# Social Citation: Finding Roles in Social Networks. An Analysis of TV-Series Web Forums

Nikolay Anokhin[1], James Lanagan[1], and Julien Velcin[2]

[1] Technicolor
1, avenue de Belle Fontaine - CS 17616
35576 Cesson-Sévigné Cedex, France
{nikolay.anokhin,james.lanagan}@technicolor.com
[2] ERIC Lab
5 av. Pierre Mendès-France
69676 Bron Cedex, France
{julien.velcin}@univ-lyon2.fr

**Abstract.** In this paper we present preliminary work studying the interactions of a community of focussed forum users and their discussions around several television series. We use k-means clustering and a number of novel citation-analysis inspired measures to perform bottom-up role detection on this community of TV fans, and show that these emergent roles correspond well with the positions assigned to users using traditional graph-based measures.

**Keywords:** Citation Analysis, Role Detection, Social Network Analysis

## 1 Introduction

When looking to spread information about new products or services, companies seek to find the most influential or connected people within their target community so as to achieve the greatest return for any investment. In the current work we look to the interactions of users on a forum and show that these interactions place certain users into certain roles. Citation analysis is an interesting approach to the problem of role identification as it allows us to build on the inherent ideas of importance of the work of an author. The primary goal of this work is to bring together existing approaches from citation analysis and machine learning so as to discover the different roles which forum users play. The major contributions of this paper are two-fold:

– We propose two measures adapted from the citation analysis domain – Node g-Index and Catalytic Power – and combine these with 2 existing measures – Cross-Topic Entropy and Generalised Degree – so as to quantify the interaction and importance of users within a community of forum users.

– By considering the contributions of users to a forum within a sliding time window, we are able to perform unsupervised clustering on the temporal representations of users to assign a principal (most representative) role. We show that the assigned roles correspond well with the positions assigned to users using traditional graph-based measures.

In the next section we shall detail some of the related work before describing the collection and statistics of our experimental corpus. Section 4 details the new adapted measures that we have created so as to characterise our forum users. The experiments performed in this work are detailed in Section 5. Finally we discuss our conclusions and present some future directions for further research.

## 2 Related Work

We see the study of importance as a side-effect of social role detection. As with much past work [1, 2], we shall focus on the identification of 'important' actors from within a network. This does however provide a good starting point for future work on role categorisation and identification.

There has been relative little work on studying roles through time or the dynamics of social roles [3]. We look to incorporate the temporal aspect of roles into our work by examining the interactions of users as a function of the current state of the network. We will not focus on the dynamics of role attribution but feel that this provides an important distinction from current approaches.

Garfield [4] noted many reasons for the citation of articles within a work that are strongly linked to the reasoning behind online conversations. We wish to use the approaches from citation analysis to provide a measure of interaction and importance to our forum users. This differs from purely graph-based measures as we intend to take into account only those interactions that have been judged sufficiently interesting.

## 3 Corpus Creation

Our corpus was created by crawling the TWOP[3] website. The website is designed for entertainment content providing forums for communities to converge and discuss TV shows, and contains dedicated sub-fora each focussed on a single television series. Each sub-forum contains many threads of conversation allowing us to tag each thread in our corpus with a single series of interest. We focus our analysis on a year of forum posts discussing 6 television series each having their own dedicated sub-forum. The shows were featured in the "Top Shows" categories (containing 8-10 television series) on the site (Table 1). A targeted crawl of the 6 sub-forums of interest retrieved 58,994 posts created by 7,066 authors. This number is different from that in Table 1 as some authors are active in more than one sub-forum; we use this to our advantage when identifying the activity profile of users.

[3] http://forums.televisionwithoutpity.com/

**Table 1.** Breakdown of corpus by sub-forum.

| Forum | Genre | Threads | Posts | Authors |
|---|---|---|---|---|
| American Idol | Reality TV | 40 | 15,907 | 2,354 |
| Dexter | Suspense | 44 | 3,296 | 596 |
| Grey's Anatomy | Drama | 52 | 12,319 | 1,926 |
| House | Drama | 58 | 11,920 | 1,371 |
| Mad men | Drama | 27 | 2,943 | 471 |
| The Office | Comedy | 71 | 12,609 | 1,692 |
| Totals | | 292 | 58,994 | 8,410 |

### 3.1 Corpus Preparation

After crawling the forum, it was necessary to reconstruct the threads of discussion. We build a user network that takes into account communications between users: the more reply-response pairs there are between users $A$ and $B$, the stronger their relationship. Having this idea in mind, we build our user network as an undirected weighted graph[4].

