Investigating Information Diffusion in a Multi-Social-Network Scenario via Answer Set Programming

Giuseppe Marra[1], Antonino Nocera[1], Francesco Ricca[2], Giorgio Terracina[2], and Domenico Ursino[1]

[1] DIIES, University Mediterranea of Reggio Calabria, Via Graziella, Località Feo di Vito, 89122 Reggio Calabria, Italy
[2] Dipartimento di Matematica, University of Calabria, Via Pietro Bucci, 89136 Rende (CS), Italy

Abstract. Information Diffusion is a classical problem in Social Network Analysis, where it has been deeply investigated for single social networks. In this paper, we begin to study it in a multi-social-network scenario, where many social networks coexist and are strictly connected to each other, thanks to those users who join more social networks. In this activity, Answer Set Programming provided us with a powerful and flexible tool for an easy set-up and implementation of our investigations.

1 Introduction

Information Diffusion has been largely investigated in Social Network Analysis [5,7,8,10,11]. However, all the investigations about this problem performed in the past analyze single social networks, whereas the current scenario is multi-social-network [2,4]. Here, many social networks coexist and are strictly connected to each other, thanks to those users who join more social networks, acting as bridges among them. But, what happens to the Information Diffusion problem when passing to this new scenario? New aspects must be taken into account and new considerations are in order. In this paper[1] we investigate the problem of Information Diffusion in a multi-social-network scenario (MSNS, for short). For this purpose, first we propose a graph-based model for an MSNS. This model takes into account the existence of more social networks, as well as the presence of bridges and topics of interest for MSNS users. Then, we provide a formal definition of the Information Diffusion problem in an MSNS. In order to implement our approach, and perform an analysis on real world data, we applied Answer Set Programming [1,6] (ASP). ASP is an ideal framework for the rapid development and implementation of programs solving complex problems [9] given its declarativity, expressive power, and availability of efficient ASP systems [12]. After describing the ASP specification solving the Information Diffusion problem, we also present the results of an experimental campaign conducted on an

[1] A preliminary version was submitted to SEBD 2014, which has informal proceedings.

R. Kontchakov and M.-L. Mugnier (Eds.): RR 2014, LNCS 8741, pp. 191–196, 2014.

MSNS based on four social networks, namely LiveJournal, Flickr, Twitter and YouTube. As far as the system for running our logic program is concerned, we used the ASP system DLV [12]. Our experimental campaign allowed us to draw an identikit of the best nodes for spreading information in an MSNS.

2 Our ASP-Based Approach to Information Diffusion in an MSNS

An MSNS Ψ, consisting of n social networks $\{S_1, S_2, \ldots, S_n\}$, can be modeled by a pair $\langle G, T\rangle$. Here, T is a list $\{t_1, t_2, \ldots, t_p\}$ of topics of interest for the users of Ψ. It is preliminarily obtained by performing the union/reconciliation of the topics related to the social networks of Ψ. G is a graph and can be represented as $G = \langle V, E\rangle$. V is the set of nodes. A node $v_i \in V$ represents a user account in a social network of Ψ. $E = E_f \cup E_m$ is a set of edges. E_f is the set of friendship edges; E_m is the set of `me` edges. An edge $e_j \in E$ is a triplet $\langle v_s, v_t, L_j\rangle$. v_s and v_t are the source and the target nodes of e_j, whereas L_j is a list of p pairs $\langle t_{j_k}, w_{j_k}\rangle$, where t_{j_k} is a topic and w_{j_k} is a real number between 0 and 1 representing the corresponding weight. This weight depends on both t_{j_k} and the ability of the user associated with v_t to propagate, to the user associated with v_s, information related to t_{j_k}.

Thus, an MSNS models a context where several social networks coexist and are strictly connected to each other, thanks to those users who join more social networks. Indeed, when a user joins more social networks, her multiple accounts allow these networks to be connected. We call *bridge user* each user joining more social networks, *bridge (node)* each account of such a user and `me` *edge* each edge connecting two bridges.

