Journal of Universal Computer Science, vol. 22, no. 3 (2016), 394-415
submitted: 29/10/15, accepted: 23/1/16, appeared: 1/3/16 © J.UCS

Mining Social Networks for Calculation of SmartSocial Influence

Vanja Smailovic
(Ericsson Nikola Tesla, Zagreb, Croatia
vanja.smailovic@ericsson.com)

Vedran Podobnik
(University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
vedran.podobnik@fer.hr)

Abstract: In today's networked society where everybody and everything becomes inter-connected, it is very important to be able to identify *key actors* and *key relationships* in such a complex multi-layered eco-system. This paper focuses on the specific research challenge of identifying the *most influential actors* in a social network built through combining relationships among same actors in two different domains – *communication domain* (proxied through real-world mobile phone communication data) and *social networking service domain* (proxied through real-world Facebook data). A practical aspect of the paper is evaluated through the SmartSocial Platform, which uses methodology and implements algorithms that enable: i) joining multiple relations among actors across different social networks into the single unified social network; as well as ii) mining created unified social network for identification of most influential actors. Evaluation of the proposed approach is based on the social experiment with 465 users. Experiment results underline two important paper contributions: i) posting frequency sensitivity analysis shows a significant effect of posting frequency on social influence scores; and ii) interdependency analysis shows a synergic effect of combining data from communication and social networking service domains when it comes to calculating influence scores.

Keywords: social networking, Facebook, telecommunications, social influence, social network analysis, SmartSocial, user profiles
Categories: H.1.2, H.3.1, H.4.3, K.4, M.4

1 Introduction

Today, every person simultaneously participates in numerous social networks which span through various perspectives of our lives – family, friends, hobbies and work – to name just a few. Some of the mentioned social networks are formed in the physical world (e.g., a network where connections among actors represent family relationships) while others are virtual (i.e., the Facebook network where connections among actors represent acquaintances). Very often the same relationships exist in both domains (e.g., "physical" family relationship which connects brother and sister is mirrored into "virtual" relationship of Facebook siblings). In today's networked society where everybody and everything becomes inter-connected, it is very important to be able to identify *key actors* and *key relationships* in such a complex multi-layered eco-system.

A possible step towards this goal is designing a methodology and implementing algorithms which enable: i) joining multiple relations among actors across different social networks into the single unified social network; and ii) mining created unified social network for identification of key actors and key relationships. This paper will propose the SmartSocial Platform as a proof-of-concept Information Technology (IT) artefact with the described capabilities. In order to not just present the feasibility of an idea, but demonstrate its real-world instantiation, this paper will focus on the specific research challenge of identifying the most *influential* actors in a social network built through combining relationships among same actors in two different domains. Namely, we will analyze how relationships in the *communication domain* (which are proxied through real-world mobile phone communication data) correlate with relationships in the *social networking service domain* (which is proxied through real-world Facebook data). Based on the described analysis conclusions will be made whether analyzing additional multi-source data which characterizes connections among same actors in different social networks has an impact on identifying most influential actors. This is also one of the reasons how we have chosen the domains which are going to be analyzed – they partially overlap, but have their own specificities as well.

Original scientific contribution of this paper is twofold: i) *sensitivity analysis* of the algorithm for calculating user's influence in a social networking domain, aimed towards answering the research question whether type or frequency of social activities have more significant impact on the social influence score; and ii) *interdependency analysis* showing how relationships in the communication domain correlate with those in the social networking domain, aimed towards answering the research question whether analyzing multi-source data which characterizes connections among same actors in different social networks has an impact on identifying the most influential actors.

The paper is structured as follows. Section II gives an overview of related work on social influence. In Section III social influence will be defined in the context of the modern ICT user through the SmartSocial Influence Model. Section IV will describe algorithm implementations for calculating user's telco and social influence based on user profiles stored in the SmartSocial Platform. In Section V a real-world experiment of calculating user's SmartSocial Influence on 465 modern ICT users will be presented. Section VI will discuss the results of the SmartSocial Influence real-world experiment. Finally, Section VII concludes the paper and announces our future work.

2 Related work

Social influence is "a measure of how people, directly or indirectly, effect the thoughts, feelings and actions of others" [Turner 91] and has a broad potential business application which makes it an important part of a company's *decision support systems*. For example, from the business perspective it would be very beneficial for telecommunication operators to include such a measure in their churn prevention activities where they will proactively focus on the most influential subscribers to keep them satisfied because there is a high risk that their churn would have negative impact on other subscribers as well [Nadinic, Buzdon 05]. More

general example could include advertising industry where businesses would aim social media marketing campaigns directly towards most influential actors in the network because such an approach would enable the most efficient spreading of their messages throughout the whole network [Podobnik et al. 13]. Finally, the ability to calculate social influence would have potential positive impact on internal company processes as well because it could be used for detecting key employees [Humski et al. 13].

Social influence calculation has seen a great rise with services and algorithms such as Klout, Kred, PeerIndex or Tellagence, all of which have demonstrated the central role of empowered users in everyday lives of ordinary people. This is especially important when it comes to online marketing and commercial value of the peer-to-peer influence. Customer value initiative such as American Express campaign of rewarding users with a 10$ free credit for a branded tweet clearly shows the trend of redefining metrics related to customer importance [Huszar 13]. The fact that today consumers create approximately 20% of all brand impressions through social network services serves as a confirmation of that trend. Furthermore, empowered users who make up less than 10% of all social networking users and create 80% of these impressions [Bernoff, Schadler 10] serve as an additional proof why measuring user social influence is not just a challenging research topic but a highly-relevant business issue as well.

Certain similarities connect Klout and the proposed SmartSocial Influence Algorithm (SSIA). First, they both use a scale of 1 to 100 for social influence. Second, the Klout score is about quality, not quantity – having interactions with an influential individual can have a much larger impact on the score than interacting with a group of people all having lower influence. Third, adding more networks into calculation (e.g., Twitter and LastFM alongside Facebook) changes social influence score. The difference between Klout and the proposed SSIA is that SSIA enables utilization of an additional data source – telecommunication operator's network.

Unlike Klout, Kred uses a transparent, openly published algorithm and unlike SSIA, the influence scale ranges from 1 to 1000. Kred defines influence as "the measure of what others do because of you", similarly as SSIA does. Kred uses Facebook (or additionally, Twitter) as user data source and normalizes the total Kred Influence. Again, as opposed to SSIA it does not make use of telecommunication operator's network as a data source.

Social influence calculation based on the limited recursive algorithm (LRA) [Hajian, White 11] showed that network's structural information alone (e.g., number of friends a node has) cannot be used to predict social influence accurately; instead, the interactions between the nodes are of greater importance, as seen in the field of social recommenders [Ting et al. 12]. The proposed SSIA uses the LRA approach to mine social networks both from structural perspective as well as from the perspective of analyzing type (i.e., quality) and time-dimension (i.e., frequency) of interactions between users.

