Connectivity, Online Social Capital and Mood: A Bayesian Nonparametric Analysis

Dinh Phung, IEEE Member, Sunil Kumar Gupta, Thin Nguyen and Svetla Venkatesh, IEEE Fellow
Centre for Pattern Recognition and Data Analytics (PRaDA)
Faculty of Science and Technology, Deakin University, Australia.
Email: {dinh.phung,sunil.gupta,thin.nguyen,svetha.venkatesh}@deakin.edu.au

Abstract—Social capital indicative of community interaction and support is intrinsically linked to mental health. Increasing online presence is now the norm. Whilst social capital and its impact on social networks has been examined, its underlying connection to emotional response such as mood, has not been investigated. This paper studies this phenomena, revisiting the concept of “online social capital” in social media communities using measurable aspects of social participation and social support. We establish the link between online capital derived from social media and mood, demonstrating results for different cohorts of social capital and social connectivity. We use novel Bayesian nonparametric factor analysis to extract the shared and individual factors in mood transition across groups of users of different levels of connectivity, quantifying patterns and degree of mood transitions. Using more than 1.6 million users from LiveJournal, we show quantitatively that groups with lower social capital have fewer positive moods and more negative moods, than groups with higher social capital. We show similar effects in mood transitions. We establish a framework of how social media can be used as a barometer for mood. The significance lies in the importance of online social capital to mental well-being in overall. In establishing the link between mood and social capital in online communities, this work may suggest the foundation of new systems to monitor online mental well-being.

I. INTRODUCTION

Social creatures we are. We smile, we laugh, with others. We cry - the world seems a better place if we are with someone. We just feel better when we are supported, when we interact. It is strange that we do not always remember how important this is.

Mortality has long been taken as an indication of community health and well-being [55]. As early as 1945, Spitz [45] showed a startling result: mortality rates for infants in institutional care was as high as 71% - without social contact, infants died. In 1997 Kawachi et al. [22] show strong relation between mortality and social capital, measured by trust levels and membership of community groups. For example, a 10% increase in trust would lead to a reduction in .06 deaths per thousand people [22]. This link has been repeatedly established (e.g., [39] [5]). More recently, Holt-Lunstad [18] declare a sobering result: “... data across 308,849 individuals, followed for an average of 7.5 years, indicate that individuals with adequate social relationships have a 50% greater likelihood of survival compared to those with poor or insufficient social relationships. The magnitude of this effect is comparable with quitting smoking and it exceeds many well-known risk factors for mortality (e.g., obesity, physical inactivity)”.

Kawachi et al. [21] define social capital as “... consists of features of social organization – such as trust between citizens, norms of reciprocity, and group membership – that facilitate collective action”. Social disconnectedness can be measured through size of social network, level of infrequent participation in social activities and perceived isolation through aspects of loneliness and perceived lack of social support [5]. Social capital, in main part includes: social inclusion, social participation and social support [53]. Social inclusion implies access - for example, older people may have less physical access to social groups. Social participation describes how much a person engages with a community. And, social support implies how much support one can get from the community.

This support and interaction, traditionally, has been derived from physical social networks - family, churches, hobby groups. We have, on the other hand, increased online connectivity evidenced by the unprecedented growth of many social networking sites. What, if any, is the relation between aspects of mental health, say mood, and “online social capital”? At a more fundamental level, can we measure it through social media? This paper examines these questions.

At the level of population scale, social media offers us a rich opportunity. People blog. They join groups, they comment on blogs. Others comment on their blogs. They follow groups, stay or leave them. Interestingly, they are willing to tell others how they feel! This media offers a rich opportunity to explore the relation between social capital and critical aspects of mental well-being - mood and mood swings. Are moods different for groups with different levels of social connectedness? What are the main components, latent factors of mood swing? Essentially, how good a sensor is social media for social capital’s connection to mood and well-being? We define online social capital in online communities. Continuing with the core components of social participation and support, we define three categories of LOW, MEDIUM and HIGH social capital. For each of this category, we define six measurable aspects:

- **Social participation**: number of groups joined, number of posts written, number of comments made.
- **Social support**: number of friends, number of comments received, number of followers.

We thus have 18 cohorts to study - 3 categories of social capital, and in each, 3 measures of social participation and 3 of social support. Using a dataset of more than 1.6 million users...
in LiveJournal over 10 years, we explore across all measures of social participation and social support: (a) distribution of mood over LOW, MEDIUM and HIGH social capital; (b) distribution of mood swings in LOW, MEDIUM and HIGH social capital; (c) Mood swing factors that are common and different for cohorts. From more than 1.6 million users, we extract relevant subsets of social capital cohorts for LOW, MEDIUM and HIGH, resulting in 90,000 users. We show through our analysis that cohorts with lower social capital when compared to cohorts higher social capital will have:

- More negative moods, and lesser positive moods;
- More mood swings, and more mood swings to negative moods;
- More underlying factors composed of negative mood swings.

To understand the underlying factors that are common in mood transitions across different cohorts, we use Bayesian nonparametric factor analysis. When there are more than one data cohort and they are related, it is useful to perform joint factor analysis that exploit shared statistical properties. A naive way to perform joint factor analysis is to augment the data from each cohort into a single dataset as if they come from the same source and perform factor analysis. However, this is often suboptimal because each data cohort has its own variations, differing in distribution. Shared subspace learning models [23], [24], [25], [1], [16] exploit the sharing strengths, whilst preserving the individual variations. They learn common factors across cohorts but also ones specific to each cohort. We show how to apply our recent model [11], [16] to this application, extracting shared and individual factors across the cohorts being considered. This model has two main advantages: it uses Bayesian nonparametric prior, thus not requiring a priori parameters such as numbers of factors, and it allows us to analyse the individual and shared aspects across 18 cohorts in a rigorous Bayesian framework. This allows us to understand the mood transition patterns. We show quantitatively that mood transition to negative moods are commonly seen in cohorts with low social capital; cohorts with high social capital mostly have positive mood transitions.