TWOP uses a 'quote' mechanism meaning that the quoted text of a message appears before the text of a reply. We used a series of regular expressions, as well as Levenshtein distance (95% overlap) to detect the parent-child relationship within the past 20 in-thread messages. This threshold was chosen empirically as it provided us with high retrieval. We manually checked the efficacy of the proposed thresholds across 10% of the corpus and found a retrieval rate 100% for all quoted parent texts. Note that mentions of authors' names within messages were not retrieved, though this method of citation/reply is used very infrequently in comparison to the quote functionality.

## 4 Measuring Influence

Influence may be defined as *"the act or power of producing an effect without apparent exertion of force or direct exercise of command"*. Several *node centrality* measures developed in graph theory are traditionally used to identify the "most important" actors in a social network. They assume that important authors occupy central positions in the network graph: *degree centrality*, *betweenness centrality* and *closeness centrality*. As stated later in Section 5, we found that the highly-ranked nodes according to these measures were clustered together by our own measures.

Citation analysis is an early form of linkage analysis [4]. In this context, we shall use messages as analogous to publications. The most widely accepted citation analysis metrics are Hirsch's *h-index* [5], as well as the *g-index* which is a direct variant of the former. It considers only those items (the H-CORE) that are significant enough to have received a predefined number of replies/citations:

[4] Although originally based on a directed graph, we shall perform citation analysis on an undirected graph as the communications/citations often form part of a larger two-way conversation

> An author has an **h-index** of $h$ if $h$ of his total contributions have received at least $h$ citations each.

The g-index was an improvement on this again as it also takes into account the distribution of the items within the h-core.

> An author has a **g-index** $g$ if $g$ is the highest rank such that the top $g$ contributions have, together, at least $g^2$ replies. This also means that the top $g+1$ messages have less than $(g+1)^2$ replies.

### 4.1 Social Citation

In the following sections we describe 4 features (including 2 novel features adapted from the citation analysis domain) selected to best represent the different aspects of a forum user with regards to their overall output and interactions within the community. By using these features we hope to identify the roles that are played by users within the community (forum) [1]. The approach is exploratory: the idea is to let the roles' typology emerge using different kinds of measures.

*Generalised Degree* The GENERALISED DEGREE of a node $A$ is defined as $D(A) = \sum_{B \in N(A)} W(A, B)$ where $N(A)$ is the set of neighbours of $A$. $W(A, B)$ is the number of communications between $A$ and $B$. This feature expresses how active the user is taking into account all communications of the user with his neighbours. On a directed graph this would be equivalent to the combined in and out degrees.

*Node g-Index* This feature evaluates how active neighbours of the node are. G-INDEX is calculated as the highest number $g$ of neighbours such that sum of their generalised degrees is $g^2$ or more.

*Catalytic Power of Users* The CATALYTIC POWER of a message $k$ reflects the amount of reaction from other users caused by the message: messages with high catalytic power produce (catalyse) large discussions. In the context of forum discussions, we use the number of direct replies, $c_k$, to estimate this power. For a user we have a list of catalytic power of her messages $\hat{c} = c^1_{k_1}, c^2_{k_2}, \ldots, c^n_{k_n}$. The *catalytic power of a user* is consequently defined as the sum of powers from the h-core of $\hat{c}$: $C = \sum_{i=1}^{h} c^i_{k_i}$

*Cross-Topic Entropy* The idea of using ENTROPY to measure user's focus on a particular topic has been used previously [6]. Let us consider a user who posted $n_i$ messages across all the threads that discuss topic $i$ (in our case the name of the discussed show was considered a topic). Let $n = \sum_i n_i$, then focus of a user is defined as $F = \sum_i -\frac{n_i}{n} \log \frac{n_i}{n}$. This measure helps to distinguish between users who contribute to many topics across the forum from users who focus on a single topic (e.g. fans).

### 4.2 Temporal Aspects of Roles

Although analysis of users' features obtained from the whole data set (the entire calendar year 2007) reveals general patterns of interactions between users, it is also important to consider how values of those features evolve over time. We observe weekly peaks of activity on the broadcast dates of new episodes of each series, and almost no activity in summer when none of the six TV shows are broadcast. In addition, activity of each user is far from uniform (see Figure 1). In the current work we shall assign a user to their most representative role, but we do so by taking into account their weekly role within the community.