3 SmartSocial Influence Model

The second decade of the 21st century in the field of Information and Communication Technologies (ICT) is marked by the *modern ICT user* – a mobile user with numerous online accounts using a plethora of services on the go daily, such as *telecom network operator services* (e.g., mobile phone calls, text messaging or mobile Internet) and various *Internet services* (e.g., YouTube, Amazon or IMDb), including social networking services (e.g., Facebook, Twitter or Instagram) [see Figure 1]. Both telecommunication network operators as well as Internet service providers regularly store rich user profiles with a purpose specific to each individual service. An approach for modern ICT user profiling which combines telecommunication operator data with Internet services data is given in our previous work [Smailovic et al. 14b].

SmartSocial Platform (SSP) is a proof-of-concept IT artefact with following functionalities: i) joining multiple relations among actors across different social networks into a single social network; and ii) mining the created single social network for identification of key actors and key relationships. Consequently, SSP infers social influence as a new knowledge from multi-source information about users. *SmartSocial Influence (SSI)* is a function of user's influence in the communication domain – *Telco Influence (TI)* – and user's influence in the social networking service domain – *Social Influence (SI)*:

$$SSI = f(TI, SI) \qquad (1)$$

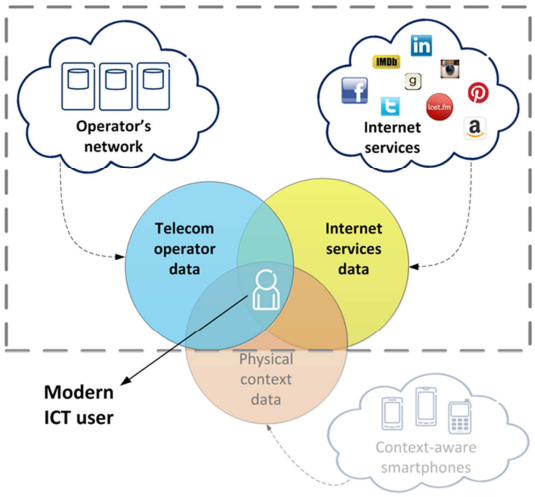

Figure 1: Modern ICT user – user data available from different sources

3.1 Telco Influence

Telco Influence Algorithm (TIA) takes Telco data as input and outputs Telco Influence score for each given user. Telco data is: i) amount of calls made or received in a certain period; ii) duration of those calls; and iii) amount of messages (SMSs) sent or received in a given period. The TIA 1.0 was based on an approach of identifying

influential users based on the amount of Telco data they generate (i.e., more is better) and is presented in more details in our previous work [Smailovic et al. 14a].

The TIA 1.0 had to be upgraded in order to take into account real-world constraints encountered during social experiment. Due to Android OS clipping the amount of possible calls in a call log to a maximum of 500 entries, the *TIA 2.0* takes *frequencies of calls (CF)*, their *duration frequency (DF)* and *frequency of SMSs (SF)* as input data instead of absolute values, respectively [see Figure 2]. This upgrade makes it possible to calculate the TI value correctly irrespective of the monitoring period in which the data was generated and makes the TIA more robust and general.

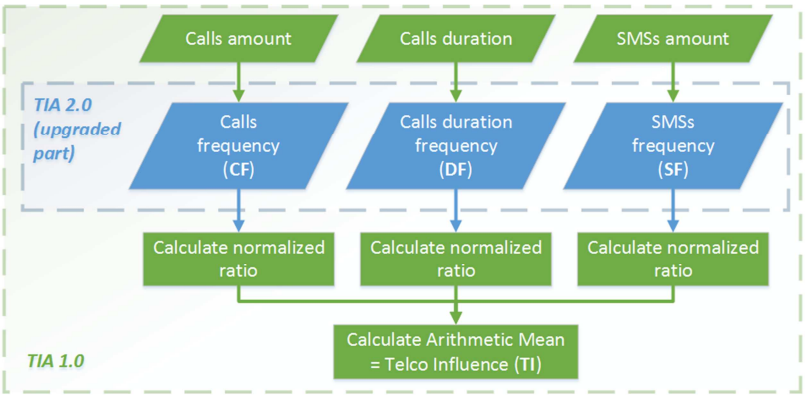

Figure 2: Telco Influence Algorithm 1.0 vs Telco Influence Algorithm 2.0

In order to have a maximum Telco Influence, the user would have to have the greatest frequencies of calls and SMSs, as well as the greatest calls duration frequency (i.e., amount of minutes in calls per day) compared to other users in the SSP database.

3.2 Social Influence

The Limited-Recursive Algorithm (LRA), which is based on the famous PageRank algorithm [Hajian, White 11], served as a basis for the *Social Influence Algorithm 1.0 (SIA 1.0)*, as described in our previous work [Smailovic et al. 14a]. The algorithm takes Facebook data as input and outputs Social Influence score for each given user. Used Facebook data is: i) *number of friends a user has (F)*; ii) *amount of posts the user has posted on his/her Wall (P)*; and iii) *amount of Likes/Comments on those posts (L)/(C)* [see Figure 3].

The *SIA 1.0* was upgraded to the *SIA 2.0* by adding two more steps. The SIA 1.0 used heuristically determined values for certain factors and weights, what was supported by the pre-experiment using real-world user data. However, based on post-analysis of the pre-experiment data as well as SIA 1.0 results, several upgrades to *SIA 1.0* were introduced. The main idea behind these upgrades was that user's *social influence* equals *content* combined with the *audience* which engages upon it. Plentiful *audience* which engages the abundant *content* results in a great *social influence.*

First, *number of friends (F)* a user has on Facebook is taken into account in the first step. In order to have a maximum *Friend Factor (FF)*, the user has to have a 1000 or more friends [Smailovic et al. 14a].

Second, *amount of Likes (L) and Comments on post (C)* taken from user's Facebook wall is used for calculating the *Audience Engagement Rate (AER)* of each post. This process is repeated a *number of posts (P)* times. One does not achieve a great *social influence* score by only having content; this content has to be *engaged* upon by the audience (i.e., user's friends) as well. The greater the number of Likes and Comments on user's posts, the greater the *AER* of that post. In order to achieve maximum *AER*, each post has to be engaged by at least 25% of users. While in the SIA 1.0 all types of posts are treated equally, in the *SIA 2.0* the *Post Type Factor (PTF)* is introduced for weighting photos, links and statuses differently.