Online social capital is nascent, in definition and scope. Our preliminary work [23] attempted to examine online social capital in two extremes, high and low, and different from this work, it is limited to the extraction of basic features, derived from the content and sentiment, for classification among capital groups. [23] propose a broad definition of social capital - a combination of user profile information, factors reflecting openness to new acquaintances, online activities such as frequency of posting or commenting and the contribution of the user to relationships. The model is broad across all types of social media, thus many components are not computable and no results are provided. Other early works include [13], [52], [7], [23], [1] - all focus on aspects of online social capital as sensors to monitor health and growth of the online social networks. Crucially, the user as the point of impact has never been considered - its connection to the emotional structures, such as user’s mood with implication to mental well-being, has never been addressed before. There is also a recent surge in information retrieval research on new media and call for further research in social media information retrieval [28]. Our work might provide a novel approach to such effort through aspects of social capital, mood and mental health of the users.

The novelty of this work lies in: a) We view social capital as a sensor to understand the users in online communities, not the social network as in previous work; b) the formulation of computable online social capital using social connectivity, and its analysis through large scale weblogs; c) extraction of a discriminative primary mood set for understanding of user’s mood across groups of social capital, and d) the demonstration of underlying mood transition factors using novel nonparametric Bayesian factor analysis techniques, enabling separation of factors across different cohorts of social capital, formally establishing for the first time the connection between social capital in online social networks and user’s mood.

The significance of this work lies in using social media as a barometer for mood. This is critical. Holt-Lunstad et al [18] extort: “... physicians, health professionals, educators, and the public media take risk factors such as smoking, diet, and exercise seriously” – the data presented here make a compelling case for social relationship factors to be added to that list. With such recognition, medical evaluations and screenings could routinely include variables of social well-being; medical care could recommend if not outright promote enhanced social connections; hospitals and clinics could involve patient support networks in implementing and monitoring treatment regimens and compliance, etc. Health care policies and public health initiatives could likewise benefit from explicitly accounting for social factors in efforts aimed at reducing mortality risk. Individuals do not exist in isolation; social factors influence individuals’ health though cognitive, affective, and behavioral pathways."

We believe that the results in this paper are the first steps in this fledgling area of measurably accounting for social factors.

II. BACKGROUND
A. Mood, Social Media and Weblogs

The Web is possibly the richest sensor the world has ever seen, monitoring a global, heterogeneous web of human and machine interactions. There are billions of daily transactions, facilitated by proliferation of modern communication devices and cheap broadband. The web today is two-way: users are not only able to read commercially owned content, they can respond to it. This adoption of user-contribution by the on-line arms of traditional media and industry seems to have been driven by the emergence of social media—YouTube, Flickr, Facebook, Twitter, Google+ and blogs—through which whole communities contribute media of various types and transact via on-line communication channels. Social media’s uniquely ‘egocentric’ nature means that these communities and the artefacts they produce constitute sentiment-laden corpora in their own right. As a representative of the new media,
weblogs tend to be more subjective than other genres. Bloggers are more comfortable in expressing their feeling, attitude and ideas towards real life events. This motivates the blogosphere for a subject of study in terms of subjectivity and sentiment analysis [55].

Existing line of research focuses on subjectivity detection, sentiment classification, joint-topic sentiment analysis and opinion summarisation (see [37] for a comprehensive survey). In this paper, we use sentiment in the form of mood – strong form of sentiment expression, conveying a state of the mind such as being happy, sad or angry.

Text-based mood classification and clustering is a sub-problem of opinion and sentiment mining, with diverse applications - automated recommendation for product websites, business and government intelligence, collection of empirical evidence for studies in psychological and behavioural sciences. Mood classification in the blogosphere can be used to filter search results - to ascertain the mental health of communities, or to gain detailed insight into patterns of how bloggers behave and relate to one another.

Early work of mood prediction in weblogs were reported in [32], [27] and [44]. Predicting a user’s mood from their written content is typically recognized as a much harder problem than traditional text categorization in machine learning and data mining due to the complexity of human emotional structure and subjective factors. For example, in the case of weblogs, the manifestation of mood is highly affected by one’s idiosyncratic vocabulary and style, with messages often reflecting social context, including community norms, history and shared understanding. Consequently, text is often short, informal, punctuated by jargon, abbreviations and frequent grammatical errors.

More recent work has utilized collective mood conveyed in social media over time for prediction and event extraction tasks [2, 25, 46]. For example, [2] use 6 collective mood states from approximately 1 million tweets to predict the closing values of the Dow Jones Industrial Average index; [25] analyzed approximately half of a billion tweets originated from the UK and found the correlation of an increase in negative moods and the government announcement time of cutting public spending; but also found that a calming effect in the public mood during the royal wedding. More advanced machine learning methods were developed in [46] to predict individuals’ emotional states in social networks that incorporates contextual information by way of designing suitable potential functions under a probabilistic factor graph formulation. Prediction performance was shown to improve when the information of users, including their activities, locations, recent moods, and the emotional states of their social friends, was incorporated. This is partly confirming the finding of Fowler and Christakis [9] that there exists the spread and influence of mood among friends and within cohorts. Different from [46] which focuses solely on the prediction task, our claimed contributions in this paper are entirely different and do not make any attempt on prediction or classification tasks; in contrast, we aim to establish link between mood and social capital and characterize mood swing patterns among various social capital groups, hence numerical comparison with the work of [46] is infeasible. Affect prediction is also found in other types of social media including the work [20] which attempted to infer emotions from images crawled from Flickr. However, these existing works focus exclusively on the problem of sentiment-based prediction task, whereas none has considered the aspect of online social capital, its connection to mood and mood swing patterns as in our work. Our early work [55] has also proposed a computational model for extraction of sentiment index over time, used a Markov model for automatically extracting sentiment events, jointly analyzing sentiment information with the contents using Bayesian topic models – however, the link between users’ mood and social capital has never been established before.

