Anger and C. Kittl. Measuring influence on twitter. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, page 31. ACM, 2011.

[3] M. Anjerani and A. Moeini. Selecting influential nodes for detected communities in real-world social networks. In *Electrical Engineering (ICEE), 2011 19th Iranian Conference on*, pages 1 –6, may 2011.

[4] D. J. Boorstin. *The image: A guide to pseudo-events in America*. Vintage, 2012.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.

[6] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *4th international aaai conference on weblogs and social media (icwsm)*, volume 14, page 8, 2010.

[7] D. Chen, J. Tang, J. Li, and L. Zhou. Discovering the staring people from social networks. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 1219–1220, New York, NY, USA, 2009. ACM.

[8] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 1029–1038, New York, NY, USA, 2010. ACM.

[9] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 199–208, New York, NY, USA, 2009. ACM.

[10] H. Deng, I. King, and M. R. Lyu. Formal models for expert finding on dblp bibliography data. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ICDM '08, pages 163–172, Washington, DC, USA, 2008. IEEE Computer Society.

[11] M. Forestier, J. Velcin, A. Stavrianou, and D. Zighed. Extracting celebrities from online discussions. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 322–326. IEEE Computer Society, 2012.

[12] M. Forestier, J. Velcin, A. Stavrianou, and D. Zighed. Extracting celebrities from online discussions. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, ASONAM '12, pages 322–326,

Washington, DC, USA, 2012. IEEE Computer Society.

[13] H. Friedman and L. Friedman. Endorser effectiveness by product type. *Journal of Advertising Research*, 19(1):63–71, 1979.

[14] R. Ghosh and K. Lerman. Predicting influential users in online social networks. *arXiv preprint arXiv:1005.4882*, 2010.

[15] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Discovering leaders from community actions. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 499–508, New York, NY, USA, 2008. ACM.

[16] B. Hajian and T. White. Modelling influence in a social network: Metrics and evaluation. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 497–500. IEEE, 2011.

[17] D. Hatcher, G. S. Bawa, and S. Barry de Ville. How you can identify influencers in sas® social media analysis (and why it matters).

[18] C. I. Hovland, I. L. Janis, and H. H. Kelley. *Communication and persuasion: Psychological studies of opinion change*. Yale University Press New Haven, CT, 1953.

[19] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[20] C. Li, S. Lin, and M. Shan. Finding influential mediators in social networks. In *Proceedings of the 20th international conference companion on World wide web*, pages 75–76. ACM, 2011.

[21] Z. Luo, M. Osborne, J. Tang, and T. Wang. Who will retweet me?: finding retweeters in twitter. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 869–872. ACM, 2013.

[22] G. McCracken. Who is the celebrity endorser? cultural foundations of the endorsement process. *Journal of Consumer research*, pages 310–321, 1989.

[23] W. J. McGuire. The nature of attitudes and attitude change. *The handbook of social psychology*, 3:136–314, 1969.

[24] C. W. Mills. The power elite [1956]. *New York*, 1981.

[25] H.-K. Peng, J. Zhu, D. Piao, R. Yan, and Y. Zhang. Retweet modeling using conditional random fields. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 336–343. IEEE, 2011.

[26] M. S. Srinivasan, S. Srinivasa, and S. Thulasidasan. Exploring celebrity dynamics on twitter. In *Proceedings of the 5th IBM Collaborative Academia Research Exchange Workshop*, I-CARE '13, pages 13:1–13:4, New York, NY, USA, 2013. ACM.

[27] G. Turner. *Understanding Celebrity*. SAGE Publications, 2004.

[28] Y. Wang, G. Cong, G. Song, and K. Xie. Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1039–1048. ACM, 2010.

[29] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM, 2010.

[30] J. Zeng, W. K. Cheung, C. hung Li, and J. Liu. Coauthor network topic models with application to expert finding. *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, 1:366–373, 2010.

[31] L. Zhou. Trust based recommendation system with social network analysis. In *Information Engineering and Computer Science, 2009. ICIECS 2009. International Conference on*, pages 1 –4, dec. 2009.