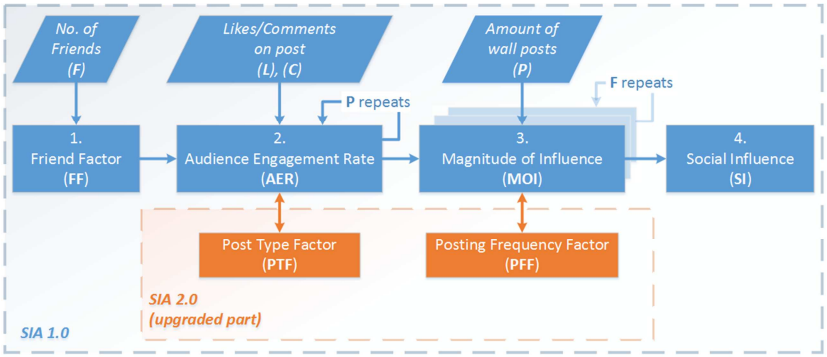

Figure 3: Social Influence Algorithm 2.0: An upgrade of the Social Influence Algorithm 1.0

Third, *amount of wall posts (P)* is used for calculating user's *Magnitude of Influence (MOI)*. *MOI* is a measure of user's content impact without help of his/her friends. It can be viewed as an averaged value of all posts' *AERs*. This is where the second SIA upgrade – *Posting Frequency Factor (PFF)* – is beneficial; the *SIA 1.0* does not take into account user's *posting frequency*. Since posting frequency determines post's impact [SocialBakers 11, TrackSocial 12], it is necessary to take it into account. Instead of solely using *amount* of posts, posting *frequency* is taken into account in the *SIA 2.0*. As will be shown in the sensitivity analysis later on, *PFF* greatly impacts the overall social influence scores distribution. In order to have a maximum *MOI*, user would have to have a maximum *AER* for any given post, and an optimal *posting frequency*, depending on the implementation.

Fourth and final, the *Social Influence (SI)* result (scaled from 0 to 100) is calculated by using a weighted sum of user's own *MOI*, together with his or hers averaged friends' *MOIs*. The process of calculating friend's *MOI* is repeated *F* times. This means that the more influential user's friends are – the more influential that user as well.

4 SmartSocial Influence Algorithms

While previous section explained the basic idea behind calculation of SmartSocial Influence, this section provides more details about TIA 2.0 and SIA 2.0 implementations.

4.1 Telco Influence algorithm implementation

Let us assume that the *monitoring period* (*mp*) is the larger value between the two – *Call log monitoring period* and *SMS log monitoring period*. The following pseudo code shows the exact implementation of the *TIA 2.0* [see Algorithm 1].

Data frequencies are *Calls Frequency (CF)*, *Calls Duration Frequency (DF)* and *SMSs Frequency (SF)*. After processing the *CF, DF* and *SF* the algorithm calculates normalized ratios for each of the data types – rescaling them by comparing them to the *minimum* and *maximum* respective values in the database. Finally, the arithmetic mean of those normalized ratios equals *TIA 2.0* score.

Algorithm 1: Telco Influence Algorithm 2.0 implementation pseudo code

```
Proc TIA-2.0 (User u ∈ Users, mp, Calls amount, Calls duration, SMSs amount)
  Proc calculateFreq (amount, period)
    frequency ← amount / period
    return frequency

  CF ← calculateFreq (Calls amount, mp)
  DF ← calculateFreq (Calls duration, mp)
  SF ← calculateFreq (SMSs amount, mp)

  Proc calcNormRatio (frequency)
    normalizedRatio ← ln (frequency – freqMin) / ln (freqMax – freqMin)
    return normalizedRatio

  TI ← (calcNormRatio(CF) + calcNormRatio(DF) + calcNormRatio(SF)) / 3
  return TI
```

4.2 Social Influence algorithm implementation

Let us assume the following definitions:

- *F* is the amount of friends the respective *user* has;
- *L* is the amount of distinct likes the respective *post* has;
- *C* is the amount of distinct comments the respective *post* has;
- *P* is the amount of posts the respective *user* has;
- *amountMin, amountMax* are the minimum and maximum respective *values* in the database.

Algorithm 2: Social Influence Algorithm 2.0 implementation pseudo code

```
Proc calculateMOI (User u ∈ Users, F, L, C, P)
  foreach post ∈ Posts do
    AER(post) ← sum(L ∪ C) / F · PTF(post)
  MOI(u) ← sqrt(sum(AER²) / P) · PFF(u)
return MOI(u)

Proc SIA-2.0 (User u ∈ Users, Friend f ∈ Friends)
  Proc calcNormRatio (amount)
    normalizedRatio ← ln (amount – amountMin) / ln (amountMax – amountMin)
  return normalizedRatio
  FF(u) ← calcNormRatio (F)
  foreach f ∈ Friends do
      friendSumMOI (f) ← friendSumMOI (f) + calculateMOI (f, F', L', C', P')
  friendsAverageMOI ← friendSumMOI (f) / F
  SI ← FF · MOI(u) + (1 – FF) · friendsAverageMOI
return SI
```

The F', L', C' and P' correspond to the assumed definitions, but of the respective *friend* instead of that of the *ego-user*. The following pseudo code shows the exact implementation of the *SIA 2.0* [see Algorithm 2].

In the *SIA 1.0*, functions *PTF*(*post*) and *PFF*(*u*) always returned the value *1.0*. In the upgraded *SIA 2.0*, they return values defined as the following.

4.2.1 Post Type Factor

Type of post determines its impact on the audience; a post can be a *link*, a *status* (i.e., text) and a *photo*. *PTF* distinguishes between posts according to Post Type [see Table 1] [TrackSocial 12].

	PTF value	
Post Type	***SIA 1.0***	***SIA 2.0***
Link	1.0	0.44
Status	1.0	0.68
Photo	1.0	2.88

Table 1: PTF values for different Post Types

4.2.2 Posting Frequency Factor

Posting Frequency Factor (PFF) can be approached in three different ways, as described below.

Sample-Literature-Optimal Posting Frequency factor (SLOF). *SLOF* is implemented with the following function [see Figure 4]. It is modelled according to the empirically-found optimal 5 to 10 posts per week [SocialBakers 11], combined with the empirically-found decreases in lower and greater values [TrackSocial 12].

Maximum *SLOF* value is 1.0 – if the user has between 5 and 10 posts per week. For values below 5, *SLOF* drops linearly to zero. For values above 10, *SLOF* drops linearly to a value of 42 posts per week, where it stagnates, as there is no further negative impact for over-posting. *SLOF* is an important basis for the remaining two *PFF* implementation variations, as its ratios between the function values are preserved throughout those variations.

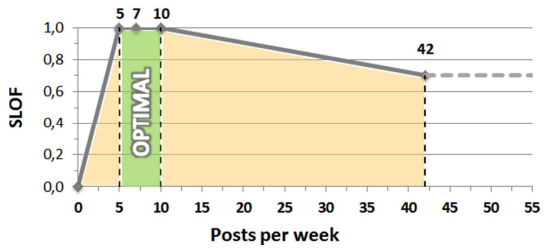

Figure 4: Function for Sample-Literature-Optimal Posting Frequency factor (SLOF)

Sample-Average-Optimal Posting Frequency factor (SAOF). *SAOF* is based on the *average* posting frequency the users in the real-world social experiment sample had, which is 1.5 posts per week [see Figure 9]. Since the ratio of *lower*, *optimal* and *greater thresholds* has to be preserved [see Figure 4], they equal to 1.1, 2.2 and 9, respectively.