B. Social capital and mental well-being

Although the concept of ‘social capital’ in physical social networks is relatively well-established, its definition to online social networks is still unclear and debatable [23]. Early proposal that the online presence increases people’s social capital and call for advancing science in this research direction were published in [54]. Specific case studies of online social capital have been conducted; e.g., [44] proposed model of social capital based on the resources and actions; [52] examines online social capital for Facebook on a cohort of college students; again [4] examined the social capital derived from Facebook for 415 users from a longitudinal survey; [6] studied online social capital and information flow specifically for Brazilian weblogs. To our knowledge, the most generic computational model of social capital in online social networks is the work of [23] which defines social capital for each user as a linear combination of the following functional components of the user: static component (doesn’t change with time, e.g., derived his or her profile), Marched-by-Search (reflecting his/her openness to new acquaintances), Activity (characterizing online activities such as frequency of posting or commenting) and Social Position (describing the value of the user in the network derived from social relationships). However, the work of [23] stops short at a theoretical model, no empirical results were reported in [23]. Furthermore, evidences of positive correlation between social capital and online interaction has been found in numerous research (e.g., [11]).

Now, turning our attention to the traditional social capital and its link to mental health before the population of social networking sites, in 1997 Kawachi et al [22] showed strong relation between mortality and social capital, measured by trust levels and membership of community groups. For example, a 10% increase in trust would lead to a reduction in .06 deaths per thousand people [22]). Piliavin et al [39] find a causal relation between better mental health and social interaction, as measured through volunteering. Cornwell and Waite [5] combine multiple indicators of social isolation to assess social disconnectedness - social network size, participation in social activities, perceived isolation. They find “that social disconnectedness and perceived isolation are independently associated with lower levels of self-rated physical health.”
D. Bayesian Nonparametric Analysis

Bayesian nonparametric models are Bayesian models with nonparametric priors so that the complexity of the model may grow unboundedly with the data. The work of Ferguson [8] has made a breakthrough in the progress of this line of research by introducing a tractable family of nonparametric prior distributions known as the Dirichlet processes over the space of discrete distributions. In a similar realm, Hort [17] introduces Beta processes as the nonparametric prior distribution for hazard functions. However, to our own bias, these modelling tools remain largely unknown to the machine learning and data mining communities until the work of [47] on hierarchical Dirichlet processes and [50], [48] on hierarchical Beta processes. In essence, these works exploit the stick-breaking construction [43], [19] to explicitly represent the prior distribution, and thus make it possible to construct efficient Markov Chain Monte Carlo inference methods. One powerful feature of Bayesian nonparametric models over their parametric counterparts lies in its ability to automatically infer the latent dimension, and more attractively the inference in most cases is no more computational demand than the parametric versions. Our work in this paper utilizes a class of Beta processes as the prior to analyze mood transition patterns, which in turn was built upon the work of [50], [48] on the hierarchical Beta processes. Details shall become clear as we progress in the paper.

To understand the basic patterns in mood transition patterns across social capital cohorts, we appeal to hierarchical factor analysis models. Given the task of exploring the mood transition factors at hand, one does not know in advance how many factors should be specified to the model as typically required in a parametric setting. A possible naïve approach to learn the number of factors is to do model selection using cross-validation. Bayesian nonparametrics is a principled way to do the model selection and avoids the model over-fitting at the same time, and thus will be employed in our work. However, most of existing Bayesian nonparametric factor models (e.g., [36], [57]) use the Gaussian distribution as the data generation mechanism, which is unsuitable for modeling mood swing in our problem as our mood transition is the count data. To overcome this difficulty, we use our recently proposed Bayesian nonparametric factor analysis which uses Poisson distribution for modeling the data and a Gamma distribution for the factors and their scores. Using gamma distribution (or a distribution with non-negative support) allows learning of factors which are parts-based akin to non-negative matrix factorizations [20]. Next we shall briefly review the recent Bayesian nonparametric factor analysis model as the building blocks for our hierarchical model whose details shall be deferred to Section IV-A so that it can be discussed in a more relevant context.

Indian Buffet Process prior: Under a Bayesian nonparametric modelling approach, a popular choice of prior for matrix factorization problems is the Indian buffet process (IBP) is a Bayesian nonparametric prior [10] used to model infinite dimensional binary matrices. Let us assume a binary matrix $Z_{K \times N}$ where $K$ can be infinitely
Table I: Statistics of the connectivity categories used to defined three social capital groups: LOW, MEDIUM and HIGH.