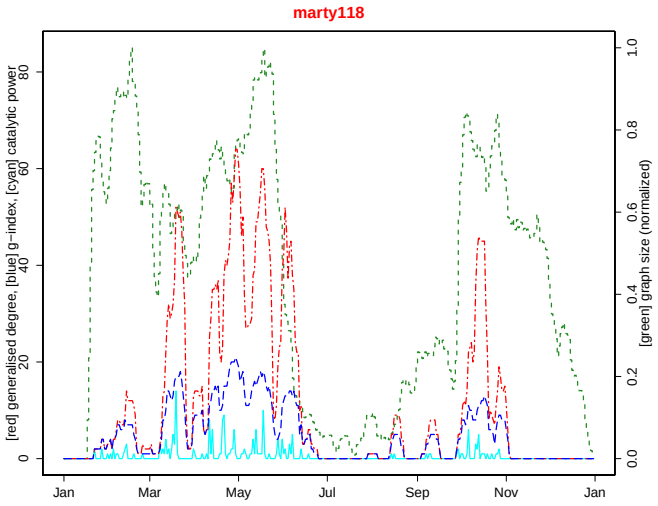


**Fig. 1.** The activity of a single user over the year, using the windowing method. General forum activity (green dash) is shown, along with g-index (blue long-dash), generalised degree (red dot-dash), and catalytic power (cyan solid).

In order to include this temporal aspect in our analysis, we need to be able to calculate the value of each of the features described in Section 4 for any user at specific moment of time.

Seven days appeared to be a natural window for our data, as new episodes of the TV shows in our data set are released on a weekly basis. For every user we calculated a time series of 365 feature vectors, one for each day of the year 2007 observing all interactions within the graph in the past 7 days. Each feature vector contains four values: a user's degree, g-index, catalytic power and cross-topic entropy. The all-zero feature vectors were excluded as they represent moments when users were not part of the community: a user posted a new (non-reply) message, but received no replies to this message. As a result 105,948 feature vectors that belong to 4,291 forum user were retained.

**Table 2.** Summary of cluster centroids.

| Cluster | Generalised Degree | g-Index | Catalytic Power | Cross-Topic Entropy | Number of Observations | Number of # Users |
|---|---|---|---|---|---|---|
| 1 | 0.00 | 0.00 | 0.00 | 5.80 | 16158 | 430 |
| 2 | 1.40 | 3.02 | 0.00 | 0.00 | 22235 | 848 |
| 3 | 1.39 | 2.98 | 3.10 | 0.00 | 31294 | 1879 |
| 4 | 1.64 | 3.29 | 3.22 | 5.73 | 4867 | 32 |
| 5 | 2.49 | 4.17 | 3.58 | 0.00 | 19463 | 654 |
| **6** | **4.04** | **5.31** | **4.52** | **0.00** | **6649** | **108** |
| **7** | **3.48** | **5.02** | **4.03** | **5.37** | **2501** | **17** |
| 8 | 1.45 | 3.07 | 0.00 | 5.64 | 2781 | 10 |

## 5 Experiments

Every user in our collection is now represented by a time-series of four-dimensional daily feature vector observations. Users may take different roles at different times throughout the year, but remain predominantly a single role.We choose to assign the most frequently-occurring and representative role to a user.

In order to surface these roles we must identify the number of possible roles that exist. We perform a k-means clustering (re-initialised 200 times) on the 105,948 vectors created in the previous section. Although k-means is an unsupervised algorithm, it is necessary to provide the initial number of clusters, $k$, to be created. As we do not make any a priori assumptions on the number of social roles, we have to infer $k$ from the given data.

Out of several existing techniques we chose, $H(k)$, the Hartigan Index. This is a rule-of-thumb proposed by Hartigan, that has been shown to be a highly-effective approach for finding the correct number of clusters [7]. The optimal value for $k$ is the value that maximises $H(k)$. Applying Hartigan's Index (re-initialised 20 times) on the given data set resulted in the optimal number of clusters $k = 8$. For each of our features, $x$, we performing min-max normalisation (0-128) and transform the value to $log_2 x$ before clustering.

### 5.1 Clustering Roles

Table 2 presents the cluster centroids along with the numbers of feature vector observations and users within the respective clusters. Within Table 2 we can see a number of obvious divisions that have been captured. Cluster 1 for example contains all observations where a user has posted to many sub-fora (hence the high entropy) but received no replies. Cluster 2 by contrast shows examples of users having replied to well connected (due to the high g-Index and generalised degree) users' posts, but without managing to generate conversation. Cluster 3 is the largest cluster and represents the largest section of the user population. Here we can see users who have been reasonably active talking to many different users of no specific prominence. There is a low generalised degree meaning that they do not receive a lot of replies, but they are capable of creating interesting/catalytic

**Table 3.** Top 10 users according to PageRank. Node centrality values are shown for users who appear in the corresponding Top 10 rankings.

