Towards a Set Theoretical Approach to Big Data Analytics

Raghava Rao Mukkamala
IT University of Copenhagen
Rued Langgaardsvej 7,
2300 Copenhagen, Denmark
rao@@itu.dk

Abid Hussain[1] and Ravi Vatrapu[1,2]
[1]Copenhagen Business School
Howitzvej 60, 2000 Frederiksberg, Denmark
{ah.itm, rv.itm}@cbs.dk
[2]Norwegian School of Information Technology (NITH), Norway

***Abstract*—Formal methods, models and tools for social big data analytics are largely limited to graph theoretical approaches such as social network analysis (SNA) informed by relational sociology. There are no other unified modeling approaches to social big data that integrate the conceptual, formal and software realms. In this paper, we first present and discuss a theory and conceptual model of social data. Second, we outline a formal model based on set theory and discuss the semantics of the formal model with a real-world social data example from Facebook. Third, we briefly present and discuss the Social Data Analytics Tool (SODATO) that realizes the conceptual model in software and provisions social data analysis based on the conceptual and formal models. Fourth and last, based on the formal model and sentiment analysis of text, we present a method for profiling of artifacts and actors and apply this technique to the data analysis of big social data collected from Facebook page of the fast fashion company, H&M.**

***Index Terms*—Formal Methods, Social Data Analytics, Computational Social Science, Data Science, Big Social Data.**

I. Introduction

The growth of social media use in society is generating large quantities of new digital information about individuals, organizations and institutions that is now commonly labeled *Big Social Data*. Social media analytics is a term we use here to refer to the collection, storage, analysis, and reporting of these new data [1]. These social data sets carry valuable information and if analysed utilizing proper methods, techniques, and tools of computational social science in particular and data science in general. They can provide meaningful facts and actionable insights that go beyond traditional social science research methods. For example, recent studies have shown that social data on Facebook can be analysed for investigating political discourse on online public spheres for the United States Election [2], [3] and social data from twitter has been used for predicting Hollywood movies' box-office revenues [4].

Conte and colleagues [5] also point that Computational Social Science is a model based science that analyses electronic trace data, builds predictive models and intends to provide instruments for enabling social science to inform decision makers for societal and organisational challenges.

A. Formal Models

Formal modeling is a process of writing and analyzing formal descriptions of models and systems that represent real-world processes. It is a technique to model complex phenomena as mathematical entities so that rigorous analysis techniques can be applied on the models to understand the reality of the complex phenomenon. Moreover, formal specifications are abstract, precise and to some extent complete in nature [6], [7]. The abstraction of a formal specification allows to comprehend a complex phenomenon, where as the precise semantics eliminates ambiguity in the model. The completeness ensures the study of all aspects of the behavior in the model [7].

Having said that, computational methods, formal models and software tools for big social data analytics are largely limited to graph-theoretical approaches [8] such as social network analysis [9] informed by the social philosophical approach of relational sociology [10]. There are no other unified modelling approaches to social data that integrate the conceptual, formal, software, analytical and empirical realms [11]. Our objective in this paper is to present, discuss, and empirically demonstrate an alternative holistic approach to predominant triumvirate of relational sociology, graph theory, and social network analysis. Our approach is based on the alternate triumvirate of associational sociology [12], set theory and fuzzy set theory [13], and formal modelling of big social data [11].

B. Advantages of the Set Theoretical Approach

For the purposes of this paper, set-theoretical approach includes both classical (also known as crisp) as well as fuzzy sets. Smithson and Verkuilen [14] articulated the following five reasons for applying set theory in general and fuzzy set theory in particular to social science research:

1) Set-theoretical ontotology is well-suited to conceptualize vagueness which is a central aspect of social science constructs
2) Set-theoretical epistemology is well-suited for analysis of social science constructs that are both categorical and dimensional. That is, set-theoretical approach is well-suited for dealing with different types as well as degrees of a particular type.

This work was partially supported by the grant titled *Social Data Analytics Tool for Sustainability* from the Sustainability Business-in-Society Platform of the Copenhagen Business School, Denmark.

3) Set-theoretical methodology can help analyse multivariate associations *beyond the conditional means and the general linear model* (p.1)
4) Set-theoretical analysis have high theoretical fidelity with most social science theories that are usally expressed logically in set-terms
5) Set-Theoretical approach systematically combines set-wise logical formulation of social science theories and empirical analysis using statistical models for continuous variables

As we show in this paper, a set-theoretical approach to sentiment analysis of big social data will be able to analyse not only the different categories of sentiments (positive, negative, and neutral) but also their dimensions(actors, artifacts, actions, activities, probabilities etc.). As Ragin [15] argues this allows for a new paradigm of social science research termed *diversity-oriented research* to bridge the theoretical, methodological, and interpretive divide of variable-oriented research and case-oriented research.

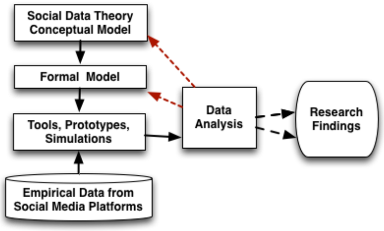


Figure 1. Overall Methodology

This paper seeks to address this problem by proposing an integrated modeling approach involving a conceptual model for social data, a formal model of the conceptual data based on set theory, a schematic model of a software application informed by the conceptual and formal models as shown in Fig. 1.

The remainder of the paper is organized as follows. We present related work in Sec. II and then we present and discuss theory of social data (Sec. III) and a conceptual model of social data (Sec. IV). Second, we outline a formal model based on set theory and discuss the semantics of the formal model with a real-world social data example from facebook in Sec. V. In Sec. VI, we present results of data analysis on big social data from Facebook page of H&M company using a method developed based on the formal model. Finally, we conclude our paper in Sec. VII.

II. Related Work

The use of Social Network Analysis can be traced back to 1979, where Tichy et.al. [16] used it as a method of examining the relationships and social structures for the analysis of organisations. Later in 1987, David Krackhardt [17] proposed cognitive social structures as a solution for social network related problems.