<table>
<thead>
<tr>
<th>Category</th>
<th>LOW</th>
<th>MEDIUM</th>
<th>HIGH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of friends</td>
<td>66,417</td>
<td>70,928</td>
<td>82,059</td>
</tr>
<tr>
<td>Number of groups</td>
<td>70,016</td>
<td>70,688</td>
<td>78,364</td>
</tr>
<tr>
<td>Number of followers</td>
<td>68,421</td>
<td>71,227</td>
<td>79,064</td>
</tr>
<tr>
<td>Number of posts written</td>
<td>60,674</td>
<td>68,703</td>
<td>89,224</td>
</tr>
<tr>
<td>Number of comments made</td>
<td>65,581</td>
<td>70,003</td>
<td>81,296</td>
</tr>
<tr>
<td>Number of comments received</td>
<td>66,055</td>
<td>69,824</td>
<td>80,804</td>
</tr>
</tbody>
</table>

Table III: The number of posts made by users in each social capital groups across six connectivity categories.

large. In modeling applications, $N$ usually denotes the number of data points and $K$ denotes the number of factors or dictionary elements which may be present/absent to represent a data point. IBP has been used in a range of applications (e.g., nonparametric independent component analysis [24] and collaborative filtering [31]). The generative model behind IBP is a beta-Bernoulli process [10]. An extension was developed by [51], which can model integer valued matrices replacing the beta-Bernoulli process with a Gamma-Poisson process. Stick-breaking for IBP was proposed in [49] which was used in our previous work [13], [16]. These models are further described in Section IV-A and are used for our analysis task.

IBP-Based Factor Analysis: The IBP had been widely used in nonparametric matrix factorizations and factor analysis applications where the main focus had been towards modeling the dyadic data using a linear Gaussian model. Consider a typical matrix factorization problem

$$X_{D\times N} = \Phi_{D\times K} H_{K\times N}^T + E$$

where $X_{D\times N}$ is a matrix containing $N$ data points lying in $D$-dim Euclidean space. The matrix $\Phi_{D\times K}$ contains the factors or basis vectors of the transformed subspace (i.e., the subspace spanned by the columns of the matrix $\Phi$, denoted by $S$). The matrix $H_{K\times N}$ contains co-ordinates of the data in $S$ (usually $K < N$). The matrix $E$ denotes the factorization error.

Usually the dimensionality (K) of $S$ is not known a priori and model-selection needs to be performed. The goal of using IBP is to automatically infer $K$ using the data. The factorization of (1) can be re-written as

$$X_{D\times N} = \Phi_{D\times K} (Z_{K\times N} \odot W_{K\times N})^T + E$$

where $H \triangleq Z \odot W$ and $\odot$ denotes element-wise product of two matrices. The matrix $W$ contains the co-ordinates while $Z$ is drawn from IBP and $Z_{ki}$ indicates the presence or absence of $k$-th basis vector in $\Phi$ for $i$-th data point $X_{i,:}$. Similar interpretation is used throughout this paper.

III. CONNECTIVITY, SOCIAL CAPITAL AND MOOD

A. Dataset Statistics

LiveJournal supports one directed person-to-person link type. For a given user, we term incoming links as followers, and outgoing links as friends. LiveJournal also allows users to join communities that discuss topics of interest. From our dataset, on average, each blogger has 10 followers and 23 friends, and joins 6 communities. In addition, only users posting no less than 15 posts and active for more than one year are included in our experiment. The final dataset consists of 1,616,625 users.

We examine social capital defined based on connectivity using one of the following six indicators: numbers of friends, number of community memberships, number of followers, numbers of posts written, number of comments made per day, and number of comments received per day. For each indicator, three social capital subsets are defined: LOW (containing 5000 members above 2.5 percentile), MEDIUM (containing 5000 members spread out from the median, including 2500 members below and 2500 members at or above the 50th percentile cut-off point), and HIGH (contains 5000 members below the 97.5 percentile). This process results in a total of 18 corpora: three social capital groups for each of the six connectivity categories. Table I shows the cut-off statistics used to define for each social capital sub-group. Table III further shows the statistics obtained within each social capital group and its notation to be used throughout the paper. There are 90,000 users, 15000 in each of the social connectivity categories being considered. The number of posts are shown in Table III.

B. Extraction of Primary Moods

Fig. 3: Visualization of 24 primary moods extracted from the data on the core affect circle of emotion structure using valence and arousal proposed in [41].

Livejournal specifies a set of 132 pre-defined moods that can be used to tag a post. Figure 2 shows the histogram of
Intra-cluster distance
Inter-cluster distance

<table>
<thead>
<tr>
<th>Category</th>
<th>Intra-cluster distance</th>
<th>Inter-cluster distance</th>
</tr>
</thead>
<tbody>
<tr>
<td># friends</td>
<td>0.5</td>
<td>933.3</td>
</tr>
<tr>
<td># followers</td>
<td>0.7</td>
<td>536.1</td>
</tr>
<tr>
<td># posts written</td>
<td>0.6</td>
<td>585.1</td>
</tr>
<tr>
<td># comments received</td>
<td>0.6</td>
<td>782.1</td>
</tr>
</tbody>
</table>

Fig. 2: Histogram of moods tagged in a corpus of 18,774,223 posts used in this paper. There is a strong pattern of exponential decay, suggesting the existence of a small subset of primary moods that dominantly explain the data (best viewed in color; pink curve: staircase plot of the true data, red curve: exponential distribution fitting with the mean = 30 after re-numbering the mood label from 1 to 132).

