Paper 319-2011

How You Can Identify Influencers in SAS® Social Media Analysis (And Why It Matters)

Don Hatcher, Gurpreet Singh Bawa, Barry de Ville, SAS Institute Inc., Cary, NC

ABSTRACT

There are many ways to calculate influence in social media. We use an approach that distinguishes between content generators and conversation generators. This allows us to distinguish between the generation of potentially passive content versus the generation of conversations that include content with “pass along” appeal. The “pass along” content is potentially more influential since it exposes the content to more actors and embodies a social “referral” component which has been demonstrated to affect the information and behavioral impact of messages.

Content Generator Metric. In its simplest form, this is the number of communications over a given time period. A social media user who posts comments to a huge follower community generates a large content influence score. Unless the information is re-tweeted or mentioned in someone else’s post, the “conversation” part of the influence score will be relatively low.

Conversation Generator Metric. Social media users who drive conversations tend to generate a large number of file forwarding activities. They are mentioned a lot. They might also motivate a large number of replies and comments.

Our original construction of metrics to operationalize these metrics used direct measures of tagged content in the social media source (in this case Twitter). Later, we experimented with the construction of metrics that used a standard network analytic framework. One nearly-universal framework uses 1st and 2nd degree inbound and outbound calculations to compute the metric.

Our experimentation allows us to demonstrate that the original “custom” approach and the standard network-analytic approach are functionally equivalent. This confirms that our network-analytic approach to measuring influence taps into the notions of capturing both content and conversation. Further, that adoption of a standard approach enables us to extend our notions of measuring various aspects of influence – first developed in the Twitter Social Media data source – to other sources in Social Media. It also allows us to extend advances in social network analytics that are based on the inbound and outbound model to influence metrics that we build in our solution.

In summary, our work strengthens our confidence and understanding of both custom and standard approaches, and provides additional insight to users of these metrics in our Social Media Analytics solution.

INTRODUCTION

Influence is an important driver to understand and assess observed behavioral patterns in social media. Some studies, for example, “Do Friends Influence Purchases in Social Networks?” (Sanman and Gupta, 2009) indicate that the relationship between influence and behavior is curvilinear – influential messages can create a “saturation effect” after which behavior that was previously reinforced is now discouraged.

Other research indicates that following alone cannot be used to determine influence, “Measuring User Influence in Twitter: The Million Follower Fallacy” (Cha, et. al., 2010). This research indicates that influence is both a function of outbound content – as in the case of twitter users tweeting to a large group of followers – as well as conversation generated – as is the case when a tweet leads to replies, retweets or author mentions.

Because of considerations such as “The Million Follower Fallacy” we initially set out to use an approach to the calculation of influence in social media that distinguishes between content-generators and conversation-generators. Although this metaphor applies to all content in social media, and social networks in general, we demonstrate the evolution of our implementation of this approach within the Twitter data source. Later we show how this approach can be adapted to apply across different social media site types and sizes to ensure reliability and comparability. We show how the metric is calculated and used in social media and what the bottom line is for people who rely on social media in their social and professional life.

ORIGINAL CONTENT-CONVERSATION FORMULATION

Our original Twitter Influence Score is based on two concepts that drive influence – content and conversation. Content is ~~only~~ valuable only if people are listening. As such, we only give content a 25% weight (sending tweets) and conversation a 75% weight (made up of mentions, retweets and replies.)

Influence is calculated separately for Topic Relevant content and Non-Topic Relevant content.

Topic Relevance. A relevant topic determined by the user; for example, "accommodation" topics, sentiment and so on (here the user is interested in hotel commentary).

Content Generator Metric. In its simplest form, this is the number of tweets over a given time period. This is reflected in the "Twitters" column of our influence report.

Conversation Generator Metrics. Twitter users who drive conversations tend to generate a large number of re-tweets. They are mentioned a lot. They might also motivate a large number of replies. The "conversation" part of the influence score therefore reflects the metrics of Replies, Retweets and Mentions.

The first step in deriving the four categories that make up the influence score is to calculate the daily raw total for each category. The raw total is calculated for each category as follows:

- Tweets – we take the number of tweets sent and multiple it by the number of followers for their raw tweet score.
- Mentions

  When a person is mentioned in a tweet, they get a mention score which is the sum of all the followers that received tweets in which the person was mentioned.

  The person sending the tweet just gets a basic Tweets score.
- Retweets

  When a person has their tweet re-tweeted, they get a retweet score which is the sum of all the followers that received the retweet.