Due to the advent of internet and the online social media in the last decade, the field of social computing attracted many researchers. It is not possible to refer to an extensive list of research articles in this emerging area, however we refer some of the important works here. First of all, Justin Zhan and Xing Fang in [18] provided an detailed overview about state of art in social networking analysis, social and human behavioural modeling and security on social networks. A framework for calculating reputations in multi-agent systems using social network analysis has been proposed in [19], where as social network analysis based on measuring social relations using multiple data sets has been explored in [20]. An algorithm to find overlapping communities via social network analysis was explored in [21]. Moreover, analysis of sub-graphs in the social network based on the characteristic features: leadership, bonding, and diversity was studied by the authors in [22]. More over, several researchers have developed formal techniques for network analysis [23]–[25] and applied those techniques to social networks. All these works are primarily focussed on analysing the social networks based on the structural relationships between the actors only. On the other hand, our work primarily focussed combining the structural aspects of social data with the content analysis of social text, to study the behavioral aspects and to further develop advanced analysis techniques for social data.

Semantic-level precedence relationships between participants in a blog network are studied in [26], where the authors proposed a methodology for the detection of bursts of activity at the semantic level using linguistic tagging, term filtering and term merging. They used a probabilistic approach to estimate temporal relationships between the blogs. However in an another interesting work, Sitaram Asur and Bernardo A. Huberman [4] showed that social media feed can be used as effective indicators of the real-world performance. In their work, they used analysis of sentiment content on urls, retweets and their hourly rates of Twitter to estimate to forecast the box-office movies revenue.

We find that the extant literature is primarily focused on using social network analysis and other graph theory related formalisms. In contrast, we propose to use Set Theory for the formal modelling of associations between actors, actions, artifacts, topics and sentiments.

III. Theory of Social Data

Social media platforms such as Facebook and Twitter, at the highest level of abstraction, involve individuals interacting with (a) technologies and (b) other individuals. These interactions are termed ***socio-technical interactions***. There are two types of socio-technical interactions: 1) interacting with the technology per se (for example, using the Facebook app on the user's smartphone and 2) interacting with social others using the technology (for example, liking a picture of a friend in the Face book app of the user's smartphone). These socio-technical interactions are theoretically conceived as (a) perception and appropriation of socio-technical affordances, and (b) structures and functions of technological intersubjectivity. Briefly, socio-technical affordances are action-taking possibilities and meaning-making opportunities in an actor-environment system bounded by the cultural-cognitive competencies of the actor and the technical capabilities of the environment. Technological intersubjectivity (TI) refers to a technology supported interactional social relationship between two or more actors. A more detailed explication of the theoretical framework in terms of its ontological and epistemological assumptions and principles is beyond the scope of this paper but for details, please confer [27], [28].

Socio-technical interactions as described above result in electronic trace data that is termed ”social data”. For the example discussed of a Facebook user liking a friend’s picture on their smartphone app, the social data is not only rendered in the different ”timelines” of the user’s social network but it is available via the Facebook graph API. Large volumes of such micro-interactions constitute the macro world of big social data that is the analytical focus of this paper. Based on the theory of social data described above, we present a conceptual model of social data below.

IV. Conceptual Model

Social data consists of two types: *Social Graph* and *Social Text* as shown in the Fig. 2. Social Graph maps on to the first aspect of socio-technical interactions that involve perception and appropriation of affordances (which users/actors act up on which technological features to interact with what other social actors in the systems). Social Text maps on to the second aspect of socio-technical interactions that constitute the structures and functions and technological intersubjectivity (what the users/actors are trying to communicate to each other and how they are trying to influence each other through language).

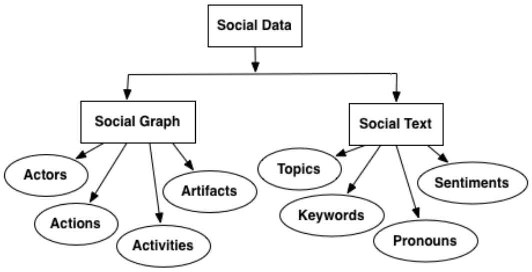


Figure 2. Social Data Model [29]

Social graph consists of the structure of the relationships emerging from the apprproiation of social media affordances such as posting, linking, tagging, sharing, liking etc. It focuses on identifying the **actors** involved, the **actions** they take, the **activities** they undertake, and the **artifacts** they create and interact with. Social text consists of the communicative and linguistic aspects of the social media interaction such as the **topics** discussed, **keywords** mentioned, **pronouns** used and **sentiments** expressed.

We now turn our attention to formalizing the conceptual model as we believe that formal models are essential for the application of computational techniques and tools, given not only the large volumes of data involved but also their ambiguity and unstructured nature.

V. Formal Model

In this section, we will provide formal semantics for social data model, which was initially presented in an internal technical report [11], which is an unpublished and non peer reviewed report.

Notation: For a set A we write $\mathcal{P}(A)$ for the power set of A (i.e. set of all subsets of A) and $\mathcal{P}_{disj}(A)$ for the set of mutually disjoint subsets of A. The cardinality or number of elements in a set A is represented as $\mid A \mid$. Furthermore, we write a relation R from set A to set B as $R \subseteq A \times B$. A function f defined from a set A to set B is written as $f : A \rightarrow B$, where a if f is a partial function then it is written as $f : A \rightharpoonup B$.

First, we define type of artifacts in a socio-technical system as shown in Def. 5.1.

Definition 5.1: We define $\mathbb{R}$ as a set of all artifact types as $\mathbb{R}$ = { status, comment, link, photo, video }.

Definition 5.2: We define $\mathbb{A}_{\mathbb{CT}}$ as a set of actions that can be performed as $\mathbb{A}_{\mathbb{CT}}$ = {*post*, *comment*, *share*, *like*, *tagging*}.