TABLE II: Statistics of the 18 cohorts. Labels of cohorts used throughout paper.

these moods. The plot was created from our corpus of over 18 millions posts, each was tagged by the user. The exponential decay of the histogram suggests that there is a small set of moods that dominantly represent typical discriminative type of emotions across social capital groups. We therefore extract the moods that constitute 50% of the population by placing a threshold over the histogram and lowering it until more than half of the posts are accounted for, yielding a set of 24 moods, termed the primary mood set. Using dimensional representation of the circumplex model of affect for emotion structure proposed in [41], [42], this mood set is visualized to show valence and arousal in Figure 3. It can be seen that these moods are reasonably diverse, including positive (right hand side) and negative (left hand side) sides of the affect circle.

To empirically inspect the usage of primary moods across social capital groups, Figure 4 shows their distribution for LOW, MEDIUM and HIGH social capital groups (correspond to each column from left to right) across 6 connectivity categories (correspond to each row). For each mood, we show the number of posts tagged with that mood, for a particular social connectivity index group. Negative valence moods are marked in blue and positive valence in red, and this is continued through all figures in this paper. One can observe a weak pattern - higher proportions of positive moods (blue) in higher social capital groups.

Another good indicator of whether the primary moods separates the classes is to consider inter- and intra- class distances for each social connectivity categories. Table IV show inter- and intra- class distances in the usage of moods among users in all social capital cohorts. As can be seen, the inter-cluster distance between (LOW vs HIGH) is much greater than the intra-cluster distance for LOW and HIGH, in each social connectivity categories.

3To derive these distances, each user is represented by a count of vector of the number of times he or he has tagged a primary mood, then the usual definition of inter- and intra- distance as in the popular k-mean setting were used.
Fig. 4: Distribution of primary moods for three social capital groups (column-wise from left to right): LOW, MEDIUM, HIGH across six connectivity categories (row-wise, from top to bottom): #friends, #groups, #followers, #posts written, #comments made and #comments received (red indicates negative valence moods and blue positive valence moods)

Fig. 5: Moving from LOW to MEDIUM social capital.
Fig. 6: Moving from MEDIUM to HIGH social capital.

Fig. 7: Moving from LOW to HIGH social capital.
C. Discrimination of Primary Moods Among Social Capital Groups

Next we examine the mood differences across social capital groups. Our intention to understand if there is any pattern of changes in the mood usage as the social capital moving from LOW to HIGH. Figures 5, 6, and 7 shows the change in percentage, between percentage of blogs tagged with a particular mood as social capital moves from LOW to MEDIUM, from MEDIUM to HIGH and from LOW to HIGH cohorts, across six social connectivity measures. It can be seen that the negative valence moods (RED) dominantly decline, whilst the positive valence moods (BLUE) dominantly increase, as one moves from low to higher social capital across all measures of social connectivity.

This is further illustrated in Figure 8(a), which shows the the tag difference moods when moving from the LOW to HIGH cohort, accumulated over across all measures of social connectivity. Decline is shown in black, increase in white. There is a marked decline in negative moods (for example, ‘depressed’) and marked increase in positive moods (for example, ‘amused’). This is further verified quantitatively in Figure 8(b) and (c), showing the valence and arousal differential. It is clear that valence and arousal are greater in the HIGH groups vs the LOW group. The MEDIUM vs LOW shows a similar trend.

IV. PATTERNS OF MOOD TRANSITION ACROSS SOCIAL CAPITAL GROUPS

To further establish the connection between mood and social capital groups, we examine the pattern of users’ mood changes over time within each group. To this end, using the primary mood set constructed earlier, we constructed a $24 \times 24$ matrix of mood transition for each social capital group. Each element at the position $(i,j)$ of the matrix represents the number of times the users within the respecting social group has changed his or her post tag from $i$-th mood to $j$-th mood consecutively. Note that this transition matrix is asymmetric, implying the imbalances among the pattern of mood changes; for example a high frequency of moving from sad to happy implies a different effect of moving from happy to sad.

Figure 9 shows typical mood to mood transitions for LOW, MEDIUM and HIGH social capital groups defined using the number of friends. A column in the matrix implies the tendency of transiting into the corresponding mood from other moods. One may observe weak patterns of transiting into more positive or high-valence moods as the social capital moves from LOW to HIGH. It can be seen that in the group of HIGH social capital, there is a strong tendency for all moods to make transition to “amused”; whereas, this transition is much weaker for the LOW group. Instead a strong transition pattern of moving into the mood “tired” is observed for LOW group.

These weak patterns of mood transitions are also repeatedly observed across six categories of social connectivity. For completeness, similar visualizations of these transition matrices are further supplied and discussed in the Appendix I.

A. Hierarchical Bayesian Nonparametric Analysis of Mood Changes

Our analysis in the previous section suggests that there are indicative patterns in the mood swings among the social capital groups. In particular, HIGH social capital groups tend to transit to more positive, high-valence mood whereas LOW social capital groups exhibit more diverse patterns and tend to use more negative moods. To further distill these patterns quantitatively, we would like to learn the basic set of latent factors that explain these observations. This can generally be treated as a standard dimensionality reduction problem in
machine learning and numerous modelling tools are available. However, since our data comes in the form of matrices, is strictly non-negative and the data type is count, modelling tools such as PCA are known to be ineffective.\(^1\)

One may attempt to approach this from a non-negative matrix factorization approach\(^2\) by concatenating these matrices together and perform NMF on this augmented matrix. This is a reasonable approach, however experimental attempts have shown that it is difficult to pre-specify the number of latent factors. Furthermore, the learned factors were found to be less distinctive across social capital groups. This suggest that we need a stronger modeling approach that should meet three requirements: a) it can automatically discover the number of latent factors, b) it can handle matrices of count data directly and c) being able to explicitly learn the shared and individual factors used across data matrices. To this end, we appeal to Bayesian nonparametric factor analysis, in particular making use of the Restricted Hierarchical Beta Processes (RHBP) developed in our previous works\(^3\),\(^4\).