The Information Diffusion problem in an MSNS takes as input: *(i)* An MSNS Ψ, consisting of n social networks $\{S_1, \ldots, S_n\}$. *(ii)* A list D of n elements. The generic element D_h of D consists of a tuple $\langle S_h, p_h, c_h\rangle$. Here, p_h denotes the priority of S_h and is an indicator of the relevance of this social network in Ψ. It is an integer from 1 to n, where 1 (resp., n) is the maximum (resp., minimum) priority. c_h is the minimum desired coverage for S_h, i.e., the minimum number of nodes of S_h which must be reached by the information to spread throughout Ψ. *(iii)* A list τ of q elements. The generic element $\tau[k]$ of τ is a pair $\langle t_k, \omega_k\rangle$. Here, t_k corresponds to the k^{th} element of the set of topics T of Ψ. ω_k is a real number, belonging to the interval $[0, 1]$ and indicating the weight of t_k in the information to spread throughout Ψ. The Information Diffusion problem in Ψ requires to find the minimum set of the nodes of Ψ allowing the maximization of the coverage of the social networks of Ψ, taking into account the minimum allowed network coverage, the network priorities (as expressed in D), and the topics characterizing the information to spread (as expressed in τ).

Our ASP-based solution of this problem is based on the following guidelines. First, a support graph $G' = \langle N', E'\rangle$ is constructed starting from G. Specifically, there is a node $n' \in N'$ for each node $n_i \in N$, and an edge $e'_j = \langle v_s, v_t, w_j\rangle \in E'$ for each edge $e_j = \langle v_s, v_t, L_j\rangle \in E$. w_j is obtained as: $w_j = \sum_{k=1}^{p} w_{j_k}\omega_k$. In other

words, w_j measures the relevance of e'_j in the current Information Diffusion activity. It depends on both the importance of each topic of T in the information to spread throughout Ψ and the ability, of the user associated with v_t, to propagate, to the user associated with v_s, the topics of T. The Information Diffusion model adopted in our approach is the well known Linear Threshold one [8]. In this model, a node is considered *active* if the sum of the weights of the friendship edges from it to the already active nodes is higher than a certain threshold. In an MSNS, the Linear Threshold model must be extended in such a way as to consider `me` edges. Given a `me` edge from a bridge b_s to a bridge b_t, our extension requires that the information to spread propagates from b_s to b_t if the weight associated with the `me` edge is higher than a certain threshold (different from the one concerning friendship edges). Starting nodes are randomly selected. However, since in an MSNS the number of bridges is extremely low [3] w.r.t. the number of non-bridges, we first state the percentage of bridges and non-bridges which must be present in the set of starting nodes, and then randomly select the two kinds of node accordingly.

The problem we are considering is extremely complex. Since this paper represents a first attempt of investigating the possibility of using ASP in this context, in the following we perform some simplifications. Specifically, we assume that all the topics of Ψ have the same weight in the information to spread; this implies the removal of the topic dependency of the problem which, thus, becomes only structural. The important consequence of this choice is that all the friendship edges in G' have the same weight (we assign a weight equal to 1 to them). As for `me` edges, we assume that all of them always propagate the information to spread. Also in this case, the consequence is that we assign a weight equal to 1 to all `me` edges. A final simplification regards the node activation policy. In fact, we assume that a node is activated when at least two edges, outgoing from it, are pointing to already activated nodes. This corresponds to set the threshold to 2 in the Linear Threshold Information Diffusion model mentioned above.

It is important to stress that the adoption of ASP allowed an easy and fast set-up of the approach implementation, while attaining acceptable performances. The logic program designed to solve our problem is as follows (see [1] for a nice introduction to ASP):

```
1. in(X) v out(X) <- starting_node(X).
2. active(X) <- in(X).
3. active(X) <- active(X1),edge(X,X1,me).
4. active(X) <- edge(X,X1,friendship),edge(X,X2,friendship), X1!=X2,
          active(X1),active(X2).
5. hasActiveNodes(Sn)<-node(N,Sn),active(N).
6. <- D(Sh,Ph, Ch), Ch!=0, not hasActiveNodes(Sh).
7. <- D(Sh,Ph,Ch), #count{N:active(N),node(N,Sh)}<Ch.
8. <∼ in(X). [1:2]
9. <∼ node(N,Sh), not active(N). [1:1]
```