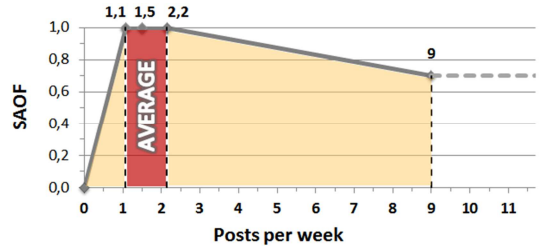

Figure 5: Function for Sample-Average-Optimal Posting Frequency factor (SAOF)

Sample-Median-Optimal Posting Frequency factor (SMOF). *SMOF* is based on the *median* posting frequency in the social experiment sample, which is 0.93 posts per week [see Figure 9]. Since the ratio of *lower*, *optimal* and *greater thresholds* has to be preserved [see Figure 4], they amount to 0.7, 0.9 and 5.6, respectively.

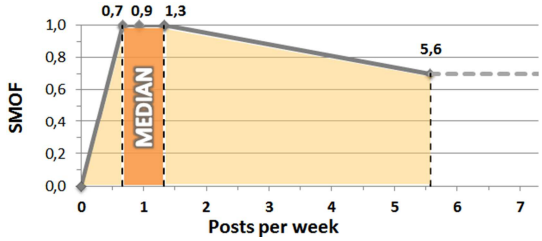

Figure 6: Function for Sample-Median-Optimal Posting Frequency factor (SMOF)

5 Real-world SmartSocial experiment

What is a SmartSocial Influence algorithm without the means to show its purpose? The experiments described below were conducted in order to collect real-world, actual data, with actual people, so the importance of Posting Type and Posting Frequency Factor as well as synergy between Telco and Social data could be analyzed.

5.1 Pre-experiment

The *SIA 1.0* and the *TIA 1.0* algorithms were implemented in the SmartSocial Platform by mid-2014 and were ready for input data. By participating in the pre-experiment, a total of 123 users provided their personal multi-source data from Facebook and Telco sources by using a website created for this sole purpose, as described in our previous work [Smailovic et al. 14a]. The pre-experiment was conducted in order to test algorithms' performances, as well as gain insight into the results they would produce regarding SmartSocial Influence scores.

The Facebook data contained over 5,000 posts in total. The average Posting Frequency was measured at 3.1 posts per week. The median Posting Frequency was much lower at 1.5 posts per week. *SmartSocial Influence (SSI)* was calculated according to the Formula (1) as an averaged sum of user's *Telco Influence (TI)* and *Social Influence (SI)*:

$$SSI = \frac{TI+SI}{2} \qquad (2)$$

Detailed description of the remaining pre-experiment results is thoroughly described in our previous work [Smailovic et al. 14a].

5.2 SmartSocial Platform

SmartSocial Platform (SSP) is a platform for context-aware social networking of modern ICT users, as described in our previous work [Smailovic et al. 14a, 14b]. It uses information-rich user profiles in order to provide new, inferred knowledge about information and communication service users. The *SmartSocial Influence* is one of examples of such inferred knowledge.

SSP is built on a thin-client heavy-backend principle. User first installs the *Android app* [Smailovic, Striga 15] in order to provide personal user data (Telco and Social data). Telco data is fetched from the Android smartphone itself. Social data is fetched from the *Facebook server* after accepting the usage terms through the *Android app*. Afterwards, the data is stored at the *Main server*, which is comprised of two components: i) *SIA 2.0* component for processing user data; and ii) *SmartSocial.eu* component which enables users to check their *SmartSocial Influence* score.

Android app [Smailovic, Striga 15] is available to all who wish to participate in the SmartSocial experiment. *Android app* comprises several steps which enable: i) inputting user's smartphone number; ii) providing Facebook data through login; iii) accepting usage terms of user data; iv) uploading Telco and Social data; v) receiving a notification of successful completion and the ability to view the created profile and *SmartSocial Influence* scores.

After successfully providing Telco and Social data, the user can view his or her collected data, as well as results on the *User Portal*. The generated user profile of the author Vanja Smailovic is accessible by using the profile password "255ev" at Login (www.smartsocial.eu/login.php).

5.3 Collected real-world sample analysis

The main experiment was conducted in the period from September 2014 until January 2015. A total of 465 user profiles were created. Out of those, 104 contained only Telco data, as these users did not provide their Facebook data. Real-world sample is comprised of the remaining 361 profiles with complete, personal multi-source data necessary for the algorithms to run – both Facebook, as well as Telco personal data. These were used for the sample analysis that follows.

The biggest node (i.e. user) or hub in the graph is the one with the biggest degree-centrality, i.e. amount of Facebook friendships [see Figure 7]. Color-coding is depicted for readability of the graph and represents modularity class, i.e. people that are interconnected together and form a group. The graph is undirected, as friendship on Facebook is a symmetrical relationship. Not surprisingly, the biggest node is the author Vanja Smailovic, which led the experiment and sought participants. As an effect, almost 20% of all 361 nodes are members of his ego-network, and as such, they do not represent a truly random sample. On the other hand, a somewhat biased sample is necessary for the purpose of interconnectedness; a truly random sample (of 361 out of 1.5 billion existing Facebook nodes) would not contain many edges as those users would not be friends on Facebook which was a prerequisite for the experiment.

Average number of Facebook friends in the sample is 483 with the median being 363, which means the majority of the participants (62%) are distributed below the average, while 35% of participants have more than 500 friends. This corresponds well to the relevant Facebook statistics [Smith 14].

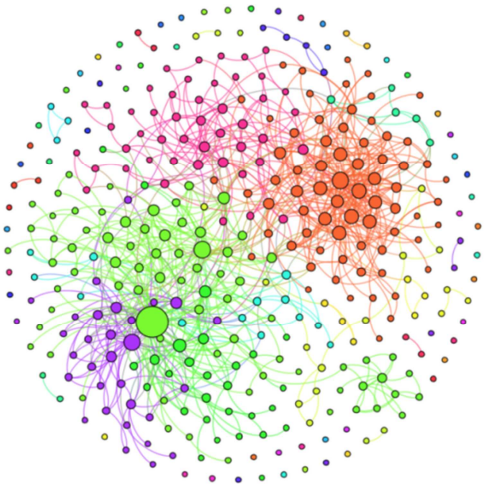

Figure 7: Graph depicting 361 anonymized SmartSocial users and their friendships

Comparing the real-world sample to the results of different models for generating graphs enables classifying the graph as being non-random, small-world network or even scale-free network. Important parameters for this classification are – the *degree distribution*, *average clustering coefficient* and *average path length*.