Let \( J \) be the number of social capital cohorts and denote by \( X_1, \ldots, X_J \) the mood transition matrices.\(^5\) Our goal is to seek a Bayesian nonparametric factor model to jointly model the factorization of these matrices:

\[
\begin{align*}
X_1 &= \Phi H_1 + E_1 \\
&\vdots \\
X_J &= \Phi H_J + E_J
\end{align*}
\]

where \( \Phi = [\phi_1, \ldots, \phi_K] \) the matrix of latent factors to be learnt, whose number of factors \( K \) can grows unboundedly and to be inferred automatically from the data. We represent the coefficient matrix \( H_j \) as an element-wise multiplication \( H_j = Z_j \odot W_j \) where \( Z_j \) is a binary matrix indicating the presence or absence of factors with respect to cohort \( j \)-th; \( W_j \) is the corresponding weight matrix and \( E_j \) is the error matrices independently distributed according to multivariate Gaussian with zero mean. Each latent factor \( \phi_k \in \mathbb{R}^{M \times 1} \) is a

\(^4\)i.e., To provide additional intuition and understanding of the data, we plot those matrices in Figure 9 and Figure 16\(^6\),\(^7\),\(^8\) in the appendix.
Fig. 12: An illustration of factors and their contributions to each social capital cohorts (12 cohorts of LOW and HIGH). The contributions are derived from the matrices $Z_j \odot W_j$. For each factor, we further provide a tag cloud visualization of the primary moods in which their contributions are proportional to the font size. This figure should be read together with Figure 11 to examine the emotional bearing of each factor.

A column vector of size $M = 24$ (the number of primary moods) representing a latent mood swing pattern. Some of the factors in $\Phi$ may be shared amongst social capital cohorts whereas some may be specific to individual cohort. This is modelled implicitly in the model by inspecting the indicator matrix $Z_j$ and will also be learned from the data.

To allow the number of factors grow infinitely, we allow the number of columns in each binary matrices $Z_j$ to grow infinitely and impose a Bayesian nonparametric prior on them and linked them via R-HBP prior, i.e.,

$$B \sim \text{BP}(1, B_0), \quad A_j \sim \text{BP}(a_j, B), \quad Z_{ji} \sim \text{BeP}(A_j)$$

where $\text{BP}$ and $\text{BeP}$ denotes the Beta and Bernoulli processes respectively. $B$ is a draw from a Beta process with the base measure $B_0$ which emanates $J$ children Beta processes $A_j$. Each $A_j$ is used to parameterize a Bernoulli process indicating which factors will be used to generate data. The support of $B$ is the same as that of the base measure $B_0$ and is passed on to each $A_j$, thus latent factors are naturally shared due to this common support and the discreteness of the Beta processes. Finally, to directly model count data we propose to employ a linear Poisson-Gamma model. The whole generative process can be summarized as:

$$\phi_{ik} \sim \text{Gamma}(a, b)$$

$$W_{ji} \sim \prod_{k=1}^{K} \text{Gamma}(a_{w_{jk}}, b_{w_{jk}})$$

$$X_{ji} \sim \text{Poisson}(\Phi(Z_{ji} \odot W_{ji}) + \lambda_j)$$

where $a, b, a_{w_{jk}}, b_{w_{jk}}$ and $\lambda_j$ are hyper-parameters which will be fixed during the inference and $Z_{ji}$ (s) in Eq (6) are sampled according to a hierarchical stick-breaking construction [16], [49]. Inference in this model can be efficiently carried out by using Markov Chain Monte Carlo methods. Typically, it involves iterative sampling of one hidden variable conditionally on the previous samples of other hidden variables until convergence. Technical details are beyond the scope of this paper; however we provide key sampling equations in...
... toудent mood transition pattern. Our model automatically infers a total of \( K = 20 \) latent factors (e.g., matrix \( \Phi \)) from 18 social capital cohorts of LOW, MEDIUM and HIGH.

We now apply our framework of Bayesian nonparametric factor analysis to discover and understand the distinctive factors that underlie mood transitions across social capital cohorts. First, we analyze the mood transition matrices for LOW and HIGH social capital groups, deferring the inclusion of the MEDIUM groups to the next lot of experiment. This results in 12 transition matrices across six social connectivity categories (cf. Table II) for analysis. Our intention is to see whether the model will be able to uncover a strong distinctive mood swing patterns between LOW and HIGH before introducing more complexity into the problem by the inclusion of the MEDIUM groups. As in a standard MCMC inference for Bayesian models, we ran the MCMC chain for 780 times, observing a good convergence behaviour (at after 100 iterations, see Appendix II for more results) and then use the last sample for our analysis. We shall mainly use the estimated factor matrix \( \Phi \) and coefficients \( Z_j \odot W_j \) to interpret the results.