Here, the input is given as a set of facts of the form `edge(X,X1,K)` modeling edges from `X` to `X1`, where `K` specifies the edge kind (`me` or `friendship`); `node(N,Sn)` denotes the set of nodes in the social network S_n; `starting_node(X)` is the set of starting nodes. Finally, `D(Sh,Ph,Ch)` identifies the desiderata. Rule `1.` guesses the nodes that must be selected in the best solution. Rules `2.` to `4.`

Table 1. Effectiveness of our approach

Number of Social Networks	2	3	4
Percentage of activated nodes	70%	81%	81%
Number of necessary starting nodes	2	3	3

compute active nodes, based on the guess; in particular, a node is active if either it is a selected one, or it reaches an active node through a `me` edge, or, finally, it reaches two active nodes through friendship edges. Constraint rules `5.` to `7.` impose admissibility conditions, as specified in D. Weak constraint rules `8.` and `9.` implement the optimality requirements for consistent solutions, so that the minimum sets of nodes providing consistent solutions are identified first (`8.`), and, among them, the ones minimizing non-active nodes are selected (`9.`).

3 Experimental Campaign

To test our Information Diffusion approach we performed an experimental campaign on an MSNS consisting of four social networks, namely LiveJournal, Flickr, Twitter and YouTube. Our MSNS has 93177 nodes and 146957 edges. All the corresponding data can be downloaded from the following address: `www.ursino.unirc.it/RR2014.html`. The password the Referee must specify is "85749236". We performed a large number of runs of our ASP program. In these runs we considered different configurations of the starting nodes. They differed for the number of nodes, the percentage of bridges (this, very important, parameter ranged from 0 to 100 with a step of 10), and the number of the social networks to cover (this number ranged from 2 to 4). We constructed more sets of starting nodes for the same configuration in such a way as to reduce the influence of possible outliers. The whole number of runs we have performed was 576. These runs allowed us to carry out several investigations, the most significant of whom are reported below (the other ones cannot be shown due to space limitations).

In a first test we computed the percentage of the nodes of our MSNS activable by our approach that were really activated by it, as well as the number of nodes necessary for this activation, against the number of social networks to cover. The corresponding results are shown in Table 1. This table evidences that our approach is really effective. Interestingly, 3 starting nodes are sufficient to cover 4 social networks. This could not have been obtained without the presence of bridges. In fact, without this kind of node, at least 8 nodes would have been necessary to cover 4 social networks.

In a second test we computed the variation of the average percentage of bridges present in the optimal solution of runs against the variation of the average percentage of bridges present in the sets of starting nodes. Obtained results are

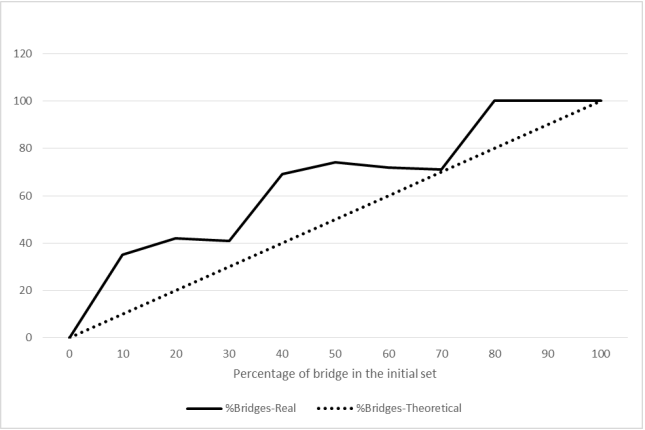

Fig. 1. Average percentage of bridges in the optimal solutions

Table 2. Composition of optimal solutions

(i)	*(ii)*	*(iii)*	*(iv)*	*(v)*	*(vi)*	*(vii)*	*(viii)*	*(ix)*	*(x)*
75%	24%	86%	22%	75%	86%	0 %	11%	14%	87%

shown in Figure 1. Observe that the percentage of bridges in the optimal solutions is generally higher, or much higher, than the percentage of bridges in the sets of starting nodes. This information is precious for drawing an identikit of the most influential nodes for Information Diffusion in an MSNS.