First, by analyzing the *degree distribution* of the real-world sample, it is observed that it resembles a power-law distribution. This is a strong indicator of small-world as well as scale-free networks. Random graphs do not display such distributions, but instead follow a Poisson distribution or similar. One of the reasons behind this is that they do not possess *hubs* (i.e., nodes with a large number of edges when compared to the rest of them). Out of 361 nodes, 75 are isolated and have a minimum degree of 0, meaning they do not have any edges connected to them; these are excluded from the graph classification as necessary for the graph-generating models to work as defined.

Second, by analyzing the *average clustering coefficient*, it is possible to determine whether the real-world sample graph is random or not. Random graphs, compared to small-world or scale-free networks exhibit very small *average clustering coefficients.* The reason behind this lies in the way the nodes connect to each other – each node has an independent, constant and random probability of connecting to another node. This is not so in real-world networks, which tend to exhibit *preferential attachment* and *growth* mechanisms when connecting nodes. The Barabasi-Albert model is able to generate a random scale-free network with these mechanisms in mind [Barabási, Albert 99]. Real-world networks (both small-world and scale-free included) exhibit high local node clustering and therefore have a much bigger *average clustering coefficient.*

Average clustering coefficient of the real-world sample graph equals 0.381. For observing graph's randomness, the Erdos-Renyi graph-generating model is used [Erdős, Rényi 59], which is able to generate a random graph with the same number of

nodes (286) and edges (890) for comparison. As expected, the measured *average clustering coefficient* of the Erdos-Renyi graph is 0.024 which is much lower than 0.381. The conclusion that the real-world sample is non-random holds true.

		Graph-generating model	
	SmartSocial experiment	**Erdos-Reny**	**Watts-Strogatz**
Generated graph type	-	Random	Small-world
Nodes	286	286	286
Edges	890	890	858
Degree distribution	Power-law	Poisson	-
Avg. clustering coeff.	0.381	0.024	0.393
Avg. path length	3.952	3.283	4.347
Conclusion for SmartSocial sample	→	is Non-random	is Small-world

Table 2: SmartSocial real-world experiment sample is non-random and small-world

Third, by analyzing the *average path length* together with the *average clustering coefficient,* one is able to distinguish small-world networks from the rest. For observing the small-world effect, the Watts-Strogatz (Beta) graph-generating model is used [Watts, Strogatz 98], which is able to generate a small-world graph with the same number of nodes (286) and similar number of edges (858) for comparison. The measured *average clustering coefficient* of the Watts-Strogatz graph is 0.393 which is very close to the sampled 0.381. The measured *average path length* of 4.347 is reasonably close to the sampled 3.952. Since the degree distribution resembles power-law, graph is not random and the *coefficients* are close to those of a small-world network – the conclusion that the social experiment sample is a small-network holds true.

In summary, the explanation for classifying the real-world social experiment sample as *non-random*, *small-world* network is given above [see Table 2].

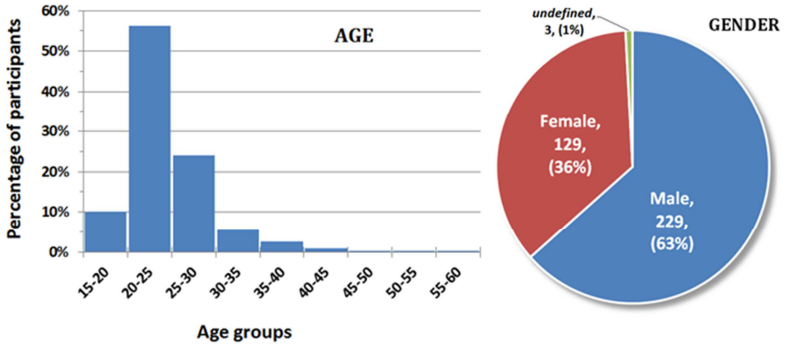

Figure 8: Age and gender distribution in the social experiment sample

Some interesting facts about the social experiment sample include *Age* and *Gender* distributions [see Figure 8]. Most of the participants were adults between 20 and 25 years of age, with the average being 25. Most of them were male (63%), with the rest being female (36%) and undefined (1%), meaning they did not enter gender data into their Facebook profile.

Total amount of Facebook posts from 361 participants was 12,074. Amount of Likes and Comments per post follows a power-law distribution as expected. Majority of the posts are rarely liked or commented upon, e.g. more than 80% of posts have less than 30 likes. Same is true for Comments as more than 95% of posts have less than 15 comments. It is interesting to note that posts usually have more Likes (max. value 189) than Comments (max. value 124).

Optimal Posting Frequency equals between 5 and 10 posts per week, with 1 post per day being optimal for engagement [SocialBakers 11, TrackSocial 12]. Only 6.6% of social experiment participants had this amount of average posts per week, while majority of them posted less than once per week. Average number of posts per week amounted to 1.5 with the median being 0.93 [see Figure 9].

This is quite different from the *pre-experiment* average of 3.1 and median of 1.5. Two different social experiments led to two different real-world samples; it could easily be the age difference that produced such discrepancy (as younger population, sampled in the pre-experiment, tends to post more frequently).

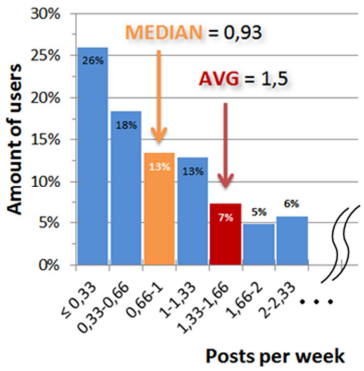

Figure 9: Average and Median Posting Frequency per week in real-world sample

6 Experiment results, impact and contribution

Experiment results underline two important paper contributions: i) Posting Frequency sensitivity analysis which shows the significant effect of Posting Frequency on Social Influence scores; and ii) interdependency analysis which shows synergy between Telco and Social data-sources when it comes to total SmartSocial Influence scores.

The basic *SIA 2.0* produced the initial results for further benchmarking [see Figure 10]. This algorithm, unlike the upgraded Social Influence algorithms, does not take into account the *Post Type factor (PTF)* or any kind of *Posting Frequency Factor*

(PFF). This results in a bimodal distribution which is very wide; the median value is 39, with the average being 41.

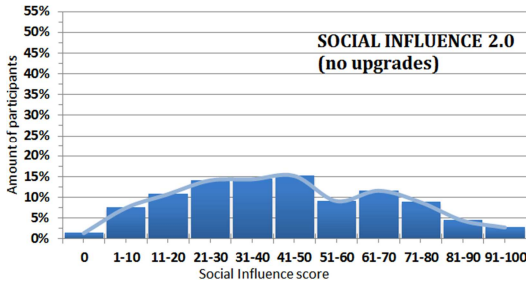

Figure 10: Social Influence initial distribution (basic Social 2.0 algorithm)

6.1 Post Type Factor sensitivity analysis

Adding the first upgrade *PTF* to the basic *SIA 2.0* resulted in a minor change in the results [see Figure 11]. Social Influence distribution saw a minor rise in values; average 48, median 48. This is due to the fact that most of the posts participants hold are Photos, which the *PTF* boosts over Statuses or Links in calculating the *Social Influence (SI)* values. It is important to notice that the similarity between the basic *SIA 2.0* and *PTF-enabled* distributions does not imply similarity in their respective *SI* values per user. On the contrary, *PTF* re-ranks the *SI* values of the users by taking the type of each of their posts into account.