As explained in the conceptual model, the social data model contains Social Graph and Social Text, which is formally defined in Def. 5.3 as follows,

Definition 5.3: Formally, Social Data is defined as a tuple $\mathsf{S} = (\mathsf{G}, \mathsf{T})$ where

(i) G is the social graph representing the structural aspects of social data as defined further in Def. 5.4

(ii) T is the social text representing the content of social data and is further defined in Def. 5.5

As shown in the first two items (i, ii, x) of Def. 5.4, the social graph primarily contains a set of actors or users (U), a set of artifacts or resources (R) and a set of activities (Ac). Each artifact is mapped to an artifact type (such as status, photo etc) by artifact type function (Def. 5.4-iv). In addition to that, some of the artifacts are mapped to their parent artifact (if exists) by parent artifact function $\triangleright$ (Def. 5.4-v). For example, if the artifact is a comment on a post, then it is mapped to its parent (which is the post), on the other hand, if the artifact is a status message or a new post, then it will not have any parent.

Furthermore, each artifact is posted by single actor. As shown in Def. 5.4-vi, the $\rightarrow_{post}$ is a partial function mapping actors to mutually disjoint sub sets of artifacts, each set containing artifacts created or posted by an actor. On contrary, the $\rightarrow_{share}$ indicates a many-to-many relationship, indicating that an artifact can be shared by many actors and similarly each actor can share many artifacts (Def. 5.4-vii). Even though *share* and *post* actions seems to be similar, the $\rightarrow_{post}$ signifies the creator relationship of an artifact, where as $\rightarrow_{share}$ indicates share relationship between an artifact and an actor which can be many-to-many.

Similar to the *share* relation, the *like* relation ($\rightarrow_{like}$) models mapping between the artifacts and actors, indicating the artifacts liked by the actors. The *tagging* relation ($\rightarrow_{tag}$)

is a bit different, which is a mapping between actors, artifacts and power set of actors and keywords (Def. 5.4-ix). The basic intuition behind the tag relation is that, it allows an actor to tag other actors or keywords in an artifact. Finally, the $\rightarrow_{\mathsf{act}}$ relation indicates a mapping between artifacts to activities (Def. 5.4-x).

Definition 5.4: The Social Graph is defined as a tuple $\mathsf{G} = (\mathsf{U}, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \triangleright, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$ where

(i) U is a finite set of actors/ users ranged over by u,

(ii) R is the finite set of artifacts (resources) ranged over by r,

(iii) Ac is a finite set of activities,

(iv) $\mathsf{r}_{\mathsf{type}} : \mathsf{R} \rightarrow \mathbb{R}$ is the artifact type function mapping each artifact to a artifact type defined in 5.1,

(v) $\triangleright : \mathsf{R} \rightharpoonup \mathsf{R}$ is parent artifact function, which is a partial function mapping artifacts to their parent artifact if defined,

(vi) $\rightarrow_{post} : \mathsf{U} \rightharpoonup \mathcal{P}_{disj}(\mathsf{R})$ is a partial function mapping actors to mutually disjoint subsets of artifacts,

(vii) $\rightarrow_{share} \subseteq \mathsf{U} \times \mathsf{R}$ is a relation mapping users to artifacts,

(viii) $\rightarrow_{like} \subseteq \mathsf{U} \times \mathsf{R}$ is a relation mapping users to the artifacts indicating the artifacts liked by the users,

(ix) $\rightarrow_{tag} \subseteq \mathsf{U} \times \mathsf{R} \times (\mathcal{P}(\mathsf{U} \cup \mathrm{Ke}))$ is a tagging relation mapping artifacts to power sets of actors and keywords indicating tagging of actors and keywords in the artifacts, where Ke is set of keywords defined in Def. 5.5,

(x) $\rightarrow_{\mathsf{act}} \subseteq \mathsf{R} \times \mathrm{Ac}$ is a relation mapping artifacts to activities.

As explained in the conceptual model, the Social Text mainly contains set of *topics* (To), *keywords* (Ke), *pronouns* (Pr), and *sentiments* (Se) as defined in Def. 5.5. The $\rightarrow_{\mathsf{topic}}$, $\rightarrow_{\mathsf{key}}$, $\rightarrow_{\mathsf{pro}}$ and $\rightarrow_{\mathsf{sen}}$ relations map the artifacts to the *topics* (To), *keywords* (Ke), *pronouns* (Pr), and *sentiments* (Se) respectively. One may note that all these relations allow many-to-many mappings, for example an artifact can be mapped to more than one sentiment and similarly a sentiment can contain mappings to many artifacts.

Definition 5.5: In Social Data $\mathsf{S} = (\mathsf{G}, \mathsf{T})$, we define Social Text as T= (To, Ke, Pr, Se, $\rightarrow_{\mathsf{topic}}, \rightarrow_{\mathsf{key}}, \rightarrow_{\mathsf{pro}}, \rightarrow_{\mathsf{sen}}$) where

(i) $\mathrm{To}, \mathrm{Ke}, \mathrm{Pr}, \mathrm{Se}$ are finite sets of topics, keywords, pronouns and sentiments respectively,

(ii) $\rightarrow_{\mathsf{topic}} \subseteq \mathsf{R} \times \mathrm{To}$ is a relation defining mapping between artifacts and topics,

(iii) $\rightarrow_{\mathsf{key}} \subseteq \mathsf{R} \times \mathrm{Ke}$ is a relation mapping artifacts to keywords,

(iv) $\rightarrow_{\mathsf{pro}} \subseteq \mathsf{R} \times \mathrm{Pr}$ is a relation mapping artifacts to pronouns,

(v) $\rightarrow_{\mathsf{sen}} \subseteq \mathsf{R} \times \mathrm{Se}$ is a realtion mapping artifacts to sentiments.

A. Operational Semantics

In this section, we will define the operational semantics of the model. More precisely, we define how actors perform actions on artifacts.

As formally defined in Def. 5.6, the first action is *post*, which accepts a pair containing an actor and a new artifact (u, r). First, the actor will be added to the set of actors (i) and then the new artifact will be added to the set of artifacts (ii). Finally the post relation ($\rightarrow_{post}$) will be updated for the new mapping (iii).