Each estimated factor \( \phi_k \) can be thought of as one basic latent mood transition pattern. Our model automatically infers a total of \( K = 20 \) of such latent factors, each is a column vector of 24 dimensions whose values represent the contribution of primary moods to that factor. These factors are illustrated in Figure 10, each column has been normalized to unity, so that a latent factor is a probability distribution over the primary mood set. A diverse and interesting set of latent factors can be observed. For example factor 14 scatters across all moods, whereas factor 13 strongly expresses the tendency of transit to mood ‘annoyed’; likewise factor 6 to transit to ‘amused’, factor 15 to transit to ‘happy’ and so on. One may further want to quantitatively describe the ‘complexity’ of these mood transition patterns by computing their entropy. Here, we are further interested in describing how well each of these factors, at an aggregate level, expresses the emotion. To do so, we compute the average valence and arousal for each factor by a weighting combination of valence and arousal values of each primary moods, i.e.,

\[
\text{valence}(\phi_j) = \frac{1}{24} \sum_{i=1}^{24} \phi_{ji} \times \text{valence}(m_i)
\]

where \( m_i \) is one of 24 primary moods and valence \( (m_i) \) is its corresponding valence reported in [41]. Arousal for each factor is computed in a similar way, and again Figure 11 visualizes these latent factors on the core affect model in which the font size of each factor is proportional to the its contribution, to explain the data derived by summing the corresponding row of the coefficient matrix \( Z_j \odot W_j \) across all cohort \( j \).

To analyze how latent factors are used across social capital cohorts, we inspect the matrix \( Z_j \odot W_j \) and the value at element \((i,k)\) represents the contribution of factor \( k \) to the \( i \)-th column of cohort data \( j \). For visualization purpose, we sum over all \( i \) to arrive at an overall contribution of each factor \( k \) to cohort \( j \), resulting a matrix of \( 12 \times 21 \) where 12 is the number of social capital cohorts and 21 is the number of factors. Figure 12 illustrates the contribution of estimated latent factors to each social capital cohort. It can be clearly seen that factors HIGH social capital groups use only a subset of factors from 1 to 10, whereas LOW social capital groups use a much more diverse set of factors, scattering across all latent factors estimated from the data. It is even more interesting to note that factors from 1 to 10 being used by HIGH social capital groups are mostly positive (high valence, high arousal) as evident in Figure 11 whereas factors from 12 to 21 uniquely used by the LOW social capital groups are generally negative (low valence).

Whilst these results do not show causality, they establish for the first time, through rigorous analysis and large data sets that high social capital does indeed keeps moods stable in a positive way. Low social capital tend to oscillate between mood extremes.

Next we show results for our Bayesian nonparametric analysis on the mood transition matrices across all 18 social capital cohorts of LOW, MEDIUM and HIGH. Our model learns a total of \( K = 20 \) latent factors shown in Figure 13. Figure 14 visualizes these factors on the core affect model in which the font size of each factor is proportional to the...
its contribution to explain the data derived by summing the corresponding row of the coefficient matrix $Z_j \odot W_j$ across all cohort $j$. Figure [15] illustrates contributions of estimated latent factors to each social capital cohort. Similar results can be seen across the 3 groups: LOW, MEDIUM and HIGH.

V. CONCLUSION

We address the formulation of online social capital, and establish the link between online social capital defined from social connectivity and users’ mood. We show the intuition through quantitative methods, and support the thesis through application of rigorous Bayesian analysis. Social media, indeed, can be a barometer of mood. And, if used wisely, can play an important part in monitoring well-being.

We conclude by noting what Lomas observed profoundly in 1998: “Despite a history in public health dating back to John Snow that underlined the importance of social systems for health, an imbalance has developed in the attention given to generating “social capital” compared to such things as modification of individual’s risk factors. In an illustrative analysis comparing the potential of six progressively less individualised and more community-focused interventions to prevent deaths from heart disease, social support and measures to increase social cohesion fared well against more individual medical care approaches. In the face of such evidence public health professionals and epidemiologists have an ethical and strategic decision concerning the relative effort they give to increasing social cohesion in communities vs expanding access for individuals to traditional public health programs.” [29].

We hope we have established a framework to do just this.

VI. APPENDIX I: MOOD TRANSITION MATRICES ACROSS COHORTS

Mood transition matrices for LOW, MEDIUM and HIGH social capital groups spanning across six connectivity categories (left-to-right, top-to-bottom): #groups joined, #posts written, #comments made (social participation) and #friends,
The Gibbs sampling update for $W_{ji}$ conditioned on the remaining variables is given as

$$p\left(s_{ji}^{K_1+1}, \ldots, s_{ji}^{K_1} \mid \text{rest}\right) \propto \frac{X_{ji}^{K_1}}{\prod_{k=1}^{K_1} \Phi_{jk}^{l} H_{jk}^{l} \lambda_{j}^{s_{ji}^{K_1+1}} \sum_{k=1}^{K_1+1} s_{ji}^{l(k+1)}}$$

where $K_1$ is the current number of active factors (i.e., being used) and the auxiliary variables $s = \{s_{ji}^{l(k)}, \forall j, l\}_{k=1}^{K_1+1}$ are drawn as

$$p(\Phi_{i,:} \mid Z_{1:j}, W_{1:j}, X_{1:j}, \lambda_{1:j}, a, b, s) \propto \prod_{k=1}^{K_1} \left(\Phi_{i,k}\right)^{a+s_j^{il}+1} \exp \left\{ -\left(b + \sum_{j,l} H_{jk}^{l}\right) \Phi_{i,k} \right\}$$

VII. Appendix II: Key Sampling Equations

Under the model described in Section IV-A Gibbs conditional posterior of $\Phi$ can be written as:

$$p(\Phi_{i,:} \mid Z_{1:j}, W_{1:j}, X_{1:j}, \lambda_{1:j}, a, b, s) \propto \prod_{k=1}^{K_1} \left(\Phi_{i,k}\right)^{a+s_j^{il}+1} \exp \left\{ -\left(b + \sum_{j,l} H_{jk}^{l}\right) \Phi_{i,k} \right\}$$

where $K_1$ is the current number of active factors (i.e., being used) and the auxiliary variables $s = \{s_{ji}^{l(k)}, \forall j, l\}_{k=1}^{K_1+1}$ are drawn as

Again, one may observe a weak tendency of transiting into moods of low valence for the users in this LOW social capital group. This is indicated by the high values of columns corresponding to low-valence moods such as ‘tired’ and self-transition of moods such as ‘depressed’ and ‘bored’. In contrast, users in HIGH social capital group have a tendency of transiting into more positive and high-valence moods. These factors are further quantitatively analyzed by our Bayesian nonparametric factor analysis models which suggest an interesting connection between the degree of user’s mood swings and their social capital.

