Sockpuppet paper — need a good title

NAMES

October 3, 2020

1 Introduction

2 Methods

2.1 The DSOCK Experiment

From vs to all: describe the experiment here.

2.2 Notation

We use $\mathcal{U}$ to denote the set of all user accounts and $Norm, Obs, Sock$ respectively to denote the pairwise disjoints subsets of normal, observer and sockpuppet accounts in $\mathcal{U}$. For each sock account sa, we use $sa.op$ to denote the user in $Norm$ who operates sa. We use $u.Socks$ to denote the set of socks operated by a user $u \in Norm$. If u does not operate any socks, then we set $u.Socks = \{u\}$. Each user $u \in \mathcal{U}$ has an associated set of posts, comments and likes denoted respectively by $u.posts$, $u.comments$ and $u.likes$. Each sock $u \in Sock$ also has an associated type $u.type \in \{\text{Covert}, \text{Overt}, \text{Unrestricted}\}$.

We likewise use $Posts$ and $Comments$ to denote the set of all posts and the set of all comments (on posts). Each post or comment has several associated attributes (or fields) such as $author$, $PostTime$, $Text$ denoting the author, time of posting and the textual content of the post/comment. In addition, comments have a Id field in order to identify the post about which the comment was made. Thus, $c.Id$ refers to the post about which a given comment c was made, while $c.Text$ shows the text of the comment c.

From vs to all: need to expand the notation and clean it up

2.3 Measuring Influence

An *event* e in the DSOCK experiment is any one of the following:

1. Post event: a user makes a post.
2. Comment event: a user makes a comment about a post. From vs to all: did we allow comments on comments?
3. Like event: a user "likes" a post or comment.
4. View event: a user views a post or a comment.
5. Report event: a user reports an account as being a sockpuppet.

The first three types of events are *expression* events where a user expresses his/her sentiment explicitly. View events are impressions which tell us what all posts/comments the user actually looked at — and which could therefore have elicited his subsequent expressions.

Our influence measurement method tries to assess the influence of users u on an explicit *expression* event e' by user u'. **We will only consider the cases when:**

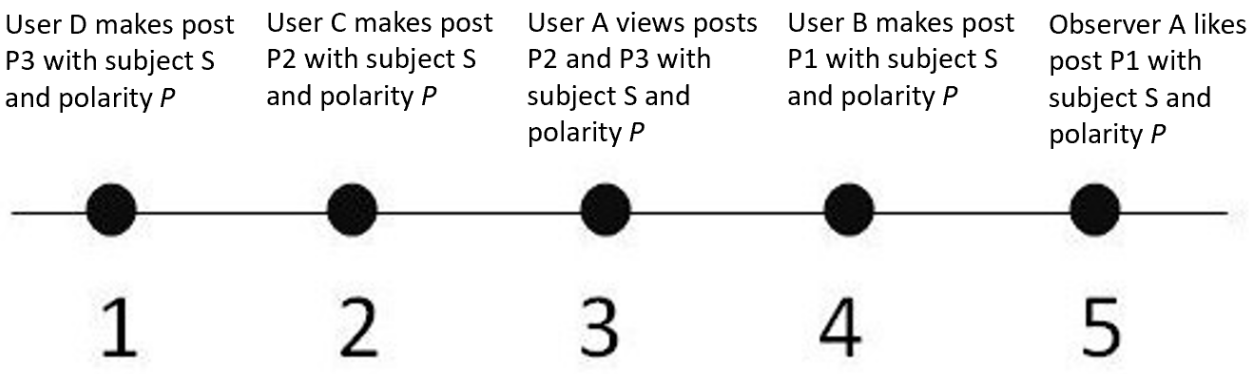

Figure 1: *Sample Timeline of Events Leading up to user A liking post P1 with subject S and polarity p.*

1. u' is an observer and
2. the event e' involves expression of an opinion by u'either on a topic on which e' had no prior opinion (as ascertained by the surveys conducted prior to the DSOCK experiment) or where the prior opinion held by u' had a different polarity than the one expressed in event e'.