Definition 5.6: In Social Data $\mathsf{S} = (\mathsf{G}, \mathsf{T})$ with $\mathsf{G} = (\mathsf{U}, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \triangleright, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$, we define a **post** operation of posting a new artifact r ($r \notin \mathsf{R}$) by an user u as $\mathsf{S} \bigoplus_p (u, r) = (\mathsf{G}', \mathsf{T})$ where $\mathsf{G}' = (\mathsf{U}', \mathsf{R}', \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \triangleright, \rightarrow_{post}{}', \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$,

(i) $\mathsf{U}' = \mathsf{U} \cup \{u\}$

(ii) $\mathsf{R}' = \mathsf{R} \cup \{r\}$

(iii) $\rightarrow_{post}{}' = \begin{cases} \rightarrow_{post}(u) \cup \{r\} & \text{if } \rightarrow_{post}(u) \text{ defined} \\ \rightarrow_{post} \cup \{u, \{r\}\} & \text{otherwise} \end{cases}$

The *comment* action (e.g. on a post) accepts a tuple containing an actor, the parent artifact (on which the comment is made) and the comment content itself as shown in the Def. 5.7. As it creates a new artifact, it will first apply a *post* action to create the comment as a new artifact with the actor (i) and then followed by an update to the parent artifact function ($\triangleright$) by adding the respective mapping for comment with its parent (ii).

Definition 5.7: Let Social Data be $\mathsf{S} = (\mathsf{G}, \mathsf{T})$ with $\mathsf{G} = (\mathsf{U}, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \triangleright, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$, the **comment** operation on an artifact r_p ($r_p \in \mathsf{R}$) by an user u for a new artifact r is formally defined as $\mathsf{S} \bigoplus_c (u, r, r_p) = (\mathsf{G}', \mathsf{T})$ where $\mathsf{G}' = (\mathsf{U}', \mathsf{R}', \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \triangleright', \rightarrow_{post}{}', \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$,

(i) $\mathsf{S} \bigoplus_p (u, r) = (\mathsf{G}'', \mathsf{T})$ where $\mathsf{G}'' = (\mathsf{U}', \mathsf{R}', \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \triangleright, \rightarrow_{post}{}', \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$,

(ii) $\triangleright' = \triangleright \cup \{r, r_p\}$

As mentioned before, the *share* operation does not create any new artifact, but it will updates the actors set and then makes an update to the share relation ($\rightarrow_{share}$) as formally defined in Def. 5.8.

Definition 5.8: Let Social Data be $\mathsf{S} = (\mathsf{G}, \mathsf{T})$ with $\mathsf{G} = (\mathsf{U}, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \triangleright, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$, then we define the **share** operation consisting of sharing an artifact r by an user u as $\mathsf{S} \bigoplus_s (u, r) = (\mathsf{G}', \mathsf{T})$ where $\mathsf{G}' = (\mathsf{U} \cup \{u\}, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \triangleright, \rightarrow_{post}, \rightarrow_{share} \cup \{(u, r)\}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$.

In the Def. 5.9, we formally define the *like* and *unlike* operations as an update to the like relation ($\rightarrow_{like}$). A *like* action on an artifact will add a mapping to like relation ($\rightarrow_{like}$) (in addition to adding the actor to the actors set), where as an *unlike* action will simply remove the existing mapping.

Definition 5.9: In a Social Data $\mathsf{S} = (\mathsf{G}, \mathsf{T})$ with Graph $\mathsf{G} = (\mathsf{U}, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \triangleright, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$, we define the **like** operation by an user u on an artifact r as $\mathsf{S} \bigoplus_l (u, r) = (\mathsf{G}', \mathsf{T})$ where $\mathsf{G}' = (\mathsf{U} \cup \{u\}, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \triangleright, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like} \cup \{(u, r)\}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$.

Similarly, we also define the **unlike** operation on $\mathsf{S} = (\mathsf{G}, \mathsf{T})$ with Graph $\mathsf{G} = (\mathsf{U}, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \rhd, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$, as $\mathsf{S} \ominus_{l} (u, r) = (\mathsf{G}', \mathsf{T})$ where $\mathsf{G}' = (\mathsf{U}, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \rhd, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like} \setminus \{(u, r)\}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$.

Finally, the *tagging* action accepts a tuple ((u, r, t)) containing an actor, an artifact and a set of hash words (i.e. keywords and actors) and an update to tagging relation ($\rightarrow_{tag}$) will be applied as shown in the Def. 5.10.

Definition 5.10: In a Social Data $\mathsf{S} = (\mathsf{G}, \mathsf{T})$ with Graph $\mathsf{G} = (\mathsf{U}, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \rhd, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$, we define the **tagging** operation by an user u on an artifact r with a set of hash words $t \in \mathcal{P}(\mathsf{U} \cup \mathrm{Ke})$ as $\mathsf{S} \bigoplus_{t} (u, r, t) = (\mathsf{G}', \mathsf{T})$ where $\mathsf{G}' = (\mathsf{U} \cup \{u\}, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \rhd, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag} \cup \{(u, r, t)\}, \rightarrow_{\mathsf{act}})$.

Finally, we also define a function ($\mathsf{T}_{\mathsf{ime}}$) to keeps track of the timestamps of the artifact's created time.

B. Example

In this section, we exemplify the formal model by taking an example from the Facebook page of H&M cloth stores as shown in the figure 3. In order to enhance the readability of the example, the artifacts (e.g. texts) have been annotated as $r1, r2$ etc and the annotated values will be used in encoding the example using the formal model.

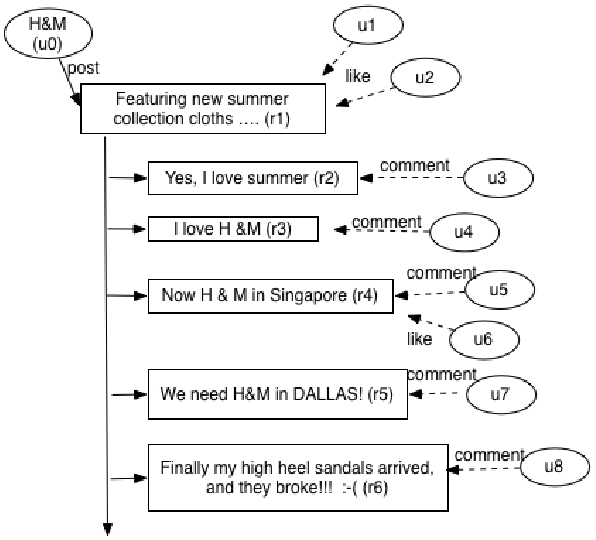

H&M
(u0)
post
u1
like
u2
Featuring new summer collection cloths .... (r1)
Yes, I love summer (r2)
comment
u3
I love H &M (r3)
comment
u4
Now H & M in Singapore (r4)
comment
u5
like
u6
We need H&M in DALLAS! (r5)
comment
u7
Finally my high heel sandals arrived, and they broke!!! :-((r6)
comment
u8

