Generic Multilayer Network Data Analysis with the Fusion of Content and Structure

Xuan-Son Vu[1], Abhishek Santra[2], Sharma Chakravarthy[3], Lili Jiang[1]

[1]Department of Computing Science, Umeå University, Sweden;
[2,3]Information Technology Laboratory, University of Texas at Arlington, USA;
[1]{sonvx, lili.jiang}@cs.umu.se,
[2]abhishek.santra@mavs.uta.edu,[3]sharma@cse.uta.edu

Abstract. Multi-feature data analysis (e.g., on Facebook, LinkedIn) is challenging especially if one wants to do it efficiently and retain the flexibility by choosing features of interest for analysis. Features (e.g., age, gender, relationship, political view etc.) can be explicitly given from datasets, but also can be derived from content (e.g., political view based on Facebook posts). Analysis from multiple perspectives is needed to understand the datasets (or subsets of it) and to infer meaningful knowledge. For example, the influence of age, location, and marital status on political views may need to be inferred separately (or in combination). In this paper, we adapt multilayer network (MLN) analysis, a non-traditional approach, to model the Facebook datasets, integrate content analysis, and conduct analysis, which is driven by a list of desired application based queries. Our experimental analysis shows the flexibility and efficiency of the proposed approach when modeling and analyzing datasets with multiple features.

Keywords: Social network analysis, Multilayer networks, Content analysis

1 Introduction

Analysis of complex datasets containing multiple heterogeneous features such as numeric, categorical, text-based features etc. has become relevant. Social networks (e.g., Facebook, Twitter etc.) fall in the category, and a framework that seamlessly integrate multiple features is desired. In this paper, our goal is to adapt an approach that allows us to efficiently and flexibly analyze social network data using explicit (known or given) as well as implicit (derived or extracted) features of the datasets. It is imperative that these datasets be analyzable in a flexible manner as different features have different impacts and importance on the information that can be inferred. For example, for advertising in social networks, influential communities are sought (based on age, gender, friends, interests, political views etc.). For quick propagation of information centrality nodes may be useful. The approach taken in this paper is generalized although illustrated on a specific real-world, large data collection.

Several approaches are available for analyzing datasets, each with its own advantages and limitations. Database Management systems (DBMSs) have been around for a while and they are good at answering *exact queries*. They are good

for analysis that needs exact answers from the dataset. Statistical, mining, and other approaches have been around, also for quite a while, and they are good for understanding aggregate characteristics – an example being one that is carried out on census datasets to extract inferences and trends over time periods. For datasets that have relationships as part of them, graph modeling and mining became useful and a number of analysis techniques have been developed (e.g., clustering, communities, hubs, triangles,...). As the datasets have become more complex in terms of the number of features, modeling them as a single graph makes the graph complex – with multiple edges between nodes corresponding to different relationships, multiple node & edge labels, etc. Modeling becomes complex and applying traditional algorithms for flexible analysis is either impossible or more complex and inefficient. We term this as the *traditional approach.*

The approach used in this paper, termed **M**ulti**L**ayer **N**etwork (MLN) analysis with decoupling, is in its early stages and being researched actively. The MLN approach does not change the analysis, *except how* datasets are modeled and analyzed. It has been receiving a lot of attentions in the last decade due to its advantages: i) *allows modeling of a complex dataset using a set of user-definable simple, single graphs termed layers)* , ii) *allows the same analysis as the traditional approach on this model without loss of accuracy*, iii) is amenable to parallelism (for scalability) and has been shown to be better in storage requirements and efficiency. There are other advantages as well [27, 16].

In this paper, we plan to showcase the above advantages of MLN approach. This paper is the first one, to the best of our knowledge, to apply this approach for the analysis of *one of the largest/densest real-world social network data collection*, although it has been used in several experimental studies on smaller/sparser datasets [6, 4].

The contributions of this paper include: **(1)** using a novel, emerging MLN approach for flexible analysis of a large complex real-world dataset, **(2)** establishing its modeling benefits, flexibility of analysis, and efficiency of computation, **(3)** integrating content analysis seamlessly with structural network analysis, and **(4)** extensive analysis and result validation for the social network work datasets.

The remainder of the paper is organized as follows. Section 2 discusses related work. Section 3 states the general problem and proposed approach. Section 4 elaborates modeling and computation aspects of our approach. Section 5 details the use of content analysis to integrate into multilayer network approach. Section 6 showcases analytical results of queries. Section 7 discusses computational advantage of the adapted approach. Section 8 concludes the paper.

2   Related Work

We provide an overview of some related work including multilayer networks, community detection, and content-based analysis.
Multilayer Networks or Multiplexes: Significant amount of work has been done in the area of multilayer networks [16, 3]. They have been categorized into homogeneous, heterogeneous, and hybrid multiplexes depending on whether the layers have same, different, or a combination of entity sets, respectively [26]. In

this paper, we focus on homogeneous multiplexes where the entity set (i.e., people in our case) is fixed, but the features are numerous giving rise to a layer for each feature. Considerable research has been done to handle varying interactions among the same set of entities such as co-authorship in different conferences [4] and city connectivity based on different airlines [6]. In order to understand the effect of multiple features using composition of multiplex layers, a principled approach [27, 28] has been proposed to arbitrarily combine features (or layers) and then analyze them without having to construct combined layers. The composition approach leads to efficient (both computation and storage) and flexible analysis of combinations of features using Boolean operations.