Consider an expression event e' in which the user u' expresses a given sentiment polarity $pol \in \{+, -\}$ about a subject sub (e.g. GMO foods). Our first task is to identify the set $PossInflEvents(e', pol)$ of events that could potentially have led to the occurrence of the expression event e'. $PossInflEvents(e', sub, pol)$ is defined as the union of 4 sets:

1. $ViewedLikeEvent(e', pol)s$ is the set of all "like" events before $e'.Time$ involving posts/comments that have the same polarity pol on sub which $e'.author$ has previously viewed;
2. $ViewedPostEvents(e', pol)$ is the set of all "post" events before $e'.Time$ involving posts/comments that have the same polarity pol on sub which $e'.author$ has previously viewed;
3. $ViewedCommentEvents(e', pol)$ is the set of all "comments" events before $e'.Time$ involving posts/comments that have the same polarity pol on sub which $e'.author$ has previously viewed.

$PossInflEvents(e', sub)$ is the union of the above three sets. Figure 1 shows a sample scenario where, at time 5, an observer A "likes" a post P1 which was made at time 4 by user B. However, there are many possible answers to the question of who should get credit for the "like" expressed by A. There are many possible answers:

1. *Earliest Author.* According to this strategy, the first author who made a post that A viewed and that expressed the same sentiment on the subject as A should get credit. In this case, user D should be credited for A's like because D was the first to express that sentiment (at time 1) and because we know that A viewed D's post (P3) at time 3.
2. *Liked Author.* According to this strategy, the author whose post is actually liked by A should get credit. In this case, user B should be credited for A's like because B's post was the one that A liked. However, a proponent of the Earliest Author Credit might argue that B was just repeating what A had already said.

However, a complaint can be made about *both* these strategies by user C who was neither the first to express the given sentiment, nor was he the one whose post was liked by A, but his post expressing the same sentiment was viewed by A before B ever made his post.

2.3.1 Direct Influence Model.

Our first influence model ignores follower-followee relationships (more on that later).

Attention Window. The direct influence model assumes the existence of an *attention window* of length w. The idea is that the only events that can possibly influence an event e' are those that occurred during the $[e'.Time - w, e'.Time - 1]$ window. For instance, suppose $w = 3$ in the example shown in Figure 1. In this case, the attention window for influencing the event at time 5 is $[2, 4]$. According to this time window, the events at time 1 are assumed to not influence the event e' *unless* something happened during the attention window that caused that event to be recalled. For instance, in thie $[2, 4]$ window, user D's post P3 is viewed by user A, so even though the event at time 1 is forgotten by A during this window, post P3 and its author still potentially influence A through the view action (on P3) at time 3.

Influence Parameters. The direct influence model assumes the existence of three parameters α, β, γ such that $\alpha+\beta+\gamma = 1$ — these parameters respectively apportion a credit to each event and thus to the author of that event. α (resp. γ) assign a credit to the first author to express the sentiment of interest on the subject (resp. the author whose post is explicitly liked or re-posted) within the attention window whose sentiment on subject *sub* is replicated in event e'. β is equally distributed to those in the "middle".

In our example, suppose $\alpha = 0.4, \beta = 0.2, \gamma = 0.4$. In this case, e_1 and e_4 each get 40% credit for e'. The remaining 10% is split equally between e_2, e_3. These allocations allow us to infer that under these parameters, B ends up with 40% of the credit, C gets 50%, while D gets 10%.

On the other hand, if w had been 4, our attention window would have been $[1, 5]$ and the $\alpha = 0.4, \beta = 0.2, \gamma = 0.4$ parameters settings would have allocated 40% of the credit to each of e_1 and e_5. The remaining 20% would have been split 10% each to e_2, e_3. The final allocation of credit would have been 40% to B, 10% to C, and 50% to D. These numbers are called the *base influence scores* of these events (and by extension of the users).

Direct Influence Graph. We define the novel concept of a direct influence graph associated with an event e'. The vertices in this graph are events in the set $PossInflEvents(e', sub)$. There is an edge from event $e \in PossInflEvents(e', sub)$ to e' with a weight corresponding to the base influence score of u. Figure **??** shows the direct influence graph associated with the timeline in Figure 1 with $w = 4$ and $\alpha = 0.4, \beta = 0.2, \gamma = 0.4$.

Note that direct influence graphs are associated with a single event e'.

2.4 Indirect Influence Model.

A problem with direct influence graphs is that they only consider those who directly influenced event e'. However, in our running example, it might be the case that there was another user E who had made a post with subject S and polarity P which user C viewed before making its post at time 2. In this case, an argument could be made that E influenced the event e' through C even though E does not appear anywhere in the direct influence graph of event e'.