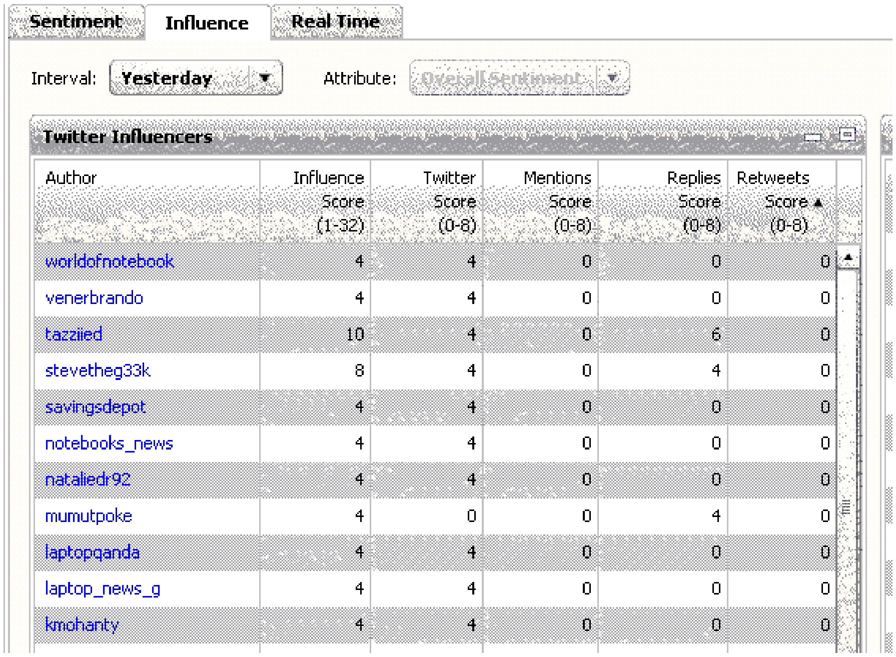
  The person sending the retweet just gets a basic Tweets score.
- Replies

  When an individual gets replies, it is unclear how many followers get a reply, so we just count the number of replies.

  The person sending the reply just gets a basic Tweets score.

This data is stored daily for each user that tweets, is mentioned, is re-tweeted and is replied that is Topic Relevant. For a standard interval (that is, 1, 7, 30, 60 or 90 days in the system today) or customer interval (user defined), the raw totals are summed for each tweeter. These numbers are plotted on a standard bell shaped curve and a score between 0-100 is assigned for each category. This score represents the area under the normal distribution curve that the standard score implies. Zero means they had none in that category.

Finally, we sum the four calculated scores and divide by four. Figure 1 provides an example of influence display in SAS Social Media Analytics.



Interval: Yesterday

Attribute: Overall Sentiment

Twitter Influencers

Author	Influence Score (1-32)	Twitter Score (0-8)	Mentions Score (0-8)	Replies Score (0-8)	Retweets Score (0-8)
worldofnotebook	4	4	0	0	0
venerbrando	4	4	0	0	0
tazziied	10	4	0	6	0
stevetheg33k	8	4	0	4	0
savingsdepot	4	4	0	0	0
notebooks_news	4	4	0	0	0
nataliedr92	4	4	0	0	0
mumutpoke	4	0	0	4	0
laptopqanda	4	4	0	0	0
laptop_news_g	4	4	0	0	0
kmohanty	4	4	0	0	0