Community Detection is a well-studied problem in monoplex (single network) analysis. It involves identifying groups of nodes that are more connected to each other than to other nodes in the network. Some work has investigated community detection methods for multilayer networks [13, 20]. But there is hardly any work for determining the communities for different multiplex layer compositions, until recently where a novel approach was proposed to infer communities of a combined network from communities of individual layers using Boolean compositions [27]. We posit in this paper that this approach is appropriate for analyzing multi-feature datasets flexibly and efficiently.

Content Analysis is a research method for studying documents and communication artifacts, which might be texts of various formats, pictures, audio or video [22]. One of the key advantages of using content analysis to analyze social phenomena is its non-invasive nature, in contrast to simulating social experiences or collecting survey answers. In this paper, we apply content analysis by text mining techniques to integrate content features into multilayer structural analysis. Motivated by the fact that, we have personality data in the given Facebook data collection and the previously preliminary studies [29, 32] observed a significant correlation between personality and privacy-concern of Facebook users, we improve [32]'s deep neural network model in this paper to detect Facebook users' privacy-concern based on their status updates and classify privacy-concern into three different levels (classes): *high, medium*, and *low*. These content-derived classes are treated as content features and modeled as layers to integrate into the multilayer structural analysis.

3   Datasets, Analytical Queries, and Problem Statement

3.1   Datasets

For this paper, we chose the Facebook (FB) data collection to showcases the power and flexibility of the MLN approach, since the data collection fulfill all the characteristics of a complex dataset in terms of a large number of features, content that could be analyzed and the requirement for analysis using combinations of features.

This data collection from *myPersonality* project [17] is one of the largest and well-known real-world social network research data collection, where the volunteers took real psychometric tests and opted in to share data from their

FB profile (period of 2007-2012). The experimental data contains four datasets: Demographic Info (D1), User's Political Views (D2), Personality (D3), and FB Status Updates (D4). We have a total of 260K individuals in the datasets predominantly from USA. Of those, about 2.6K have more common features as compared to others. Hence, as shown in Table 1, we have chosen the 2.6K subset for detailed analysis.

We use four features from dataset D1 including age, gender, relationship status, and locale. One self-declared political-view feature from D2. D3 provides five features which are the five personality traits of the Five-Factor Model (*FFM*) [18]. *FFM* is considered

Table 1: Statistics of four datasets

Datasets	#users
Demographic Info (D1)	2,676
User's Political Views (D2)	2,695
Personality (D3)	2,485
Facebook Status Updates (D4)	1,645

the most influential and standard model for personality trait prediction in psychology over the last 50 years. Based on D3 together with D4, one more feature of people's privacy-concern is inferred. For better understanding in later sections, we especially introduce the well-known five personality traits [9], which are defined as openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism.

Openness (OPN) to experience: intellectual, insightful vs. shallow
Conscientiousness (CON): self-disciplined, organized vs. careless
Extraversion (EXT): sociable, playful vs. aloof, shy
Agreeableness (AGR): friendly, cooperative vs. antagonistic, faultfinding
Neuroticism (NEU): insecure, anxious vs. calm, unemotional

3.2   Analytical Queries

The focus of this paper is to demonstrate modeling, flexible and efficient analysis of complex datasets to infer trends and establish correlations, and possible causality. A few analytical queries that are meaningful for this dataset are shown below.

(Q1) **Dominant Political Views:** How the user-declared political view (e.g., democrat, doesn't care, republican) varies across age groups in the dataset? (There are 100 political views in Facebook dataset used in this paper).

(Q2) **Relationship Status Correlation:**
  (a) With respect to age groups, how does relationship status (e.g., single, in a relationship, and married) vary?
  (b) How do the relationship statuses affect the personality traits of an individual? Does it differ based on gender?

(Q3) **Personality Trait Analysis:**
  (a) How much of the population demonstrate contrasting personality traits (e.g., OPN and NEU)?
  (b) How do the personality traits evolve with age? For example, which age group of people deals better with stress?

(Q4) **Privacy Concern Correlation:**
  - (a) How does the individuals' age correlate with their comfort level of sharing personal information on social media?
  - (b) Do personality traits have a bearing on the level of privacy-concern?

3.3   Problem Statement

Currently, the above analytical queries are done using a graph-based approach where *a single graph needs to be created for each analysis based on the involved features*. Typically, nodes represent entities (i.e., people in our case) and edges represent relationship between nodes based on feature values (e.g., same age group, the same relationship status). These graphs are analyzed using graph metrics such as community, hubs or centrality nodes, and so on. This approach entails the creation of a customized graph for each query using the features involved which can lead to an exponential number of graphs in the worst case. For the above mentioned analytical queries (Q1-Q4), multiple graphs, the number of which depends on the number of features involved, need to be created, stored, and analyzed for each query.

Using the MLN approach, *for a given dataset with M features (whether explicit or derived), determine the layers based on the analyses requirements of the dataset and use composition for analysis using Boolean or other operations.* Thus, once multiplex layers are created as shown in this paper, any number of analyses can be performed without generating additional graphs/layers. Necessitated by the queries, this paper primarily uses *AND* compositions.

4   Data Modeling and Layer Composition Using MLN

4.1   Data Modeling

It is already explained in Section 3 that we have four Facebook datasets (D1-D4) and the corresponding features. *UserID* is common to all datasets to associate with each other as shown in Figure 1. *UserID* is also used to make the dataset anonymous to mask the privacy of an individual.