Figure 3. Example in formal model

Example 5.1: The example shown in Fig. 3 will be encoded as follows,
$\mathsf{S} = (\mathsf{G}, \mathsf{T})$ where $\mathsf{G} = (\mathsf{U}, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \rhd, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$ is the social graph and T= (To, Ke, Pr, Se, $\rightarrow_{\mathsf{topic}}, \rightarrow_{\mathsf{key}}, \rightarrow_{\mathsf{pro}}, \rightarrow_{\mathsf{sen}}$) is the Social Text.

Initailly, the sets of activities, topics, keywords, pronouns and sentiments will have the following values.
$\mathrm{Ac} = \{promotion\}$,
$\mathrm{To} = \{\text{summer collection}, \text{new store request}\}$,
$\mathrm{Ke} = \{H\&M, Dallas, Singapore\}$
$\mathrm{Pr} = \{We, I\}, \mathrm{Se} = \{+, 0, -\}$,
$\mathsf{U} = \{u_0, u_1, u_3, \} \rightarrow_{\mathsf{act}} = \{(r_1, promotion)\}$

post action by u_0
$\mathsf{S} \bigoplus_{p} (u_0, r_1) = \mathsf{S}_1 = (\mathsf{G}_1, \mathsf{T})$ where
$\mathsf{G}_1 = (\mathsf{U}_1, \mathsf{R}_1, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \rhd, \rightarrow_{post\ 1}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$ with the following values
$\mathsf{U}_1 = \mathsf{U} \cup \{u_0\}$, $\mathsf{R}_1 = \mathsf{R} \cup \{r_1\}$ and
$\rightarrow_{post_1} = \rightarrow_{post} \cup \{(u_0, \{r_1\})\}$

like action by u_2
$\mathsf{S}_1 \bigoplus_{l} (u_2, r_1) = \mathsf{S}_2 = (\mathsf{G}_2, \mathsf{T})$ where
$\mathsf{G}_2 = (\mathsf{U}_2, \mathsf{R}_1, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \rhd, \rightarrow_{post\ 1}, \rightarrow_{share}, \rightarrow_{like\ 1}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$ with the following values
$\mathsf{U}_2 = \mathsf{U}_1 \cup \{u_2\}$, and $\rightarrow_{like\ 1} = \rightarrow_{like} \cup \{(u_0, r_1)\}$

comment action by u_3
$\mathsf{S}_2 \bigoplus_{c} (u_3, r_2, r_1) = \mathsf{S}_3 = (\mathsf{G}_3, \mathsf{T})$ where
$\mathsf{G}_3 = (\mathsf{U}_3, \mathsf{R}_2, \rhd_1, \mathsf{r}_{\mathsf{type}}, \mathrm{Ac}, \rightarrow_{post\ 2}, \rightarrow_{share}, \rightarrow_{like\ 1}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$ with the following values
$\mathsf{U}_3 = \mathsf{U}_2 \cup \{u_3\}$, $\mathsf{R}_2 = \mathsf{R}_1 \cup \{r_2\}$, $\rightarrow_{post\ 2} = \rightarrow_{post\ 1} \cup \{(u_3, \{r_2\})\}$ and $\rhd_1 = \rhd \cup \{(r_2, r_1)\}$.

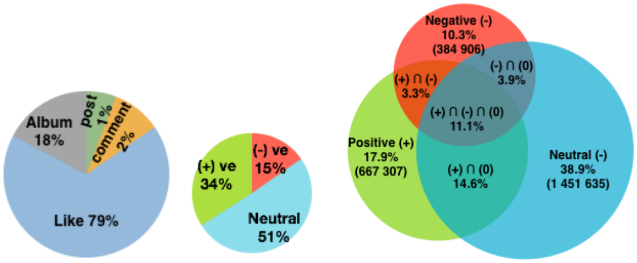
VI. Data Analysis

In this section, we present data analysis for profiling of actors and artifacts based on the formal model. First, we outline the method for calculating the sentiment for actors based on the sentiments on artifacts and then we will present results of the analysis that was carried out on the big social data extracted from the Facebook page of the H & M company.

As part of case study, Facebook data of H&M was fetched by SODATO [30] from 01-Jan-2009 to 31-December-2013. The Facebook data corpus consists of a total of 12,577,235 entries, which consists of posts, comments, likes and albums as shown in Fig. 4(a). Around 1% (112,211) and 2% (297,064) of entries are posts and comments respectively. The H & M data corpus is dominated by Likes (9, 947,567 likes on posts & comments), which is followed by the comments and likes on albums.

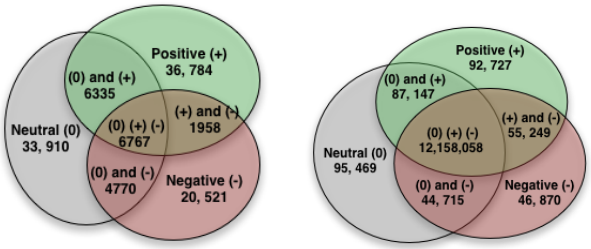
In prior work [11], we reported statistically significant correlations between real-world business outcomes (quarterly sales) and social media activities (measures of social graph (posts, likes, comments) as well as social text (positive, negative or neutral sentiment expressions). With regard to social graph, statistically significant strong correlations were observed between quarterly sales and total likes, total likes on the company's posts as well as users' posts and total comments on users' posts [11]. With regard to social text, statistically significant strong positive correlations were observed for positive sentiment expression only for Comments on Posts by Non-H&M users on the facebook wall. On the other hand, strong correlations were observed, surprisingly, for the negative sentiment expressions on Total Posts, Posts by Non-H&M and Comments on Posts by Non-H&M facebook users [11],

As we discussed under *Advantages of Set Thoeretical Approach* (Sec. I-B) we constructed crisp sets of the sentiments



(a) Artifacts (b) Artifact sentiment (c) Actor sentiment

Figure 4. Overview of Artifact and Actor Sets



(a) Posts (b) Posts, Comments, Likes and Shares

Figure 5. Overview of Artifact Sentiment Sets

expressed by actors performing actions on artifacts to better understand the statistical correlation between real-world outcomes (quarterly sales) and social media activities (facebook engagment) of H&M. Towards this end, we now report data analysis and findings from the crisp set analysis of actor and artifact sentiment that reveal seasonal variation (more peaks during the spring and fall period where the fashion industry traditionally reveals new products) as well as crisis periods (for example, garment factory accidents in Bangladesh), as shown in Fig. 8.

