============================================================================

To: Munindar Singh, Editor-in-Chief, ACM Transactions on Internet Technology (TOIT)

Date: 17/8/2016

From: Liron Samama-Kachko, Barak Hagbi, Roni Stern, Rami Puzis, and Ariel Felner

Re: Revised ACM TOIT submission TOIT-2016-0051, "TONIC: Target Oriented Network Intelligence Collection Efficient Exploration of Social Networks"

Dear Editor,

Thank you for your correspondences from June 6th, 2016, regarding our submission "TONIC: Target Oriented Network Intelligence Collection Efficient Exploration of Social Networks".

We appreciate your comments and efforts and did our best to improve our paper and address the issues raised by the previous reviewers.

In particular, we made the following main changes to the paper to account for the reviewers’ concerns:

1. Following the recommendation of all reviewers, we added experimental results on two additional online social networks and analyzed the results. We believe this also addresses reviewer's 3 concern about having enough added material over the conference versions of this work.
2. The related work section was significantly expanded, including a discussion on graph sampling methods, focused web crawling, and a broader discussion on searching in an unknown graph.
3. We defined more clearly the term "lead" to address the concerns of reviewers 1 and 2.
4. A better example is given to demonstrate the differences between the proposed aggregation methods of the promising heuristic.
5. Typos and other minor errors were fixed.

In accordance with your recommendation, we hereby submit our revised version to ACM TOIT. Below please find our detailed response to all the previous reviewers’ comments.

Thank you.

Liron, Barak, Roni, Rami, and Ariel

=======================================================================

## Associate Editor:

**AE 1.**

*The paper received varied feedback from different reviewers. While all the reviewers appreciated the contributions of this paper, there were also multiple questions raised about the methodological details as well as the novelty of the work. An important concern raised was with regards to the novelty of this paper given two prior publications with a similar theme by the authors. With a Major Revision rating, the guest editors would like to provide the authors a chance to explain their contributions w.r.t. the prior work and also respond to the other concerns raised by the reviewers.*

**Response.**

Each of the two previous conferences papers had a special focus on some aspects of the TONIC problem and the approaches to solve it. A main contribution of this current submission is that here we provide a unifying view that summarized all our past work on the subject in a coherent and comprehensive manner.

Nevertheless, to further strengthen the novelty of this paper, we now added (1) two additional online social networks to the experimental analysis, thus allowing a more comprehensive evaluation and sounder conclusions, and (2) completely re-wrote the related work section to make it much more thorough, situating our work in proper context and explaining its relation with related topics like intelligent web crawling, graph sampling, and others. In addition, we also handled all the other concerns of the reviewers to the best of our ability.

## Referee 1:

**R 1.1.** *The primary concern that the reviewer has is the identification of leads. In this paper the assumption is that if a node is in LoF of target or of a lead, it can be considered as a lead. The paper does not provide a metric to measure using isLead() or Acquire() how can a user understand that a particular node is a lead.*

**Response.** As the reviewer wrote we defined that a lead is a profile that has the target in its LoF. Thus, a particular profile is either a lead or not. We dedicate a paragraph in section 3 (starting with the sentence "Information in general and the LOF in particular can be extracted from an OSN profile in several ways") in which we list several ways to extract the LOF of a given profile (web scrapping and OSN API such as FQL).

**R 1.2.** *The paper needs to re-format its references. Currently the paper is unreadable due to its reference formating since most of the references and tables/figures do not conform to the numbering scheme provided in the body of the paper. For example page 9 line 44 "power law distribution [14],". Since the references are by author names, the reader is expected to count the reference to arrive at it. Please change the reference format for this paper.*

**Response.** We are sorry for this unfortunate compilation error. It was corrected.

**R 1.3.** *Page 2 Line 50: Please provide a well-known reference of a best-first search. For example: (Vempaty 1991). You can choose a better example. The idea is to help a reader who has not come across a best-first search.*  
**Response.** We added the reference suggested by the reviewer and an additional reference.

**R 1.4.** *Related Works: The paper mentioned Profile Attribute Prediction as a related topic to the paper. Although the reviewer agrees to this premist but feels like web search can be an ideal candidate for relevant literature review. The similarity of TONIC with web search is the fact both of these techniques look into a special graph for targeted nodes.*

**Response.** Following the reviewer's suggestion, we have extended the discussion on web crawling and moved it to the beginning of the related work section. In addition to the fish- and shark- search methods we also mention focused search methods based on TF-IDF, PageRank and others.