Unlike other features from D1 and D2, the features of personality traits (OPN, CON, EXT, AGR, NEU) and privacy-concern are not explicitly present in the given dataset and will be derived through content analysis. The derived output for each personality trait would be binary label "Yes" or "No". And the
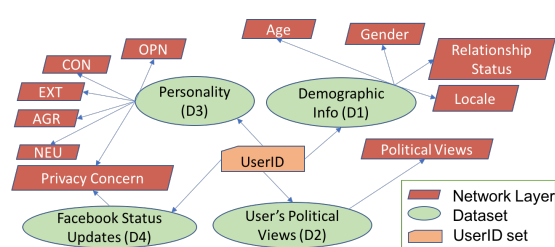

Fig. 1: Modeling the FB Data Collection for Multilayer Network Analysis

derived output for privacy-concern of each FB user in our datasets is "high", "medium", or "low". This clearly indicates the power of the approach in absorbing derived content in the same way as an explicit feature. Different types of

content extraction can be supported readily. Our approach to content extraction is presented in Section 5.

Table 2 shows the statistics regarding the generated multiplex layers along with the number of edges in each layer. The number of nodes in all layers is the same but the number of edges will be different since it depends on the available information for each feature. The 11 layers in Table 2 that are generated for the Facebook multilayer network correspond to the features in Figure 1. The semantics of the graphs in each layer are described as follows:

(L1) **Age**: Any two users are connected by an edge when they both fall into a same age group, namely [≤ 20], [21-30], [31-40], [41-50], [51-60], and [≥ 61].
(L2) **Gender**: Any two users with the same gender are connected.
(L3) **Relationship Status**: Any two users with the same relationship status are connected by an edge.
(L4) **User-Defined Political Views**: Any two users with the same political view are connected by an edge.
(L5) **Locale**: Any two users with the same locale settings (e.g., en_UK, en_US) are connected by an edge.
(L6-L10) ***FFM* (i.e., OPN, CON, EXT, AGR, NEU)**: Each personality trait of the *FFM* forms one network layer. Any two users with the same type of personality trait are connected.
(L11) **Privacy Concern**: Any two users with the same privacy-concern level (i.e., high, medium, or low) are connected.

4.2   Metric Computation & Layer Composition

Table 2: Statistics of 11 multiplex layers

Layer	From dataset	# edges
L1: Age	D1	1,228,223
L2: Gender	D1	1,813,638
L3: Relationship Status	D1	1,119,592
L4: Political Views	D2	494,974
L5: Locale	D1	2,799,160
L6: OPN	D3	1,020,306
L7: CON	D3	840,456
L8: EXT	D3	795,691
L9: AGR	D3	718,201
L10: NEU	D3	627,760
L11: Privacy Concern	D3, D4	2,191,659

In the multilayer network described above, although each layer has the same nodes, their edge connectivity will vary according to the feature value distribution. For example, groups of people having the same political view may not be present in the same age group, groups of people with the same personality may have different levels of privacy-concern and so on. For detecting the tightly connected groups of people with respect to a particular feature, we compute *communities* in the corresponding layer by applying Infomap[5]. A community in a graph translates to a group of nodes that are more connected to each other than to other nodes/communities in the graph.

Analytical queries listed in Section 3.2, require commonality of information in *at least two multiplex layers*. This corresponds to the Boolean *AND* operation for composing layers. For example, in Q4a (Section 3.2), to analyze the effect of age on level of privacy-concern, we need to compute communities where an edge
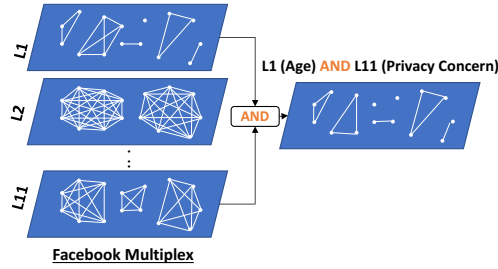
Fig. 2: An example of AND-Composition

represents people who fall into the same age group **and** have the same privacy-concern level. Figure 2 shows a simple example with several small Facebook multiplex layers, and an ***AND*** composition of L1 (Age) and L11 (Privacy Concern) layers. Table 3 shows the *AND*-compositions whose *communities* have to be generated in order to perform the flexible analysis of the Facebook multilayer network for our analytical queries listed in Section 3.2.

Table 3: *AND*-Compositions needed for analytical queries shown in Section 3.2

Ana-lysis	Required ***AND***-Compositions of Layers
Dominant Political Views	
Q1	L1 (Age), L4 (Political View), L5 (Locale)
Relationship Status Correlation	
Q2a	L1 (Age), L3 (Relationship Status), L5 (Locale)
Q2b	Five 3-layer compositions: [L2 (Gender) AND L3 (Relationship Status)] with each of L6 (OPN), L7 (CON), L8 (EXT), L9 (AGR), L10 (NEU)
Personality Traits Analysis	
Q3a	L6 (OPN), L10 (NEU)
Q3b	Five 2-layer compositions: Each of L6 (OPN), L7 (CON), L8 (EXT), L9 (AGR), L10 (NEU) with L1 (Age)
Privacy Concern Correlation	
Q4a	L1 (Age), L11 (Privacy Concern)
Q4b	Five 3-layer compositions: [L2 (Gender) AND L11 (Privacy Concern)] with each of L6 (OPN), L7 (CON), L8 (EXT), L9 (AGR), L10 (NEU)

Community Detection for *AND* Compositions: The traditional way to address any of the listed analytical queries is to first generate the required combined graph corresponding to the *AND*-composition and then detect the communities. The MLN approach pre-computes the communities of the individual layers. Based on the composition requirements of the analytical query, partial results (i.e., pre-computed communities) are intersected (for *AND*-composition) to generate the combined communities. It can be shown analytically [27] that both computationally and storage-wise, the MLN approach is more efficient. The requirement to apply composition is that the communities of the individual layers must be *self-preserving* in nature.