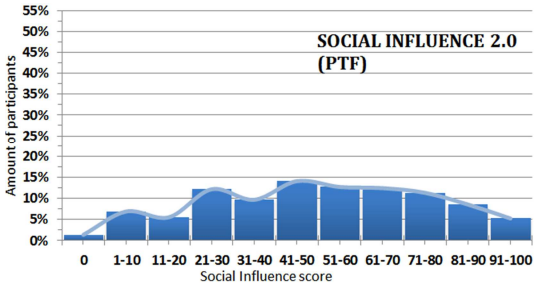

Figure 11: Post Type Factor (PTF) does not impact the initial distribution greatly

6.2 Posting Frequency Factor sensitivity analysis

Posting Frequency sensitivity analysis shows the significant impact of the *Posting Frequency Factor (PFF)* on the final *Social Influence (SI)* value for a given user.

As expected, *SLOF* impacts the initial distribution of the basic *SIA 2.0* greatly, much more than the *PTF* [see Figure 12]. This is due to the fact that the majority of users did not have the *optimal* posting frequency (5 to 10 posts per week). Therefore, the *SI* distribution is highly skewed to the left, being very "strict" towards users with below-optimal posting frequency. The average *SI* is 13 with median at 7.

The final *SI* distribution includes both upgrades to the basic *SIA 2.0* – it is *PTF-and-SLOF-enabled*. This distribution is very similar to the *SLOF-enabled* distribution, with an average *SI* of 15.8 and median of 8 [see Figure 12]. As it can be seen, not even *PTF* can "help" push the distribution to the right; the algorithm is still too "strict" due to majority of users being less-than-optimal when it comes to posting frequency.

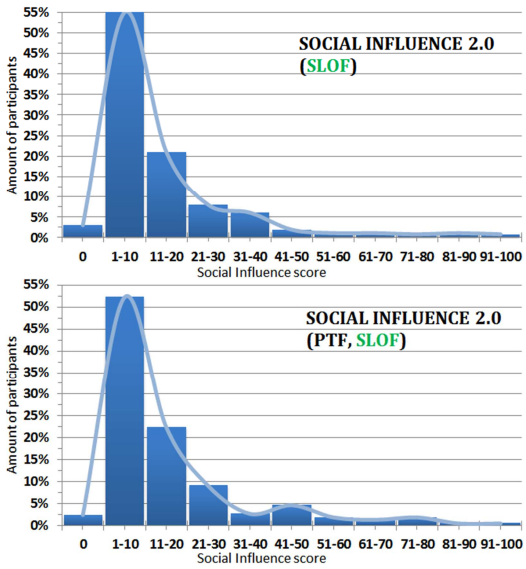

Figure 12: SLOF and PTF with SLOF upgrades

SLOF introduced a great change to the *SI* values given by the initial basic *SIA 2.0* – but *SLOF* is not usable with the real-world sample gathered through our social experiment, as it imposes unsuitably "strict" Literature-Optimal posting frequency values. What would happen if the *lower*, *optimal* and *greater threshold* values for posting frequency were taken from the measured *average* or *median* posting frequencies in the real-world social experiment sample?

As expected, *SAOF* is much more usable on the real-world social experiment participants sample than *SLOF*. Average *SI* is 37.9 and median is 30 [see Figure 13]. Still, *SAOF-enabled* algorithm produces a distribution that is more "strict", skewed-to-the-left than the basic *SIA 2.0* does [see Figure 10].

Finally, *SMOF* is the most usable on the real-world social experiment participants' sample, compared to *SLOF* and *SAOF*. Average *SI* is 39.5 and median is 35 [see Figure 13]. Furthermore, *SMOF* is the only *PFF* upgrade that produces a distribution that is similar to the basic *SIA 2.0* distribution [see Figure 10].

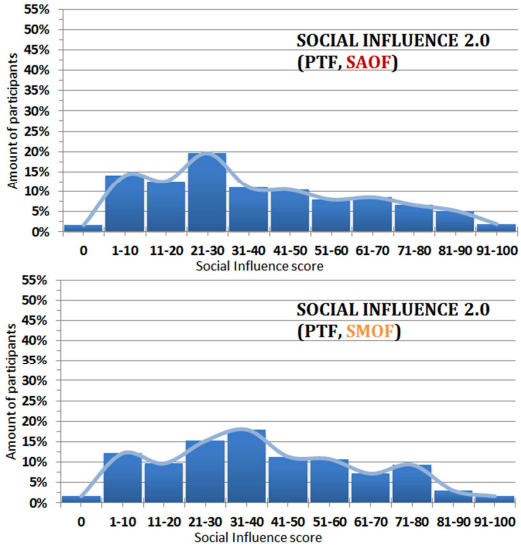

Figure 13: Social Influence final distribution (PTF and SAOF, PTF and SMOF)

It is important to notice that the similarity between the basic *SIA 2.0* and *SMOF-enabled* distributions does not imply similarity in their respective *SI* scores per user. On the contrary, *SMOF* re-ranks the *SI* scores of the users by taking their Posting Frequency into account.

	Posting Frequency factor		
Social Influence	**SLOF**	**SAOF**	**SMOF**
Ranked 1st	Person A	Person B	Person C
Ranked 2nd	Person B	Person A	Person B
Ranked 3rd	Person C	Person C	Person A

Table 3: Social Influence rankings change when utilizing SLOF, SAOF and SMOF

More interestingly, there is a difference in *SI* rankings of users between the *SLOF*, *SAOF* and *SMOF*. The same top three influencers (Person A, Person B and Person C) exchanged *SI* rank positions throughout the *PFF* sensitivity analysis [see Table 3].

In conclusion, this sensitivity analysis clearly showed that the differences between characteristics of the sample (namely *average* and *median* values of posting frequencies) result in a sample-dependent outcome regarding final *SI* values. More precisely, the differences in posting frequency values between the pre-experiment and the main real-world social experiment sample prove it is important to take them into

account when performing the *SI* calculation. Without adjusting the *PFF* according to the input sample, one would produce unusable final *SI* values.

6.3 Synergy of Telco and Social data

Telco Influence Algorithm 2.0 (TIA 2.0) final distribution [see Figure 14] shows an average of 40.9 and median of 40, a result which is very usable.