**R 1.5.** *Page 6 Para 2 Lines 18-22: The reviewer feels that the Heuristic search in an Unknown Graph is more relevant than Link Prediction (See last comment). Please elaborate on this section and (if necessary, due to page constraints) reduce section 2.1.*

**Response.** Following the reviewer's suggestion, we extended the related work section on heuristic search in an unknown graph, and the discussions on attribute and link prediction were significantly shortened focusing only on insights and measures that are utilized in current work.

**R 1.6.** *Def: 3.3 Page 6 Line 35: Once a lead is acquired, how will the user verify whether the lead contains information pertaining to the target. The paper does not provide a metric to capture this. The paper assumes that members in LoF will contain relevant information about the user, which is not always true. I have 438 friends in facebook. Approx. 25~30 out them can be considered leads if I am considered a target.*

**Response.** Following the reviewer's concern, we refined our definitions: a lead now is explicitly defined in Definition 3 as a profile that has the target in its LOF. Thus, all of your 438 friends will be regarded as leads and the task in TONIC is to find them. Afterwards, we envision a secondary process in which an offline process analyzes the data from these leads to see which one of them indeed contains valuable information about the target. This secondary phase is out of the scope of our work, and may be done, for example, manually by an analyst or automatically with some sophisticated information extraction algorithms, to identify which of the found leads provide useful information.

**R 1.7.** *How does the paper understand whether someone is a L, PL, or a NL.*

**Response.** Any discovered profile that is not one of the initial leads is in PL until an isLead() query is applied to it. After applying an isLead() query to a profile, it is either added to L if it is found to be a lead, or added to NL otherwise. This is explicitly given in the pseudo code of Algorithm 1, and we also added in the revised version an explanation of a profile's life-cycle in the paragraph before we defined the TONIC problem.

Thus, the key to understand the group in which a profile resides depends mainly on whether an IsLead() was applied to it. As we clarified above and in the revised paper, a profile is a lead if it is a friend of the target, and therefore implementing isLead() is simply to check if the target is in a profile's LOF

**R 1.8.** *Please provide numbers to each equation for easy reference.*

**Response.** Done

**R 1.9.** *Page 6 Line 51: There should be a mention that isLead being in LoF is a necessary condition for isLead(v) but not sufficient condition.*

**Response.** We clarified above and in the revised manuscript that we define lead in this work exactly as a profile that has the target in its LOF. To determine whether a lead contains valuable information about the target is therefore left to either be done manually by a human analyst or by sophisticated Information Extraction tools that are outside the scope of this work.

**R 1.10.** *Equation Page 6 Lines 52: How do you account for target profiles that suppress their LoF.*

**Response.** We were unsure if by "target profiles" the reviewer means that target profiles or the profiles that are being queried during the TONIC process. If the reviewer is asking what to do if the target profile has suppressed its LoF (i.e, that accessing the target's profile is not allowed) then this is exactly the problem we address in this work: the target's LOF is not accessible and we aim to "reconstruct" it by accessing the LOF's of other profiles, which may have accessible LOFs. Accessing these LoF's may reveal that the target profile is in them, and consequently that the corresponding profile is a lead.

If the reviewer is asking what to do if a profile that is being queried (either isLead() or acquire()) has suppressed its LoF (i.e., made its LoF unaccessible), then we consider these profiles as non-leads. This is because we cannot verify that they are leads. Indeed, this is an important observation by the reviewer and we have added footnote in the revised manuscript to clarify this.

**R 1.11.** *The paper needs to elaborate on 2nd paragraph page 7 Lines 11:14: The paper mentions: "Intial leads can, for example, be found manually by using traditional intelligence techniques". Please elaborate. Do you mean u look at the entire LoF manually? Please provide references or elaborate.*

**Response.** We elaborate on this in the revised version, and provide examples of how this can be done. For example, in a police scenario where the goal is to obtain leads of a running fugitive, then the police may know the OSN profiles of the targets family and close friends, and these may serve as initial leads.