Checking the Self-Preserving Property: By definition, *"a community is self-preserving if the nodes in it are so tightly connected such that even if only a subset of connected nodes are chosen from that community, they will form a smaller community [27]."* For each of the 11 layers, we computed the internal clustering coefficients of each node, which is the ratio of the number of edges among the neighbors of the node *in the same community* to the total possible edges among those neighbors, and it was *one*. This indicated that the *communi-*

ties in each layer were self-preserving. In Section 6, we will discuss some of the inferences that have been drawn from this analysis, and in Section 7 we highlight the performance analysis between the traditional and MLN approaches.

5   Content Analysis on User Generated Content

As briefly introduced in Section 4, we apply content analysis to derive the features of five personality traits and privacy-concern to adapt into the proposed multilayer network (MLN) approach. For detecting personality traits, [23] has identified many linguistic features associated with each personality trait in *FFM*. For instance, *Extraversion* (EXT) tends to seek stimulation in the external world, the company of others, and to express positive emotions. *Neuroticism* (NEU) people use more 1st person singular pronouns, more negative emotion words than positive emotion words. As mentioned earlier in Section 4.1, five personality trait scores for each user are given in the dataset D3. To form layer L6-L10 using these personality traits in the proposed multilayer network analysis, we followed the strategy from [8] by using five mean-values {3.8, 3.5, 3.6, 3.55, 2.8} of the given personality scores to decide "Yes" or "No" label of {OPN, CON, EXT, AGR, NEU} correspondingly. "Yes" label is assigned if the personality trait score is higher than or equal to the corresponding mean-value. Otherwise, "No" label is assigned. In the following of this section, we will mainly explain how to generate privacy-concern as one layer to adapt the multilayer network.

Previous work [32] has found that using given personality and status updates of users, privacy-concern can be predicted accurately based on UGC data. We extended their model by building a deep neural network model to automatically classify users' privacy-concern to high (HiPC), medium (MePC), and low (LoPC) based on given five personality trait scores and status updates. Motivated from previous studies [23, 32, 30], we extract the following content features from Facebook status updates, which will be used in the following deep neural network model to predict users' privacy-concern:

– **Polarity features:** Since sentiment words embody personality (e.g., *"I don't hate you. Well, you found me · · · "*), we use the number of polarity signals appearing in FB status updates as the polarity features. We identify positive and negative words using the sentiment dictionaries provided by Hu and Liu [11]. Additionally, we consider boolean features to check whether or not a negation word is in a FB status (e.g., n't).
– **Syntactic features:** We extract part-of-speech tags (POS tags) for all status updates in the dataset. Afterwards, we use all the POS tags with their corresponding term-frequency and inverse-document-frequency (tf-idf) values as our syntactic features and feature values, respectively.
– **Semantic features:** The major challenges when dealing with user generated data are: (1) the lexicon used in a status update is informal with many out-of-vocabulary words and (2) they are usually short text [31]. The lexical and syntactic features seem not to capture that property very well. To handle this challenge, we apply two approaches to compute vector representations for

FB status updates. First, we utilize Latent Dirichlet Allocation (LDA) [2] for discovering the abstract "topics" that occur in all FB status updates. Secondly, we employ 300-dimensional pre-trained embedding models at word level [19] and at character level [14] to compute a representation for a FB status update as the average of the embeddings of words and characters in the status update.

– **Lexical features:** include [1-5]-grams in both word and character levels. For each type of n-gram, we only select the top 1,000 n-grams based on *tf-idf*.

Deep Neural Network Model: The recent novel model from [32] was adapted to find users' privacy-concern level based on their social network status updates or personality-trait score. Their neural network model was proposed mainly for privacy-degree prediction, instead of privacy-concern level prediction. To adapt the proposed multilayer network, there is a need of categorized output to create a privacy-concern network layer (L11), where users with the same privacy-concern level is supposed to be connected. Thus, we extended their approach by developing a category-based privacy-concern detection model as shown in Figure 3. This extension does not only matter to the output (from discrete to categorization), but the feature representations in hidden layer of deep neural network model are different.

It is a Multilayer Perceptron (MLP) model [10], the architecture of which consists of an input layer, two hidden layers and a softmax output layer. Given all Facebook status updates of a user, the input layer represents the status update by a feature vector which concatenates lexical, syntactic, semantic and polarity feature representations. The two hidden layers with ReLU activation function [21] take the input feature vector to select the most

Fig. 3: Neural network model for privacy-concern detection

important features which are then fed into the softmax output layer for privacy-concern level detection and classification. Regarding classification performance evaluation, we split 20% of the data for a blind test. We run 10 fold cross-validation on the rest 80% to train and select the best hyper-parameters. After all, the model achieved an accuracy of 84.44% on the blind test set. Furthermore, we compare the model with other popular models (i.e., support vector machine, random forest) and the recent advanced model (i.e., C-LSTM [35]), and none of them performs as good performance as our model does. This clearly shows the effectiveness of the model to predict privacy-concern levels based on UGC data. As explained above, the predicted user privacy-concern levels are used to create the network layer L11 of the multilayer network.