Fig. 16: Mood transition matrices for LOW social capital groups spanning across six connectivity categories (left-to-right, top-to-bottom): #groups joined, #posts written, #comments made (social participation) and #friends, #comments received, #followers.

Fig. 19: Posterior distribution on the number of latent factors (run for 12 cohorts).
where $DZ_{jk}$ denotes a diagonal matrix constructed from $Z_{jk}$.
and the auxiliary variables $t = \{ t_{ip}^{il} \text{ for each } i \}_{p=1}^{K+1}$ are sampled as below

$$
p(W_{jk}^{il} \mid Z_j, \Phi, X_j, \lambda_j, \alpha_{w_j}, b_{w_j}, t) 
\propto \prod_{k=1}^{K+1} (W_{jk}^{il})^{a_{w_j} + \sum_t t_{jk}^{ilk} - 1} 
\times \exp \left\{- \sum_i (\Phi DZ_{jk}^{il})^{ik} \right\}
$$

$$
p(t_{ij}^{il} \mid \text{rest}) 
\propto X_{jl}^{il} \prod_{k=1}^{K+1} \left( \Phi^{il} \H_{jk}^{il} \right)^{t_{jk}^{il}} \lambda_j^{(K+1)}
$$

Gibbs sampling of $\lambda_j$ remains similar to the variables $\Phi$ and $W_j$. The required posterior distributions are given as

$$
p(\lambda_j \mid Z_{1:J}, W_{1:J}, \Phi, X_{1:J}, R)
\propto \lambda_j^{a_{\lambda_j} + \sum_{i=1}^M \sum_{l=1}^{N_j} r_{ij}^{il} - 1} \exp \left\{ - \left( b_{\lambda_j} + MN_j \right) \right\}
$$

For each $i, l$, the auxiliary variables $r_{ij}^{il}$ can be sampled as

$$r_{ij}^{il} \mid \text{rest} = \text{Binomial} \left( X_{jl}^{il} \frac{\lambda_j}{\Phi^{il} \left( H_{jl}^{il} \right)^{1 + \lambda_j}} \right)$$

Figures[19] and[20] show the posterior distribution on the number of latent factors estimated from the experiment on 12 social capital cohorts and likelihood curve which shows a good convergence after 100 Gibbs iteration. Full technical details and derivations can further be found in [16], [14].

**REFERENCES**


Fig. 18: Mood transition matrices for HIGH social capital groups spanning across six connectivity categories (left-to-right, top-to-bottom): #groups joined, #posts written, #comments made (social participation) and #friends, #comments received, followers.


[28] Michael S Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. Content-


Dinh Phung is currently an associate professor in the School of Information Technology, Deakin University, Australia. He obtained a bachelor of computer science with a first class honours and a PhD from Curtin University in 2001 and 2005 respectively. His primary research interest is statistical machine learning, graphical models and their applications in computer vision, multimedia and social computing with an established publication record in these areas. Before joining Deakin, he was the recipient of the Curtin Research Fellowship from Curtin University from 2006 to 2011 and was the recipient of an International Research Fellowship from SRI International in 2005. In 2008, he was invited to Dagstuhl school series on the topic of context modeling and understanding in Germany. He further received the Early Career Research and Development Award and the Curtin Innovation Award from Curtin University in 2010 and 2011 respectively.

Sunil Kumar Gupta is a lecturer at the Center for Pattern Recognition and Data Analytics (PRaDA) of Deakin University, Australia. He received his B.Tech. in Electronics engineering from Harcourt Butler Technological Institute, Kanpur, his M.E. in Signal processing from Indian Institute of Science, Bangalore and his Ph.D. from Curtin University, Australia in 2001, 2008 and 2011 respectively. His research interests include data mining and machine learning with applications in text mining, information retrieval and healthcare analytics.

Thin Nguyen is a member of the Center for Pattern Recognition and Data Analytics (PRaDA) at Deakin University, Australia. He received his PhD from Curtin University, Australia in 2012. His broad research interests lie in large-scale data mining and pattern recognition. In particular, his research work has focused on sentiment analysis, personalisation and crowdsourcing in social media.

Svetha Venkatesh is Professor of Computer Science and Director of Centre for Pattern Recognition and Data Analytics (PRaDA) at Deakin University. She was elected a Fellow of the International Association of Pattern Recognition in 2004 for contributions to formulation and extraction of semantics in multimedia data. She is a Fellow of the Australian Academy of Technological Sciences and Engineering. She is on the editorial board of IEEE Transactions on Multimedia and was on the board of ACM Transactions on Multimedia (2008-2011). She is a program member of several international conferences such as ACM Multimedia. Venkatesh has developed frontier technologies in large scale pattern recognition exemplified through 346 publications including 2 books, 12 book chapters, 78 journal and 254 conference publications. She has won over $12.1 million in competitive research funding since 1992. She has co-authored 9 patents, of which 3 are full patents, and she has co-founded 3 start-up companies.