A. Methodology

Actors perform *Actions* in *Activities* on *Artifacts*. Artifacts carry direct sentiment as they contain content, which can be analysed by machine learning tools such as sentiment engine of Google Prediction API [31] and thereby artifacts get a sentiment score and a label. Individually, an *action* does not carry any sentiment, but it is the artifacts on which these actions are carried over, that contain sentiments. Similarly, *actors* do not carry any sentiment directly, but they express their sentiments by performing *actions* on the *artifacts*, which contain the direct sentiment. Therefore, the sentiment attributed to an actor can be inferred from the artifacts on which the actions are performed. The set of sentiments in the Social Text contains some predefined labels: *positive* $(+)$, *neutral* (0) and *negative* $(-)$ as indicated in $\mathrm{Se} = \{+, 0, -\}$. Normally, the sentiment scores are expressed as real numbers (between 0 to 1), and the sum of such scores of an artifact for multiple sentiment labels will be equal to 1. As an example, the artifact $r2$ from Fig. 3 can be categorised as $\{+ : 0.65, 0 : 0.30, - : 0.05\}$. However, in this paper, we have considered the default sentiment labels only for the artifacts.

1) Artifacts: Form the formal model, one can infer a set of artifacts (posts and comments) that belong to a sentiment $(se \in \mathrm{Se})$ as follows,
$\mathsf{R}^{se} = \{r \mid (r, se) \in \rightarrow_{\mathsf{sen}}\}$.
Similarly, the set of artifacts which are posts only (as shown in Fig. 5(a)) can computed as follows,
$\mathsf{R}^{se}_{posts} = \{r \mid (r, se) \in \rightarrow_{\mathsf{sen}} \wedge \rhd(r) \text{ is undefined}\}$.
The set of artifacts that belong for a given time period $t_1 - t_2$ (e.g. quarterly as shown in Fig. 6) as follows,
$\mathsf{R}^{se}_{t_1-t_2} = \{r \mid (r, se) \in \rightarrow_{\mathsf{sen}} \wedge (t_1 \leq \mathsf{T}_{\mathsf{ime}}(r) \leq t_2)\}$.
Finally, the number of posts, comments, likes and shares for each sentiment label (as shown in Fig. 5(b)) can be computed as $|\mathsf{R}^{se}| + |\{u \mid r \in \mathsf{R}^{se} \wedge (u, r) \in (\rightarrow_{share} \cup \rightarrow_{like})\}|$.

2) Actors: Furthermore, the set of actors that are associated with any given set of artifacts (e.g. R^{se}) that pertains to a sentiment label (se) can be computed as follows, $\forall r \in \mathsf{R}^{se}$.
$\mathsf{U}_{\mathsf{R}^{se}} = \{u \mid r \in \rightarrow_{post} (u)\} \cup \{u \mid (u, r) \in (\rightarrow_{share} \cup \rightarrow_{like})\}$.

The set of actors contains actors who posted an artifact and those who shared and liked the artifact as shown in Fig. 4(c). From the set of actors ($\mathsf{U}_{\mathsf{R}^{se}}$), one could compute sets for given time periods (e.g. quarterly) to obtain a frequency distribution of actors sentiment over temporal dimension as shown in the Fig. 7.

B. H&M Case Study - Results

The sentiment analysis for the whole data corpus (artifacts) was carried and sentiment sets of artifacts were computed as explained in the previous section. The distribution of the *post* artifacts and total entries are shown in the Fig. 5(a) and Fig. 5(b) respectively. By inspecting Fig. 5(a), we find that a majority of the conversations as embodied by posts and comments on those posts belong to the exclusive sentiment categories of positive only, negative only, and neutral only. A minority of posts and comments on them display a mixture of sentiment categories (6767 posts and their comments have all three positive, negative, and neutral sentiments). These are dimensions of sentiment categories that are revealed and made available by the set-theoretical approach of this paper for further qualitative and/or quantitative analysis. Comparing the two venn diagrams, it can be noticed that even though *posts* are distributed more or less equally over exclusive sentiment labels, in the total entries category, it is interesting to note that 96% of entries (12,158,058) corresponds (mostly likes) to the *posts* in the intersection of $(+) \cap (-) \cap (0)$. In other words, the subset of 6767 posts that embody all three sentiment categories attract the most number of likes from actors compared to any other subset in the venn diagrams. The quarterly distribution of artifacts sentiments over the temporal dimension is shown in Fig. 6.

In the data corpus, there are 3,734,629 unique actors whose inferred sentiment distribution is shown in Fig. 4(c). One could notice that around 40% of the users (1,451,635)