6   Experimental Study and Query Results Analysis

This section shows the results for each analytical queries (Q1-Q4) formulated in Section 3.2. Based on the communities obtained for the required *AND*-Compositions listed in Table 3, a detailed analysis was performed to draw some insights that are discussed below. To the extent possible, we have related our analysis with various published independent surveys.

Dominant Political Views (Q1): Figure 4 shows the distribution of the top three political views over the US population active on Facebook for each age group. Some observations are:

– Among the politically interested and socially active US people across age groups, the majority supported the democrats in the period of 2007-2012. As we know, there was a lot of support for the democratic presidential candidate who got elected in 2008, and we believe this is reflected in the political views of that period. The period of 2007-2009 also indicated the same which makes sense as the campaign was underway in that period. This is confirmed in [34, 33] since Barack Obama, a democrat, who took the US presidency on January 8, 2009 was able to influence people's political leanings through his presidential campaign from February 10, 2007.

– Among the socially active youth ($\leq$ 30 years old), majority of them have the political view of "doesn't care". Although this includes people below the voting age, even the published statistics [12] show that young people are least likely to vote and may not have formed an opinion about their political leanings.

– Among politically interested youth ($\leq$ 30 years old) who are also active on social media, dominant support is for democrats. This may be attributed to a few of president Obama's youth centric movements and the significant use of social media for the first time in a US election and also to his subsequent accomplishments [15].

– Interestingly, among all age groups, only the ones above 61 years old favored republicans over democrats, which is also reflected in the election reports from 2008 [24].
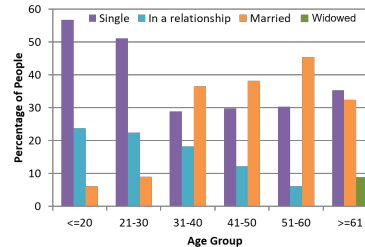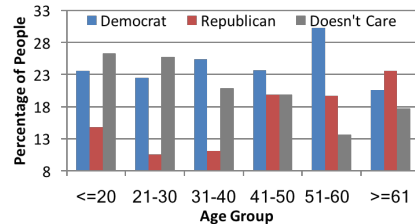

Fig. 4: Top 3 political views by age group

Fig. 5: Top 3 relationship statuses by age group

Relationship Status Correlation (Q2): The preference of a relationship status based on age and the corresponding effect on personalities of different genders were analyzed.

(Q2a) *Variation with Age:* Figure 5 shows, for each age group, the distribution of people among the group's top three relationship statuses. A few intuitive inferences that can be drawn are:

- The youth (≤ 30 years old) stay single than be in a relationship or get married, according to the given dataset.
- The percentage of married people steadily increases with age which can be attributed to the popular fact that as people age, they want to be in a longer term commitment (in a relationship or married).
- The transitions from "Single" to "In a relationship" to "Married" are clearly seen with change in age in Figure 5 which matches the social trend.
- The third largest fraction of people in age group ($\geqslant 61$) constitutes those who have lost their spouses (Widowed).

(Q2b) *Effect on Personality and Gender:* Changes in relationship status seem to have effects on the personality. Moreover, this change seems to be correlated with the gender. For the given population, we present the distribution of males and females among the top three relationship statuses - Single (S), In a relationship (R), Married (M), who display different personality traits in Table 4. The ones marked in **bold*** and *italics*† represent the category of people with highest and lowest percentages, respectively.

Table 4: % of people with different personality traits based on relationship status and gender

Trait	Males			Females		
	S(%)	R(%)	M(%)	S(%)	R(%)	M(%)
OPN	**55.7***	54.8	47.3	53.0	55.6	*43.0*†
CON	*44.2*†	50.2	52.7	46.1	51.7	**58.5***
EXT	45.0	**54.8***	*35.5*†	49.1	50.8	44.6
AGR	42.9	50.2	*31.8*†	46.4	45.9	**50.4***
NEU	36.5	*26.6*†	33.6	45.2	**53.8***	48.1

A few observations that can be made from the above analysis are: a) Married females have least openness (OPN) to experience, and highest conscientiousness (CON). b) Married females have the highest agreeableness (AGR), while married males have the least. These observations have been made with respect to the given population and need to be confirmed on a bigger population.

Personality Traits Analysis (Q3): Various analyses based on the five personality traits for each individual are discussed below.

(Q3a) *Contrasting personality traits:* Clearly, if a person feels anxious and does not stay relaxed (NEU) then he/she will try to make his/her life comfortable by indulging in less stressful activities making them be less open to new experiences (OPN). Thus, the fraction of people displaying these contrasting personality traits is supposed to be low. Our analytic results go hand in hand with this intuition as just 23.3% people belonging to this category.

(Q3b) *Personality trait evolution with Age:* Figure 6 shows how each personality trait varies with age, based on which, few interesting observations can be made as follows.

− Openness (OPN) reflects whether one prefers new experiences and to engage in self-examination. This trait increases with age and peaks around the 30s (54.2% in age group of 31-40). However, older people prefer to go with the
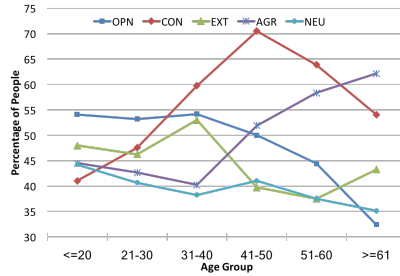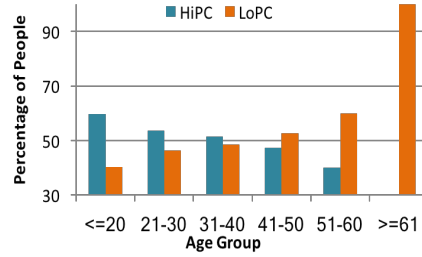
Fig. 6: Changing personality distribution with age

Fig. 7: Distribution of HiPC and LoPC by age group

tried-and-tested approach (67.6% of the people above 60 years old resist new experiences).