| User | P.Rank (x10$e^{-2}$) | Degree (x10$e^{-1}$) | Between. (x10$e^{-1}$) | Close. | Cluster (% vectors) |
|---|---|---|---|---|---|
| marty118 | 0.492 | **0.455** | 0.670 | 0.340 | 7 (53.51) |
| claudiaj | 0.420 | 0.331 | | | 6 (76.07) |
| english toffee | 0.373 | | 0.602 | 0.342 | 1 (30.02) |
| Energiya Buran | 0.301 | | 0.342 | 0.323 | 1 (32.89) |
| Maybe Once | 0.286 | | 0.249 | | 5 (21.71) |
| Nikki528 | 0.280 | 0.345 | 0.329 | 0.318 | 6 (57.89) |
| Bessie Mae | 0.270 | 0.435 | **0.758** | **0.348** | 7 (59.90) |
| jjr | 0.263 | | | | 6 (69.94) |
| Mozzy55 | 0.263 | | | | 3 (64.83) |
| LollipopGal82 | 0.262 | 0.321 | 0.357 | 0.332 | 6 (42.13) |

posts. Conversely, Cluster 4 contains observations when users create slightly more controversial messages spread across several sub-fora.

Two clusters presented in Table 2 are of special interest for us. Cluster 7 includes feature vectors with high values of all the features, and so, members of this cluster can be seen as potentially important in several topics. Members of Cluster 6 also have high characteristics in all the features except for entropy. These users may be considered important in a single topic. Note that clusters 6 and 7 contain ∼3% of users: this correlates well with past research [8].

### 5.2 Validating Social Citation

In order to compare our approach with conventional techniques for identifying important users we calculated degree, betweenness, and closeness centralities as well as the PageRank of each user [9]. Many of the same users returned in the high-valued clusters appear as highly-central nodes in the graph (Table 3). There are some users that seem to be classified very differently by the two approaches. *Mozzy55* for example is the only user from Cluster 3 to make it into the PageRank-based Top 10. PageRank is performed on a static graph containing the entire dataset. *Mozzy55* receives very few replies, but from highly connected users at sporadic intervals throughout the year.

The rule of assigning users to roles using the majority of vectors has some disadvantages. For instance, users “english toffee” and “Energiya Buran” were both assigned to Cluster 1 since this is there most prevalently and accordingly representative cluster. Upon inspection we saw that both of these users have almost as many occurrences in Cluster 7 respectively. These users appear to be inactive for large proportions of the year, though capable of generating large amounts of conversation when they are active.

## 6 Conclusions & Future Work

In this work we addressed the problem of identifying social roles of users of internet forums. We see several opportunities to improve the proposed features. In particular it may be interesting to use a directed rather than undirected network when calculating generalised degree and g-index, as it may help distinguish such social roles as "answer people" [1]. Catalytic power can also be improved by looking beyond direct replies.

Although our approach helps to identify important users despite inactivity for a period of time, we see that there is important work to be continued on the dynamic evolution/transference from one role to another. In the current work all of the features that we examine are given equal weighting within the feature vector. This may not be the optimal solution since entropy, say, may be far less indicative of a role than local interactions. We intend to examine in future different weighting schemes for our features.

In summary we have proposed four measures that aim to capture the interactive style and behaviour of forum users. We have shown that these measures perform well in identifying and clustering users of similar styles, and as a consequence help in the identification of influential users or super-spreaders.

## References

1. Welser, H.T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., Smith, M.: Finding social roles in wikipedia. In: iConference '11: Proceedings of the 2011 iConference, Seattle, WA, ACM (2011) 122–129
2. Bakshy, E., Hofman, J., Mason, W., Watts, D.: Identifying 'Influencers' on Twitter. In: Fourth ACM International Conference on Web Seach and Data Mining (WSDM), Hong Kong, China, ACM (2011)
3. Forestier, M., Stavrianou, A., Velcin, J., Zighed, D.: Roles in Social Networks: Methodologies and Research Issues. WAIS **10**(1) (2012) 117–133
4. Garfield, E.: Concept of Citation Indexing: A Unique and Innovative Tool For Navigating The Research Literature. (1997)
5. Bornmann, L., Mutz, R., Daniel, H.: Are There Better Indices for Evaluation Purposes Than the h-Index? A Comparison of Nine Different Variants of the h-Index Using Data From Biomedicine. JASIST **59**(5) (2008) 1–8
6. Jamali, S., Rangwala, H.: Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis. In: WISM'09: International Conference on Web Information Systems and Mining. (2009) 32–38
7. Chiang, M.M.T., Mirkin, B.: Experiments for the Number of Clusters in K-means. In: Proceedings of the 13th Portuguese Conference on Progress in Artificial Intelligence. EPIA'07, Guimarães, Portugal, Springer-Verlag (2007) 395–405
8. Whittaker, S., Terveen, L., Hill, W., Cherny, L.: The Dynamics of Mass Interaction. In: CSCW'98: Proceedings of the 1998 ACM Conference on Computer Supported Co-operative Work. CSCW '98, Seattle, Washington, United States, ACM (1998) 257–264
9. Page, L., Brin, S., Motwani, R., Winograd, T.: The Pagerank Citation Ranking: Bringing Order To The Web (1998)