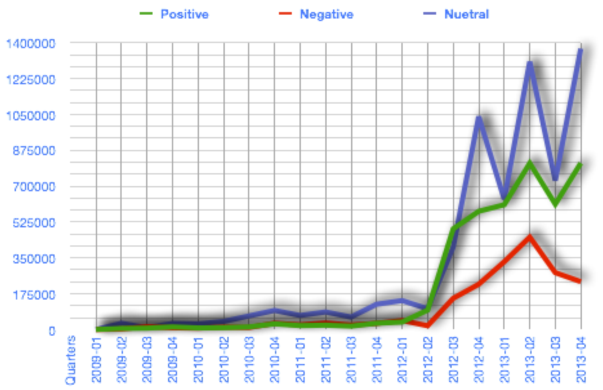


Figure 6. Quarterly Distribution of Artifact Sentiment

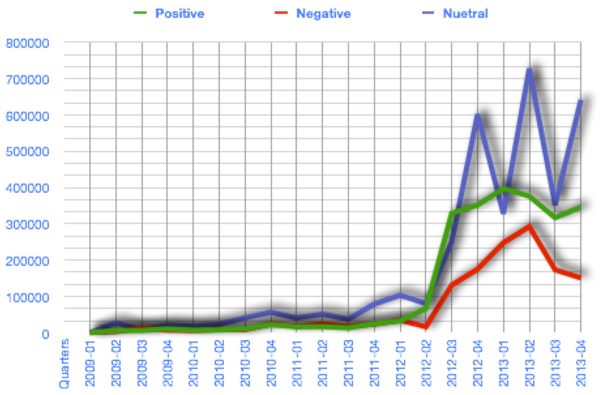


Figure 7. Quarterly Distribution of Actors Sentiment

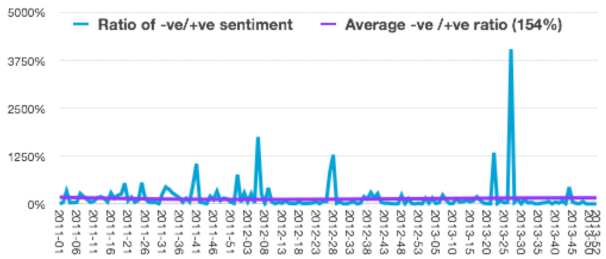


Figure 8. Weekly Distribution of -ve/+ve Sentiment ratio

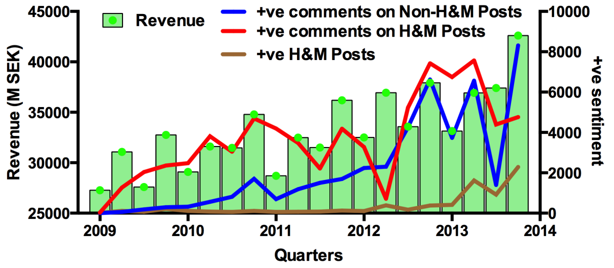


Figure 9. Quarterly distribution of H&M sales vs +ve sentiment

belong to the neutral category only performing actions on the artifacts belonging to the neutral sentiment category. The same phenomenon can be observed in Fig. 7 where quarterly

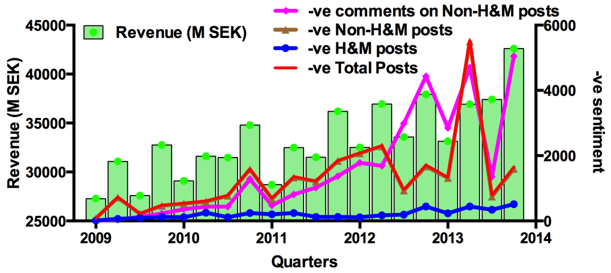


Figure 10. Quarterly distribution of H&M sales vs -ve sentiment

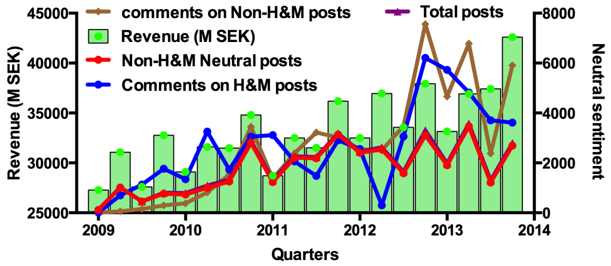


Figure 11. Quarterly distribution of H&M sales vs neutral sentiment

distribution of actors sentiment over the temporal dimension is plotted.

As reported in [11], we found surprisingly strong positive correlations of quarterly revenues with negative sentiments on total posts, posts by Non-H&M users and comments on posts by Non-H&M Facebook users. In this paper, we report on subsequent analysis (Fig. 9, 10 & 11) of the corpus based on the set-theoretical approach. The venn diagrams in Fig. 4 & 5 together with the temporal distribution of negative sentimentsin Fig. 10 provide preliminary evidence that negative sentiment category in itself is not detrimental to the brand identity and business value if it not directed towards the company. As we first observed in [11], sentiment polarity is necessary but not sufficient for predicting business outcomes. Analytical approaches based on set theory can help better understand not only the categories of sentiments but also their dimensions.

In summary, we have documented descriptive findings based on the set-theoretical analysis of the statitically signficant correlations between social data measures of sentiments expressed and real-world outcomes of quarterly sales. These descriptive findings can be turned into prescriptive recommendations and predictive analytics for companies once they are tested across other kinds of social data (twitter, pinterest, instagram), other kinds of companies the same industry sector (fast fashion), and other industry sectors.

VII. Conclusion

Set theoretical approaches to formal modelling of social data hold several advantages over graph theoretical approaches that underpin the different methods and techniques of social network analysis (SNA). To be more specific, set theoretical approaches model social associations (such as if an actor is associated with positive sentiments) rather than social relations.

This is particularly useful in analysing sentiments of temporal evolution and overall composition of artifacts and actors.

Automated sentiment annotation of social data artifacts based on computational linguistics methods such as machine learning produce both classifications of tokens into types (such as positive, negative and neutral) as well as probabilistic estimates. As we have demonstrated in this paper, these classifications and probabilities can be used to reveal historical developmental patterns as well as overlapping categories.

Practical implications from the analysis presented here could help inform an organization to assess the size of the different actor/community types such as entirely positive, partially positive, entirely negative etc. For example, investigating the absolute and relative size of entirely negative conversations might enable the organization to identify the underlying customer service issues and/or content problems. Similarly, knowing the absolute and relative number of social media users that exclusively express positive sentiments towards the organization helps identify and nurture the advocacy group.

In this paper we have presented an integrated modeling approach for analysis of social data using a conceptual model on social data, a formal model modeling the key concepts of the conceptual model and a schematic model of a software application developed based on the conceptual and formal models.