– Conscientiousness (CON) associates with achievement and working systematically, methodically and purposefully. Analysis shows that the age group with conscientiousness the most is 41-50 years old. A recent survey about founders and entrepreneurs indicated that their average age was 45 years old [1].

– Extraversion (EXT) describes one's sociability and enjoy to be the center of attention. This trait seems to peak at two age groups (i.e., [31,40] and [$\geqslant 61$]) in the dataset.

– Agreeableness (AGR) reflects a tendency to perceive others in a more positive light. Parenthood and grand-parenthood may make the elder generation more empathetic towards others as compared to the younger lot.

– Neuroticism (NEU) reflects one's ability to deal with emotion states, such as stress and anxiety. It can be observed from Figure 6 that the younger lot does not deal very well with stress. Even the studies substantiate this finding as around 80-90% adolescent suicides are linked to common psychiatric disorders, such as depression and anxiety [7]. This trait (NEU) seems to be most stable over age compared to other traits.

Privacy Concern Correlation (Q4): Three levels of privacy concern (PC) have been considered for this work - HiPC, MePC and LoPC. Here we have taken into account age, gender, and personalities as three parameters for performing analysis to understand the choice of particular level of privacy-concern.

(**Q4a**) *Variation of privacy-concern across age groups:* The concern level of sharing personal information on the social media varies with age. Irrespective of the age group, the MePC was the most dominant level of privacy. Out of the remaining individuals, Figure 7 shows the distribution of the people with extreme levels of privacy - High and Low. Few observations are discussed below.

– People (≤ 40 years old) prefer the higher level of privacy. This can be attributed to the fact that this age group is probably more aware of the cons of sharing sensitive personal information on the web such as identity theft.

– The status updates of people ($\geqslant 41$ years old) contain more personal information and this trend increases with age. This reflects a lower level of privacy-concern probably due to their unawareness of the potential harm from disseminating personal information on social media.

(**Q4b**) *Correlation of privacy-concern over personality traits:* Table 5 shows the two extreme personality traits and their corresponding privacy-concerns for males and females.

Table 5: Dominant personality traits of male and females preferring different levels of privacy-concern

Privacy	Extraversion(100%)		Neuroticism(100%)	
	Males(%)	Females(%)	Males(%)	Females(%)
HiPC	0	0	09.90	**09.99**
MePC	39.89	**45.36**	30.15	**49.96**
LoPC	07.45	07.30	0	0

− Females have higher privacy-concern than males on both extraversion and neuroticism. This observation kept consistency with the previous study in [25].
 − Both males and females on extraversion display low and some medium levels of privacy-concern. This matches the definition of extraversion from social scientists. That is, people who enjoy being the center of attention are likely to share more personal information on the web such as check-ins and day-to-day activity updates.
 − Both males and females on neuroticism tend to have predominantly high privacy-concern, without anyone having low level privacy-concern. This matches the social scientists' definition.
 Analysis Q4a and Q4b are important as it validates our content extraction approach to derive accurate privacy-concern.

7   Efficiency Analysis of The MLN Approach

So far, we have established the modeling and flexibility of analysis of the MLN approach. Below, we will establish its efficiency in general and highlight it with respect to the current dataset analysis.
Processing Layers Instead of a Single (large) Graph (SLG): Separation into layers allows one to process each layer *only once* for *all analysis* and the composition allows us to make use of the pre-computed partial results. Furthermore, in many cases, the size of each layer is likely to be smaller than the size of combinations of layers. On the other hand, a new combined graph needs to be created and processed in the traditional approach for each *unique* analytical query. Storing each layer is more compact and uses less memory than storing all the required layer combinations.
Processing Layers in Parallel: The MLN approach easily lends itself to process all (or subsets of) layers in parallel to further improve efficiency. The total cost is the processing cost of the most complex layer. This can only be done for a set of known analysis queries in the traditional approach in contrast to the MLN approach where it needs to be done only once. In the MLN approach, it is further possible to process each layer in parallel by partitioning and leveraging existing algorithms (again one time). Although this can be done in the traditional

approach, it has to be done *after* the combined graph is created which reduces its effect significantly.

Efficiency of Composition: The core of the MLN approach is its ability to compose layers pair-wise to get complete, correct results. Each composition is likely to be on fewer and smaller number of components (from each layer) thereby reducing the resources needed (both storage and processing).

Combinatorial Reduction: As the complexity of dataset increases, it translates to more layers (corresponding to more features). Based on the number of layers, in the MLN approach, there is a significant and non-linear reduction in the processing cost as the number of layers increase as compared to the traditional approach. Assuming M layers, for an exhaustive analysis, they need to be combined in $2^M - 1$ ways, each representing a unique analysis of a combination of features. This translates to creating $2^M - 1$ combined graphs in the traditional approach and processing them individually. In the MLN approach, instead, M layers are processed *once* and $2^M - 1$ combining of partial results from layers are performed. As we show below, these compositions are significantly smaller (by orders of magnitude) computationally as compared to processing of a layer. Essentially, the exponential complexity has been reduced to a linear one one with very little additional processing.