In a third test we computed the following statistics about the composition of optimal solutions: *(i)* average percentage of bridges; *(ii)* average percentage of the direct neighbors of bridges; *(iii)* average percentage of power users; *(iv)* average percentage of the direct neighbors of power users; *(v)* average percentage of nodes being both bridges and power users; *(vi)* average percentage of nodes being bridges or power users; *(vii)* average percentage of nodes being bridges but not power users; *(viii)* average percentage of nodes being power users but not bridges; *(ix)* average percentage of nodes being neither bridges nor power users; *(x)* average Jaccard coefficient[2] of bridges and power users. The corresponding results are reported in Table 2. From the analysis of this table we can observe that 99% of the nodes in the optimal solutions are either bridges or direct neighbors of bridges. Analogously, 98% of the nodes in the optimal solutions are either power users or direct neighbors of power users. Furthermore, all the bridges involved in the optimal solutions are power users, and almost all the power users involved in the optimal solutions are bridges. Finally, only a little fraction of the nodes present in the optimal solutions are neither bridges nor power users.

[2] We recall that the Jaccard Coefficient $J(A, B)$ between two sets A and B is defined as $J(A, B) = \frac{A \cap B}{A \cup B}$.

4 Conclusion

In this paper we have investigated Information Diffusion problem in a Multi-Social-Network Scenario. We have applied ASP and analyzed the properties of real MSNSs using real-world data. In the future, we plan to remove the simplifications applied to the approach introduced in this paper, design other predictive models for Information Diffusion, and, finally, apply ASP for extending Social Network Analysis investigations from single social networks to MSNSs.

Acknowledgments. This work was partially supported by Aubay Italia S.p.A., by the project BA2Kno (Business Analytics to Know) funded by the Italian Ministry of Education, University and Research, and by Istituto Nazionale di Alta Matematica "F. Severi" - Gruppo Nazionale per il Calcolo Scientifico.

References

1. Baral, C.: Knowledge Representation, Reasoning and Declarative Problem Solving. Cambridge University Press (2003)
2. Berlingerio, M., Coscia, M., Giannotti, F., Monreale, A., Pedreschi, D.: The pursuit of hubbiness: Analysis of hubs in large multidimensional networks. J. Comput. Science 2(3), 223–237 (2011)
3. Buccafurri, F., Foti, V.D., Lax, G., Nocera, A., Ursino, D.: Bridge analysis in a social internetworking scenario. Inf. Sci. 224, 1–18 (2013)
4. Buccafurri, F., Lax, G., Nocera, A., Ursino, D.: Moving from social networks to social internetworking scenarios: The crawling perspective. Inf. Sci. 256, 126–137 (2014)
5. Domingos, P., Richardson, M.: Mining the network value of customers. In: Proceedings of the Seventh ACM SIGKDD KDD, San Francisco, CA, USA, August 26-29, pp. 57–66. ACM (2001)
6. Gelfond, M., Lifschitz, V.: Classical negation in logic programs and disjunctive databases. New Generation Comput. 9(3/4), 365–386 (1991)
7. Goldenberg, J., Libai, B., Muller, E.: Talk of the network: A complex systems look at the underlying process of word-of-mouth. Marketing Letters 12(3), 211–223 (2001)
8. Granovetter, M.: Threshold models of collective behavior. New Generation Comput. 83(6), 1127–1138 (1978)
9. Grasso, G., Leone, N., Manna, M., Ricca, F.: Asp at work: Spin-off and applications of the dlv system. In: Balduccini, M., Son, T.C. (eds.) Logic Programming, Knowledge Representation, and Nonmonotonic Reasoning. LNCS, vol. 6565, pp. 432–451. Springer, Heidelberg (2011)
10. Guille, A., Hacid, H., Favre, C., Zighed, D.A.: Information diffusion in online social networks: a survey. SIGMOD Record 42(2), 17–28 (2013)
11. Kempe, D., Kleinberg, J.M., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM SIGKDD, Washington, DC, USA, August 24-27, pp. 137–146. ACM (2003)
12. Leone, N., Pfeifer, G., Faber, W., Eiter, T., Gottlob, G., Perri, S., Scarcello, F.: The dlv system for knowledge representation and reasoning. ACM Trans. Comput. Log. 7(3), 499–562 (2006)