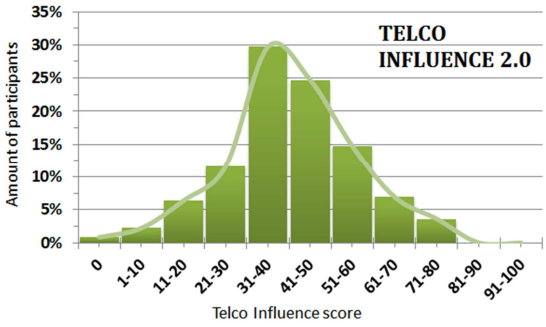

Figure 14: Telco Influence final distribution

Interestingly, the maximum *TI* value is 80, due to the fact that the user with the highest amount of Calls is not the one with the highest duration of Calls or amount of SMS messages – and vice-versa. In order to have *TI* value of 100, one needs to have the greatest amount of all three measured data.

Two scatterplot analyses of the *SI* and *TI* values are depicted below [see Figure 15]. Left-hand side shows basic *SIA 2.0* values [see Figure 10], while the right-hand side shows results of the *PTF-and-SLOF-enabled* values [see Figure 12].

It is visible that the values fill the scatterplot plot uniformly (in accordance with their respective distributions), without forming any kind of pattern, showing that *SI* does not enable prediction of *TI*. Situation in the scatterplot remains the same regardless of the variations in utilizing *PTF*, *SLOF*, *SAOF* or *SMOF*, respectively.

The non-ranked scores were thought of in the vein of the Pearson coefficient [Pearson 1895]. Similarly, the ranked scores were inspired by the Spearman coefficient [Spearman 1904]. Non-ranked scores are "absolute" and equal to the ones the respective algorithm produces. Ranked scores are "relative" and produced by ranking the "absolute" scores that the respective algorithm produced. Mathematically speaking, ranking produces monotonic relationships between the values, while non-ranking gives true values. Ranked values range from 1 to 361 (i.e., the number of participants in the real-world social experiment sample). Since the Spearman coefficient as a non-parametric test does not assume linearity and homoscedasticity of measured values as Pearson does, it is possible to plot a regression line together with the R-squared value.

Once more, the different approaches in the basic *SIA 2.0* and *PTF-and-SLOF-enabled* algorithms did not produce any significant changes in the scatterplot. The R-squared values are extremely small, showing that *SI* and *TI* scores weakly fit the linear regression model, meaning they are not linearly-interdependent. Together with

the linear regression, several other regression models were employed, namely exponential, logarithmic, polynomial (of the 2nd, 3rd, 4th, 5th and 6th degree) and power. Out of those, going through basic *SIA 2.0*, then the *PTF*, *SLOF*, *SAOF* and *SMOF* variations, the maximum R-squared was produced by the polynomial regression (of the 6th degree) each time, and equaled no more than 3.8% in any given case.

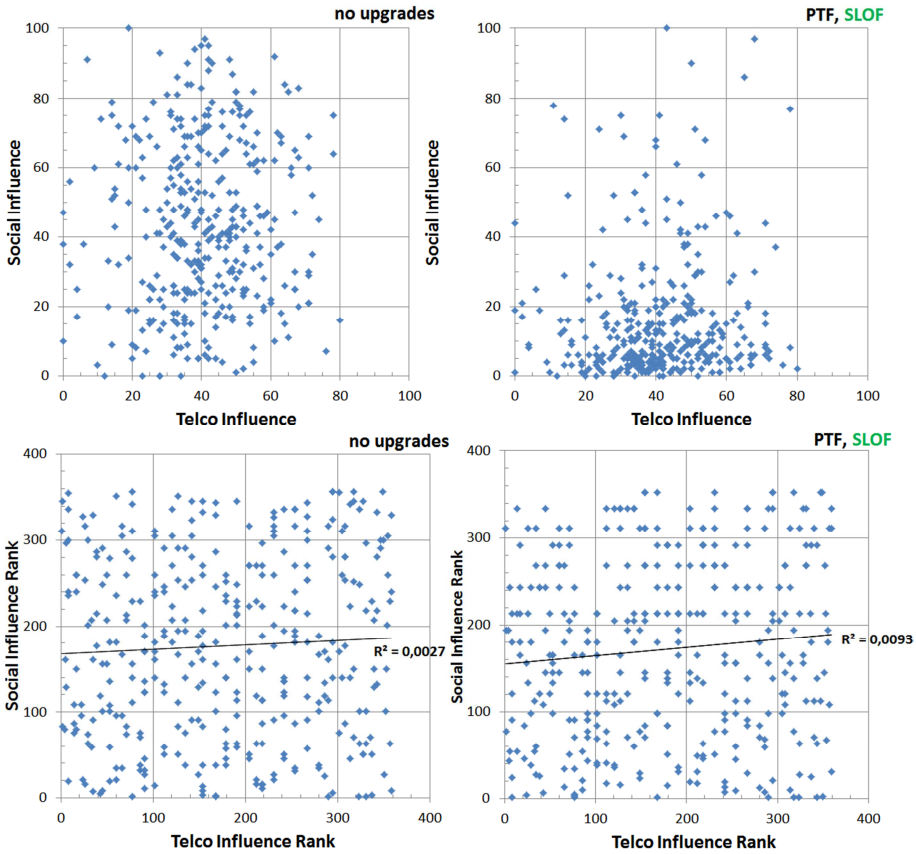

Figure 15: Scatterplots of Social vs. Telco Influence values (non-ranked and ranked)

The second contribution of this paper is the proof of synergy of Telco and Social data. The power of this synergy is clearly observed through the scatterplots [see Figure 15]. Statistically, all of the scatterplots (even the ones not depicted) clearly show that, if one knows the *SI* of a user, there is a *very small probability* of reliably predicting her/his *TI*. Vice-versa is true as well. This unambiguously means that one should utilize *both* Telco and Social data-sources in order the have the complete, final view on the user's social influence. The added value that emerges from adding Social data in the process of SmartSocial Influence calculation is evident if we compare social influence score distributions in four different scenarios where: i) SSI = f(TI); ii) SSI 1 = $f(TI, SI_{SLOF})$; iii) SSI 2 = $f(TI, SI_{SAOF})$; and iv) SSI 3 = $f(TI, SI_{SMOF})$ [see

Figure 20]. Although distributions of social scores when utilizing SAOF and SMOF resemble those when using only Telco data – significant differences in individual scores and rankings exist [see Table 3], what again supports the need to take both domains into account.