7.1   Complexity Analysis

For the complexity analysis, we assume that for a given multilayer network with fixed number of M layers, say $\{G_1, \cdots, G_M\}$, each of the N query analyses requiring K related layers on average, should return a list of communities, L.

– Single (large) Graph (SLG): In this approach, for every analysis it first generates the composed graph, G_{AND}, obtained through (K-1) 2-layer *AND*-compositions, on average. On this *AND*-composed layer, we apply the Infomap technique (InfoM) [5] to generate the list of communities, L. Thus, for N analyses the complexity of this approach will be $O(N * (AND_{i=1}^{K} G_i + \text{InfoM}(G_{AND}))$, where $K \leq M$.

– MLN: Its first step is to generate the communities for each of the M layers by applying Infomap. While generating the communities we also obtain the internal clustering coefficient for each node, which are used to determine if the communities are *self-preserving* or not. If the property of *self-preserving* is satisfied, then for each analysis, to generate the list of communities L, for the corresponding *AND*-composition G_{AND}, communities from K layers are intersected based on nodes. Thus, for N analyses the cost of this approach will be, $O(\sum_{i=1}^{M}(\text{InfoM}(G_i)) + N * \cap_{j=l}^{K-l+1} C_j)$, where $K \leq M$.

In terms of storage space, one needs to store M lists of communities for MLN, whereas in SLG, $N * (AND_{i=1}^{K} G_i)$ graphs need to be stored. Considering both space and time, if the number of analyses, N, and the average number of layers required for each analysis, K, are low then the one-time cost of performing M number of Infomap operations in MLN will dominate and make MLN more expensive as compared to SLG. However, with the increasing values of N and

K, the efficiency of MLN over SLG improves significantly, as the cost of producing the number of AND-composed graphs and applying Infomap that traverses through the edges of each of them, begins dominating. In the worst-case for $N = O(2^M)$ and $K = O(M)$, SLG will perform an exponential number of *AND*-Compositions and edge traversal based Infomap operations whereas, MLN will just perform the cost-effective node intersection of layer-wise communities.

7.2   Computational Results

Below, we show efficiency results of analysis on the given Facebook datasets.

Experimental Set up: We used a quad-core 8th generation Intel i7 processor Linux machine with 8 GB memory for all of our analysis. Based on Table 3, we computed communities for a total of 19 *AND*-compositions to answer the queries Q1 to Q4, each requiring 3 layers on an average.
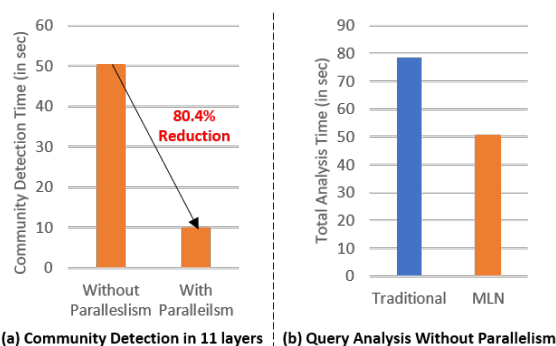

(a) Community Detection in 11 layers ┊ (b) Query Analysis Without Parallelism

Fig. 8: Efficiency Comparison of MLN and Traditional Approaches

Figure 8 (a) shows the time taken for processing all of 11 layers *with and without parallelism*. As can be seen, in the MLN approach *with parallelism*, it reduces the cost of processing the most complex layer (9.847 seconds for L5: Locale, 2.8 million edges, Density: 0.77 - *most dense*) – a ***reduction of 80.4%*** *approximately.*

The incremental computation cost for each query using the MLN approach is extremely small. This can be appreciated from the worst case scenario - comparing minimum layer processing cost with maximum composition cost. The total composition cost to answer the most complex query (Q2b) was **0.039 seconds** and the minimum layer processing cost was **1.61 seconds** (L4: Political View, 494K edges, Density: 0.14 - *least dense*). **The difference is more than two orders of magnitude.**

Figure 8 (b) shows the *total time* taken to answer the analysis queries using the traditional and the MLN approach, respectively, *without parallelism*, as 78.520 seconds and 50.354 seconds (for **36% reduction**). Further, if communities for each layer are generated in parallel, total computation time for the MLN approach reduces to 9.987 seconds (for **87% Reduction**). Also, note that the analysis shown in this paper is **less than 1% of the possible analysis**.

In summary, the experiments on the Facebook Dataset validate the *MLN approach from an efficiency perspective as compared to the traditional approach.*

8   Conclusions

In this paper, we have applied the emerging MLN approach model and analyze a social network data collection in a flexible and efficient way. We have also shown how content analysis can be readily incorporated into the proposed MLN approach. Experimental analysis and evaluation not only demonstrate the flexibility and efficiency of data analysis using the MLN approach but also validate the analysis results. For future study, we plan to (1) apply the proposed MLN approach to the bigger full data collection and (2) apply hypothesis testing on two different data distributions (e.g., the current one versus the full data collection) to see the statistical significant degree of our findings.