**R 1.12.** *Recommended: Page 7: Line 46: please italicize cost/reward. Can ignore.*

**Response.** Done.

**R 1.13.** *Page 8 Line 26: Algorithm 19? Should be Algorithm 1.*

**Response.** Fixed.

**R 1.14.** *Page 8 Algorithm 1: Line 6: In the text there is no elaboration on the ChooseBEST(OPEN) method. Please elaborate.*

**Response.** We now discuss the ChooseBest method in the text.

**R 1.15.** *Page 9: Para 2: Lines 19-23: Naming errors. b & c interchanged. Refer to figure 1.*

**Response.** Fixed.

**R 1.16.** *Page 9 line 21: initial lead instead of leads.*

**Response.** Fixed

**R 1.17.** *Page 9: Line 53: Peoples -> people.*

**Response.** Fixed

**R 1.18.** *Page 11 Line 32: Please change footnote to a symbol. superscript 2 might be mistaken for a square.*  
**Response.** We kept the footnote symbol to be a number to conform with TOIT standards. Note that we put it after punctuation marks to avoid such mistakes.

**R 1.19.** *Page 11 Section 5.1.4: Lines 45, 46: MaxP(P2) = 1/3 not MaxP(P1) = 1/3. And AvgP(P2) = 1/3 not AvgP(P1) = 1/3.*

**Response.** Fixed.

**R 1.20.** *Page 12: Lines 16-20: Please show a toy example where BysP is superior to MaxP or AvgP. In the example you provided all the 3 are pointing to the same nodes. P1 in this case which will be expanded first.*  
**Response.** Following the reviewer's comment, we replace Figure 6b with a different one that better shows the differences between the three aggregation methods.

**R 1.21.** *Page 12 Para starting with {The friend measure}: Lines 29-33: please mention the fact that it is a Jaccard measure.*  
  
**Response.** The friends measure is different from Jaccard's measure. The Jaccard's measure is the ratio between the intersection of the nodes' neighborhoods and the union of their neighborhoods, while the Friends Measure counts the **nodes** in the intersection of the neighborhoods and also counts the number of **edges** between nodes in the neighborhoods of the measured nodes. In other words, when computing Jaccard's measure between two vertices *v* and *u*, then you do not consider the edges between node in the neighborhood of u and nodes in the neighborhood of v. In contrast, Friends Measure counts these edges, in addition to the intersection between the neighborhoods. Nonetheless, in the revised version we better situate the Friends Measure w.r.t other measures, and in particular note that Friends Measure is a special case of the Katz measure.

**R 1.22.** *Table 1, Page 14: please mention in the figure title which model is used.*  
**Response.** The BTF corresponds to the column title zero, and the others are ETF with different bounds. We added this clarification to the figure caption.

**R 1.23.** *Page 16 para 1: Lines 17-25: please rephrase: sounds repetitive.*  
**Response.** We deeply apologize for this error and have removed the repeated part.

**R 1.24.** *Since this paper has a valid ethical concern, the reviewer requests a more detailed ethical concerns discussions.*

**Response.** We broadened the last section that discusses ethical concerns, explicitly stating the ethical concerns and discussing their gravity.

## Referee 2:

**R 2.1.** *1). what is the different between the task/approach in this paper and heterogeneous networks? It is clear that a lead could be a profie exposes information about the target, while heterogeneous networks could be built by extending the target by connected profiles or information. heterogeneous networks/graphs have been deeply discussed and explored in various applications and domains. I hope the authors can introduce related work and distinguish it from their own work here.*

**Response.** Heterogeneous social networks are studied mainly in context of link prediction, community detection, and influence / information diffusion. The only context relevant to current research is link prediction. Indeed, referring to multiple types of relationships could improve the crawling heuristics studied in this paper. Since such relationships are usually created by different social processes they carry more information than flat networks. This gives strong motivation to using machine learning models for guiding the search process in TONIC. We have added a paragraph along these lines to the related works section in context of link prediction.

**R 2.2.** *2). The authors introduced a hybrid approach in sectgion 6.2.2. As far as I know, there are a number hybrid heuristic approaches. Here, the authors adopted a switching method. Any other hybrid approaches can be used as alternative solutions? will the switching approach significantly increases computational costs? since it will switch as soon as BysP exhausted teh set of leads it can reach.*  
**Response.** Indeed, we have tried a number of hybrid approaches. We introduced the switching method as in our preliminary studies it worked better than other simple hybrid approaches. Further research may study more intelligent ways to combine any of the proposed heuristics, e.g., by learning from past crawls which combination of heuristics works best in every step of the search. Preliminary studies showed some potential to this approach, and we mention this in the revised manuscript's future work section.