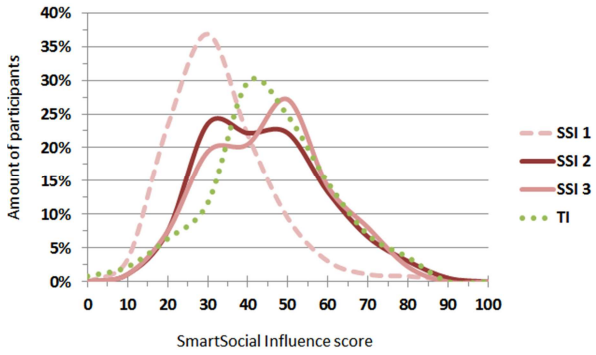

Figure 16: Three variations of the SSI vs. TI distribution

This result has a potential significant impact on design and implementation of company's *decision support systems*. Let us demonstrate this potential on the already mentioned example of telecommunication operators churn prevention activities. The question is how calculating user's social influence through the SmartSocial Platform (SSP) provides added value (when compared with traditional business practices) for telecommunication operator wanting to reduce subscriber churn. Using traditional practice (only telco data), the telecommunication operator would try to predict specific subscribers with high probability of leaving. If they are "important users" (e.g., post-paid business users that generate a lot of revenue), this group of subscribers becomes targeted for subscriber retention. On the other hand, a typical pre-paid telco subscriber merely comprises a "long tail" of a revenue distribution. However, some of those subscribers have great *social influence* – they are omnipresent, frequent and important users of online *social networking services*. If any of them churn and switch to other competing telecommunication service provider, this information might cause significant churn of the remaining subscribers as well. If using traditional business practices, telecommunication operators would fail to retain such a user. However, if they employ approach based on calculating user's social influence through the SSP (combining data from two domains – both Telco and Social), identification of such a user will become possible.

7 Conclusions and Future Work

This paper presented several upgrades to the literature-based algorithm for calculating social influence and analyzed their impact on the overall SmartSocial Influence results. The most important upgrade is introduction of the Posting Frequency Factor. Through the real-world social experiment with 465 participants, we have confirmed the major impact its introduction made on the final distribution of user *social*

influence scores, as well as user rankings based on those scores. Furthermore, the analysis of Telco Influence and Social Influence results showed that it is very improbable to predict one by knowing the other, confirming that it is necessary to utilize *both* Telco and Social data sources in order to have a full view of the user's *social influence*. This confirms the synergy of Telco and Social data in calculating user's *social influence*.

In future research, the authors consider adding more input data and parameters into the sensitivity analysis in order to evaluate their interdependence. Furthermore, the posting frequency curves might be modelled differently (i.e., using non-linear functions) and impact of such changes observed. The Telco Influence algorithm implementation upgrade is to be considered as well, with the goal of bringing it closer to the real-world telecommunication operators' algorithms for churn prevention; results are to be observed and compared to the original version of the algorithm.

Acknowledgements

The authors acknowledge support of research projects "Managing Trust and Coordinating Interactions in Smart Networks of People, Machines and Organizations", funded by the Croatian Science Foundation under the grant UIP-11-2013-8813; "Ericsson Context-Aware Social Networking for Mobile Media", funded by the Unity through Knowledge Fund; and "A Platform for Context-aware Social Networking of Mobile Users", funded by Ericsson Nikola Tesla. Furthermore, the authors would like to thank all participants who installed the SmartSocial Android app and provided their personal data which contributed to this research.

References

[Barabási, Albert 99] Barabási, A.-L. and Albert, R. 1999. "Emergence of Scaling in Random Networks." Science 286 (5439): 509–12.

[Bernoff, Schadler 10] Bernoff, J. and Schadler, T. 2010. "Peer Influence Analysis." In Empowered: Unleash Your Employees, Energize Your Customers, and Transform Your Business, 37–56. Boston, Massachusetts: Harvard Business School Publishing Corporation.

[Erdős, Rényi 59] Erdős, P. and Rényi, A. 1959. "On Random Graphs I." Publicationes Mathematicae (Debrecen) 6: 290–97.

[Hajian, White 11] Hajian, B., and T. White. 2011. "Modelling Influence in a Social Network: Metrics and Evaluation." In 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 497–500. doi:10.1109/PASSAT/SocialCom.2011.118.

[Humski et al. 13] Humski, L., Striga, D., Podobnik, V., Vrdoljak, B., Banek, M., Skocir, Z. and Lovrek, I. 2013. "Building Implicit Corporate Social Networks: The Case of a Multinational Company." In Proceedings of the 12th International Conference on Telecommunications. Zagreb, Croatia: IEEE.

[Huszar 13] Huszar, F. 2013. "With Tweet-to-Buy, American Express Values Its Community at $10." Harvard Business Review. https://hbr.org/2013/02/with-tweet-to-buy-american-exp.

[Nadinic, Buzdon 05] Nadinic, B. and Buzdon, R. 2005. "New Possibilities for Knowledge Discovery in Telecommunication Companies." In Proceedings of the 8th International Conference on Telecommunications. Zagreb, Croatia: IEEE.

[Pearson 1895] Pearson, K. 1895. "Note on Regression and Inheritance in the Case of Two Parents." Proceedings of the Royal Society of London (1854-1905) 58: 240–42.

[Podobnik et al. 13] Podobnik, V., Ackermann, D., Grubisic, T. and Lovrek, I. 2013. "Web 2.0 as a Foundation for Social Media Marketing: Global Perspectives and the Local Case of Croatia." In Cases on Web 2.0 in Developing Countries: Studies on Implementation, Application, and Use, 342–79. Hershey : IGI Global.

[Smailovic, Striga 15] Smailovic, V., Striga, D. 2015. "SmartSocial.eu." http://smartsocial.eu.

[Smailovic et al. 14a] Smailovic, V., Striga, D., Mamic, D. P. and Podobnik, V. 2014. "Calculating User's Social Influence through the SmartSocial Platform." In Proceedings of Software, Telecommunications and Computer Networks Conference (SoftCOM), 383–87. Split, Croatia: IEEE.

[Smailovic et al. 14b] Smailovic, V., Striga, D. and Podobnik, V. 2014. "Advanced User Profiles for the SmartSocial Platform: Reasoning upon Multi-Source User Data." In ICT Innovations 2014 Web Proceedings. Ohrid, Macedonia: Springer.

[Smith 14] Smith, A. 2014. "6 New Facts about Facebook." PewResearchCenter. http://www.pewresearch.org/fact-tank/2014/02/03/6-new-facts-about-facebook/.

[SocialBakers 11] SocialBakers. 2011. "How Often Should You Post on Your Facebook Pages?" http://www.socialbakers.com/blog/147-how-often-should-you-post-on-your-facebook-pages.

[Spearman 1904] Spearman, C. 1904. "The Proof and Measurement of Association between Two Things." The American Journal of Psychology 15 (1): 72.

[Ting et al. 12] Ting, I-Hsing, Pei Shan Chang, and Shyue-Liang Wang. 2012. "Understanding Microblog Users for Social Recommendation Based on Social Networks Analysis." Journal of Universal Computer Science 18 (4): 554–76. doi:10.3217/jucs-018-04-0554.

[TrackSocial 12] TrackSocial. 2012. "Optimizing Facebook Engagement." http://tracksocial.com/docs/tracksocial_whitepapers_optimizingfacebookengagement.pdf.

[Turner 91] Turner, J. C. 1991. Social Influence. Wadsworth Publishing.

[Watts, Strogatz 98] Watts, D. J. and Strogatz, S. H. 1998. "Collective Dynamics of 'small-World' Networks." Nature 393 (6684): 440–42. doi:10.1038/30918.