Acknowledgement

References

1. Azoulay, P., Jones, B., Kim, J.D., Miranda, J.: Age and high-growth entrepreneurship. Tech. rep., National Bureau of Economic Research (2018)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
3. Boccaletti, S., Bianconi, G., Criado, R., del Genio, C., Gmez-Gardees, J., Romance, M., Sendia-Nadal, I., Wang, Z., Zanin, M.: The structure and dynamics of multilayer networks. Physics Reports **544**(1), 1 – 122 (2014)
4. Boden, B., Gnnemann, S., Hoffmann, H., Seidl, T.: Mining coherent subgraphs in multi-layer graphs with edge labels. In: Proc. of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 1258–1266 (2012)
5. Bohlin, L., Edler, D., Lancichinei, A., Rosvall, M.: Community detection and visualization of networks with the map equation framework (2014), `http://www.mapequation.org/assets/publications/mapequationtutorial.pdf`
6. Cardillo, A., Gómez-Gardenes, J., Zanin, M., Romance, M., Papo, D., Del Pozo, F., Boccaletti, S.: Emergence of network features from multiplexity. Scientific reports **3** (2013)
7. Cash, S.J., Bridge, J.A.: Epidemiology of youth suicide and suicidal behavior. Current opinion in pediatrics **21**(5), 613 (2009)
8. Celli, F., Pianesi, F., Stillwell, D.S., Kosinski: Workshop on computational personality recognition (shared task). The Seventh International AAAI Conference on Weblogs and Social Media (2013)
9. Costa, P. T., J., McCrae, R.R.: The Revised NEO Personality Inventory (NEO-PI-R), pp. 179–198 (2008)
10. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Netw. **2**(5), 359–366 (1989)
11. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 168–177 (2004)

12. Jordan Fabian: Poll: Young people (still) least likely to vote (2013), `https://splinternews.com/poll-young-people-still-least-likely-to-vote-1793839910`
13. Kim, J., Lee, J.: Community detection in multi-layer graphs: A survey. SIGMOD Record **44**(3), 37–48 (2015)
14. Kim, Y., Jernite, Y., Sontag, D., Rush, A.M.: Character-aware neural language models. pp. 2741–2749. AAAI'16 (2016)
15. Kimberly Amadeo: What has obama done? 13 major accomplishments (2018), `https://www.thebalance.com/what-has-obama-done-11-major-accomplishments-3306158`
16. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., Porter, M.A.: Multilayer networks. CoRR **abs/1309.7233** (2013), `http://arxiv.org/abs/1309.7233`
17. Kosinski, M., Matz, S., Gosling, S., Popov, V., Stillwell, D.: Facebook as a social science research tool. American Psychologist (2015)
18. Mairesse, F., Walker, M.A., Mehl, M.R., Moore, R.K.: Using linguistic cues for the automatic recognition of personality in conversation and text. J. Artif. Int. Res. pp. 457–500 (2007)
19. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR **abs/1301.3781** (2013)
20. Mucha, P.J., Porter, M.A.: Communities in multislice voting networks. Chaos **20**(4), 041108 (2010)
21. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. pp. 807–814. ICML'10 (2010)
22. Pashakhanlou, A.H.: Fully integrated content analysis in international relation. International Relations **31**(4), 447–465 (2017)
23. Pennebaker, J.W., King, L.A.: Linguistic styles: Language use as an individual difference. Journal of Personality and Social Psychology pp. 1296–1312 (1999)
24. Roper Center for Public Opinion Research, Cornell University: How groups voted in 2008 (2008), `https://ropercenter.cornell.edu/polls/us-elections/how-groups-voted/how-groups-voted-2008/`
25. Rowan, M., Dehlinger, J.: Observed gender differences in privacy concerns and behaviors of mobile device end users. Procedia Computer Science **37**, 340–347 (2014)
26. Santra, A., Bhowmick, S.: Holistic analysis of multi-source, multi-feature data: Modeling and computation challenges. In: Big Data Analytics - Fifth International Conference, BDA (2017)
27. Santra, A., Bhowmick, S., Chakravarthy, S.: Efficient community re-creation in multilayer networks using boolean operations. In: International Conference on Computational Science, ICCS 2017. pp. 58–67 (2017)
28. Santra, A., Bhowmick, S., Chakravarthy, S.: Hubify: Efficient estimation of central entities across multiplex layer compositions. In: 2017 IEEE International Conference on Data Mining Workshops, ICDM Workshops (2017)
29. Sumner, C., Byers, A., Shearing, M.: Determining personality traits and privacy concerns from facebook activity. Black Hat Briefings pp. 197–221 (2011)
30. Vu, T., Nguyen, D.Q., Vu, X.S., Nguyen, D.Q., Trenell, M.: Nihrio at semeval-2018 task 3: A simple and accurate neuralnetwork model for irony detection in twitter. In: Proceedings of the 12nd International Workshop on Semantic Evaluation (SemEval-2018). pp. 525–530 (2018)

31. Vu, X.S., Flekova, L., Jiang, L., Gurevych, I.: Lexical-semantic resources: yet powerful resources for automatic personality classification. In: Proceedings of the 9th Global WordNet Conference. pp. 173–182 (2018)
32. Vu, X.S., Jiang, L.: Self-adaptive privacy concern detection for user-generated content. In: Proceedings of the 19th International Conference on Computational Linguistics and Intelligent Text Processing (2018)
33. Wikipedia contributors: Barack obama presidential campaign, 2008 — Wikipedia, the free encyclopedia (2018), `https://en.wikipedia.org/w/index.php?title=Barack_Obama_presidential_campaign,_2008&oldid=838416057`
34. Wikipedia contributors: United states presidential election, 2008 — Wikipedia, the free encyclopedia (2018), `https://en.wikipedia.org/w/index.php?title=United_States_presidential_election,_2008&oldid=850473763`
35. Zhou, C., Sun, C., Liu, Z., Lau, F.C.M.: A C-LSTM neural network for text classification. CoRR (2015)