Regarding computational costs, the proposed switching approach requires only a negligible amount of overhead of that of BysP – the overhead required to keep track of the number of IsLead() actions since the last lead was found.

**R 2.3.** *3). The experimental results are clear but only reply on the Google+ network. To make a strong claim on the results or patterns, for example, ETF outperforms RTF and even runs faster, it is better to get it evaluated based on multiple data, such as Facebook and Twitter social networks. I suggest the authors to examine at least on additional data, if it is possible to acquire the appropriate data for experiments.*  
**Response.** We took this comment of the reviewer very seriously and performed additional experimental evaluation on two additional OSNs. The results of these experiments are explained in the revised manuscript in Section 9.

## Referee: 3

**R 3.1.** *1) The novelty of this paper is limited. The technique presented in this paper is almost the same as the authors' pervious two papers [Ref 1,2]. There is lack of justification of the unique contribution of this submission. How does this submission significantly extend the pervious work?*  
**Response.** Indeed,some part of the paper were already presented in two separate conference papers. The main importance of this current submission is that here, we provide a unifying view that summarized all our past work on the subject in a coherent flowing manner. In addition, (as the reviewer noted) we added analysis from the Information Retrieval perspective. Moreover, following the reviewers comment, we also add in this revision new experimental results on two new OSNs and analyze them. These results show that out approach is robust across different OSNs.

**R 3.2.** *2) The experiment results are similar to the results presented in [1,2]. The IR task in section 8 is the only new addition. However, the section doesn't provide much insight beyond the previous sections. Further discussion may include how different methods work or fail in dealing with false positive or false negative cases.*

**Response.** As we wrote above, to address the reviewer's concern we added substantial new experimental results on two new OSNs in the revised manuscript. In addition, we expanded on the analysis of our IR results, and in particular on what lessons can be learned from comparing the AUC and DCG metrics. Regarding a deeper study of how different methods work or fail in dealing with false positive and false negatives, we note that this is in some sense related to the net-gain resultsin the cost-benefit analysis section (section 7). On this line, one may consider having different costs for querying different profiles, e.g., if some profiles are more sensitive to queries and thus adds a risk element to querying them. This is indeed an exciting line of research to pursue, and we added this as a possible future work.

**R 3.3.** *3) The proposed approach is based on the following statement: "Based on the assumption that leads tend to cluster together, …" -- the authors had not justified this assumption properly.*

**Response.** The support for this assumption comes from the homophiliy principle, which states that people tend to be friends with people with similar attributes. This principle is known to manifest in online social networks, and it translates in our context to supporting the assumption that leads tend to cluster together. This is better explained in the revised manuscript.

**R 3.4.** *4) The related work is not sufficient and appropriate. The authors should carefully review related work in graph sampling and survey sampling methodology. For example, related work like [Ref 3,4] have discussed more conditions (such as graph structure) and sampling criteria (such as representativeness) that are relevant to this problem.*

**Response.**  We thank the reviewer for pointing out this issue and have re-worked the related work section to discuss web crawling algorithms designed for graph sampling. The key difference between TONIC and web crawling for graph sampling is that in TONIC we to find profiles that are leads, while in graph sampling the purpose is to ***estimate*** a property of the OSN (e.g., how many profiles attended college). Also, the number of profiles usually crawled in graph sampling research is much more than the number of crawls in a TONIC search.

**R 3.5.** *5) The writing of the paper seems pretty rough. The citations and references are in different formats and hence are difficult to correspond.*

**Response.** We are sorry for this unfortunate compilation error. It was corrected.

**R 3.6.** *6) The relevance of this paper to the special issue has not been well justified.*

**Response.** Our paper is aimed to the general scope of ACM TOIT, which is "all areas of network and web systems". We believe that intelligent crawling for information in web-based social networks fall into this category.

**R 3.7.** *\* For different ETF heuristics, the paper provides empirical results with different iteration number (Figure 6). It would be nice to also provide the same analysis on RTF heuristics.*

**Response.** Both perspectives (percentage and leads) are interesting but we did not want to over-crowd the paper with too many experimental results variants. However, the new results for the added OSNs are given with the same analysis – where the number of iterations is on the x-axis.

**R 3.8.** *\* There are typos in this paper. For example, in the example in Page 9  b is non lead and c is lead according to figure 1, while lines 21-23 say b is lead and c is non lead.*

**Response.** We apologize for such errors and did our best to fix them in the revised manuscript.