The formalization of the conceptual model allows the necessary abstraction to comprehend the complex scenarios of social data. On top of that, the formal model also served as a bridge between the conceptual model and schematic model of the software application and helped in concretising the abstract ideas from the conceptual model to schematic model in the process of developing the Social Data Analytics Tool. Moreover, we have also presented a method for profiling of artifacts and actors and appleid this technique to the data analysis of big social data collected from Facebook page of the fast fashion company, H&M. Modeling social concepts in general involves fuzziness. As part of future work, we would like to use Fuzzy set theory to model fuzzy behaviour in the social data.

References

[1] R. Vatrapu, "Understanding social business." in *Emerging Dimensions of Technology Management*. Springer, 2013, pp. 147–158. 1

[2] S. P. Robertson, R. K. Vatrapu, and R. Medina, "Off the wall political discourse: Facebook use in the 2008 u.s. presidential election," *Info. Pol.*, vol. 15, no. 1,2, pp. 11–31, Apr. 2010. 1

[3] S. Robertson, R. K. Vatrapu, and R. Medina, "Online video friends social networking: Overlapping online public spheres in the 2008 u.s. presidential election," *Journal of Information Technology & Politics*, vol. 7, no. 2-3, pp. 182–201, 2010. 1

[4] S. Asur and B. Huberman, "Predicting the future with social media," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, vol. 1, 2010, pp. 492–499. 1, 2

[5] R. Conte, N. Gilbert, G. Bonelli, C. Cioffi-Revilla, G. Deffuant, J. Kertesz, V. Loreto, S. Moat, J.-P. Nadal, A. Sanchez, A. Nowak, A. Flache, M. San Miguel, and D. Helbing, "Manifesto of computational social science," *The European Physical Journal Special Topics*, vol. 214, no. 1, 2012. 1

[6] A. Nowak, A. Rychwalska, and W. Borkowski, "Why simulate? to develop a mental model," *Journal of Artificial Societies and Social Simulation*, vol. 16, no. 3, 2013. 1

[7] A. Hall, "Realising the benefits of formal methods," in *Formal Methods and Software Engineering*, ser. LNCS. Springer Berlin Heidelberg, 2005, vol. 3785. 1

[8] J. L. Gross and J. Yellen, *Graph theory and its applications*. CRC press, 2005. 1

[9] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca, "Network analysis in the social sciences," *science*, vol. 323, no. 5916, pp. 892–895, 2009. 1

[10] M. Emirbayer, "Manifesto for a relational sociology," *The American Journal of Sociology*, vol. 103(2), pp. 281–317, 1997. 1

[11] R. R. Mukkamala, A. Hussain, and R. Vatrapu, "Towards a formal model of social data," IT University of Copenhagen, Denmark, IT University Technical Report Series TR-2013-169, November 2013. 1, 3, 5, 7

[12] B. Latour, *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press, USA, 2005. 1

[13] C. C. Ragin, *Fuzzy-set social science*. University of Chicago Press, 2000. 1

[14] M. J. Smithson and J. Verkuilen, *Fuzzy Set Theory : Applications in the Social Sciences (Quantitative Applications in the Social Sciences)*. SAGE Publications, Feb. 2006. [Online]. Available: http://www.worldcat.org/isbn/076192986X 1

[15] C. C. Ragin, "Fuzzy sets: calibration versus measurement," *Methodology volume of Oxford handbooks of political science*, 2007. 2

[16] N. M. Tichy, M. L. Tushman, and C. Fombrun, "Social network analysis for organizations," *The Academy of Management Review*, vol. 4, no. 4, October 1979. 2

[17] D. Krackhardt, "Cognitive social structures," *Social Networks*, vol. 9, no. 2, pp. 109–134, Jun. 1987. 2

[18] J. Zhan and X. Fang, "Social computing: the state of the art," *International Journal of Social Computing and Cyber-Physical Systems*, vol. 1, no. 1, pp. 1–12, 01 2011. 2

[19] J. Sabater and C. Sierra, "Reputation and social network analysis in multi-agent systems," in *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 1*, ser. AAMAS '02. New York, NY, USA: ACM, 2002, pp. 475–482. 2

[20] J. Karikoski and M. Nelimarkka, "Measuring social relations with multiple datasets," *IJSCCPS*, vol. 1, no. 1, pp. 98–113, 2011. 2

[21] M. Goldberg, S. Kelley, M. Magdon-Ismail, K. Mertsalov, and A. Wallace, "Finding overlapping communities in social networks," in *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, 2010, pp. 104–113. 2

[22] O. Macindoe and W. Richards, "Comparing networks using their fine structure," *International Journal of Social Computing and Cyber-Physical Systems*, vol. 1, no. 1, 2011. 2

[23] A.-L. Barabsi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999. 2

[24] A. Clauset, C. Moore, and M. E. J. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, May 2008. 2

[25] M. E. J. Newman, "The structure and function of complex networks," *SIAM REVIEW*, vol. 45, pp. 167–256, 2003. 2

[26] T. Menezes, C. Roth, and J.-P. Cointet, "Finding the semantic-level precursors on a blog network." *IJSCCPS*, vol. 1, no. 2, pp. 115–134, 2011. 2

[27] R. K. Vatrapu, "Technological intersubjectivity and appropriation of affordances in computer supported collaboration," Ph.D. dissertation, University of Hawaii at Manoa, USA, 2007, aAI3302125. 3

[28] ——, "Explaining culture: An outline of a theory of socio-technical interactions," in *Proceedings of the 3rd International Conference on Intercultural Collaboration*, ser. ICIC '10. New York, NY, USA: ACM, 2010, pp. 111–120. 3

[29] A. Hussain, R. Vatrapu, D. Hardt, and Z. Jaffari, "Social data analytics tool: A demonstrative case study of methodology and software." in *Analysing Social Media Data and Web Networks*, M. C. Rachel Gibson and S. Ward, Eds. Palgrave Macmillan, 2014 (in press). 3

[30] A. Hussain and R. Vatrapu, "Social data analytics tool (sodato)," in *DESRIST 2014*, ser. Lecture Notes in Computer Science (LNCS). Springer, vol. 8463, 2014, pp. 368–372. 5

[31] Google-Inc, "Google prediction api," September 2012, https://developers.google.com/prediction/. 6