Full Influence Graph. Given an event e', the *full influence graph* associated with e' is the graph that is obtained by recursively finding the full influence graph of each node in the direct influence graph of e'. To achieve this formally, we define an operator $\oplus(G, v)$ that takes an influence graph G and a vertex v in that influence graph as inputs and generates a new influence graph as input. Suppose $G = (V, E, \wp)$ where $\wp$ is a weight function that assigns a weight to each edge, and suppose the direct influence graph G_v of vertex v is $(V', E', \wp')$. Then $\oplus(G, v) = (V \cup V', E \cup E', \wp \odot \wp')$ where

$$(\wp \odot \wp')(x, y) = \begin{cases} \wp(x, y) & \text{if } y = e' \\ \wp'(x, y) & \text{if } y = v \end{cases}$$

We can now inductively define the full influence graph associated with e' as follows. Our definition uses $\mathsf{DIG}(e')$ to denote the direct influence graph associated with event e'.

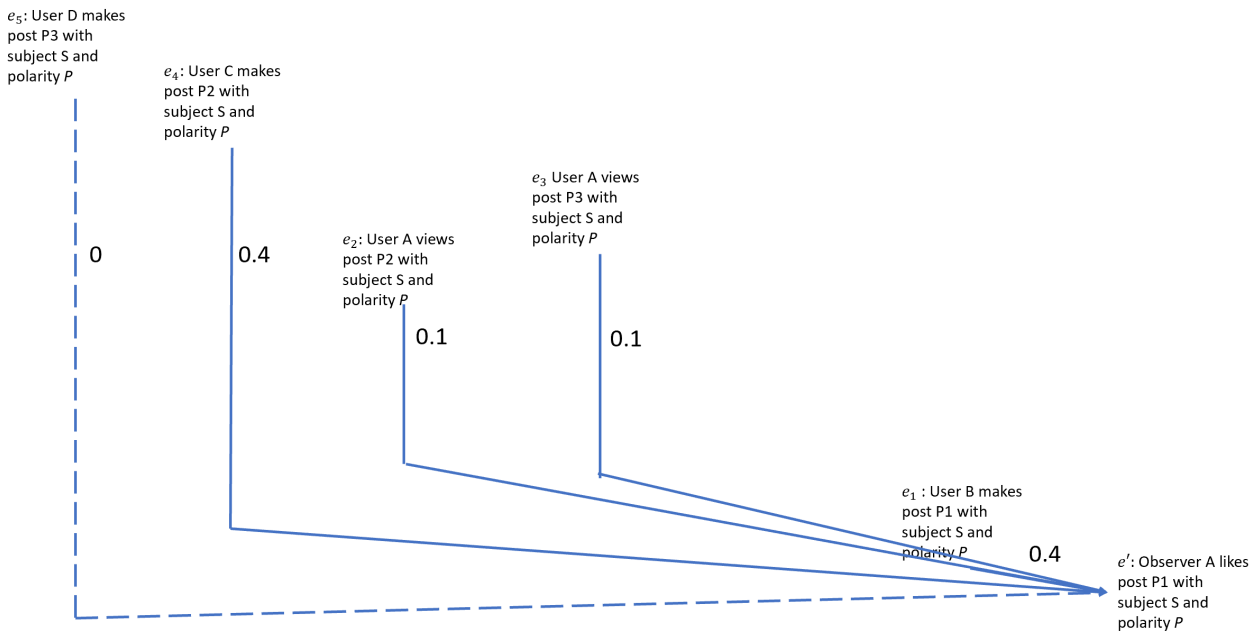

Figure 2: *Sample Direct Influence Graph for the event $e' =$ "User A likes Post P1 with subject S and polarity P using $w = 4$ and $\alpha = 0.4, \beta = 0.2, \gamma = 0.4$.*

$$\mathsf{FIG}_0(e') = \mathsf{DIG}(e').$$
$$\mathsf{FIG}_{j+1}(e') = \bigoplus_{x \in \mathsf{FIG}_j(e')} (\mathsf{FIG}_j, x).$$

We start with the direct influence graph of e' (FIG_0) which is the event of interest and expand it to include the direct influence graphs of all vertices in $\mathsf{DIG}(e')$ to get FIG_1. The same process is repeated till we eventually find a k such that $\mathsf{FIG}_k(e') = \mathsf{FIG}_{k+1}(e')$. Because the set of impressions and expressions in our data is always finite, such a k must exist. We will denote this by $F_\infty(e')$ for the sake of simplicity.

Full Influence of a User on e'. The *full influence of a user u on the event e' is the sum of the pageranks of the vertices in $F_\infty(e')$ in which user u was involved.*

From vs to all: The last definition is a bit vague as I was running out of time.

References