Figure 1: Example SMA Screen Display of Influencers

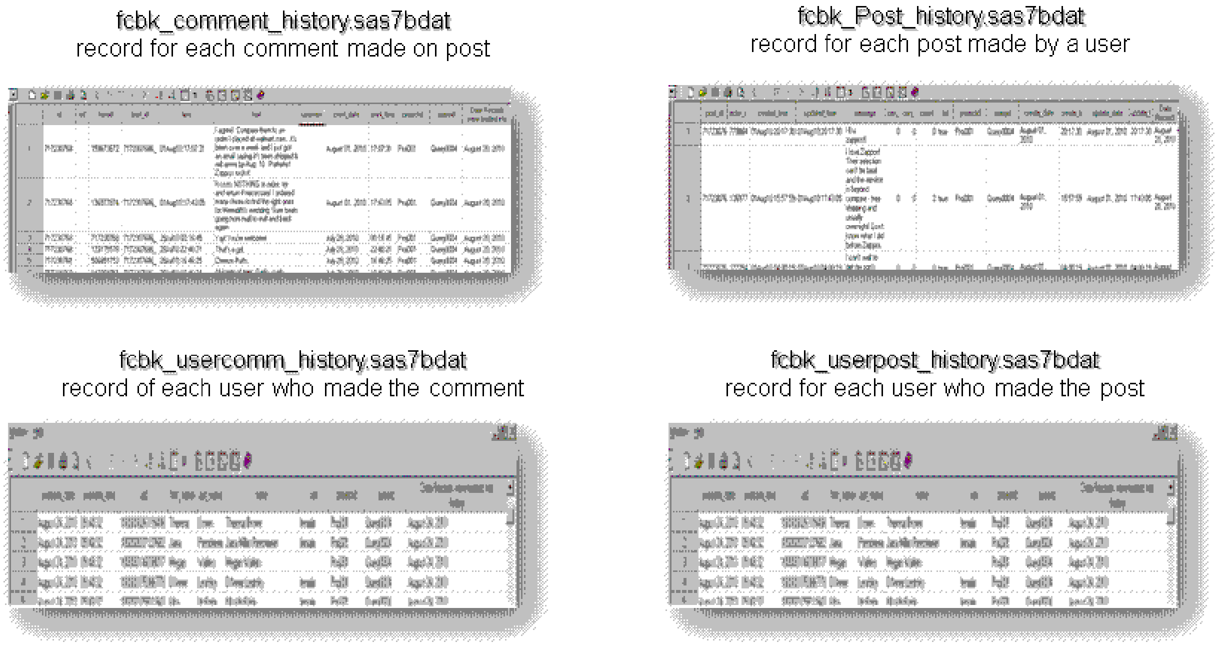
GENERALIZING INFLUENCE CALCULATIONS

The Twitter social media data feed makes the type of content versus conversation distinction easy to calculate. Tweets as content versus Replies, Retweets and Mentions as “conversation” is an easy concept to grasp and, as shown above, easy to implement. Other social media such as Facebook do not index content in this fashion. Moreover, social media network literature and social network analytics are based on directed graphs which capture in and out messages between nodes. In this representation, Tweets are out messages and Replies, re-tweets and mentions can be considered inbound messages.

We realized that if we were to recast the twitter influence calculations in this fashion – to construct outbound and inbound message counts as edges between the nodes of a graph -- then we would have a general social network representation that could apply across other social media sources as well. We understood the flexibility that this would give us and sought to ensure ourselves that we did not lose the capture of content versus conversation when the new metrics were calculated.

Facebook provides a typical example of the inbound versus outbound (directly graph) representation. Facebook fan pages, for example, consist of posts and comments to posts. The Facebook database carries post_id on the message table so every post and comment is captured by this post_id. So outbound and inbound messages are easily calculated.

Figure 2 displays the Facebook data structure.



fcbk_comment_history.sas7bdat
record for each comment made on post

fcbk_Post_history.sas7bdat
record for each post made by a user

fcbk_usercomm_history.sas7bdat
record of each user who made the comment

fcbk_userpost_history.sas7bdat
record for each user who made the post

Figure 2: Example Data Structure for Post-Comment Structure

As shown in Figure 2, each post and comment (shown as Post History and Comment History) are indexed by postid. Both the Post and Comment tables contain a link to the user ID which contains information about the user's profile (name, gender and so on). This user information is held in the Usercomm and Userpost history tables.

SECOND DEGREE METRICS

When we move to an outbound and inbound influence metric calculation, we begin to reap the advantages of the body of research conducted in social network analysis over the past decade. When we use the calculation of second degree metric scores, "degree" is a measure of separation between nodes in a network. The 1st degree metrics are the direct number of connections (in and out) between any two nodes in a network. The 2nd degree metric is the first set of indirect connections; in other words, the outbound and inbound metrics of the nodes that have a direct connection. For example, on a given day I correspond with two friends. My first degree metric will be a count of all the outbound and inbound messages that I exchange with those two friends. My two friends, in turn, also exchange messages during the day with their network of friends. So these friends also have 1st degree metrics that measure their communications.

If I include both my 1st degree metrics for my daily correspondence with my friends and, in addition, include ***their*** 1st degree metrics then the indirect metrics from my friends that I include are called my 2nd degree metrics. The inclusion of 2nd degree metrics is a way of measuring the proximate influence of my friendship network and is a way of gathering metrics that point to the "pass along" potential of messages that I exchange in my network. This formula recognizes that connected nodes in a network will pass along information that is directly exchanged between nodes. Information that is received from a direct connection is presumed to be more likely to be passed along than information that is received from nodes that are further afield (all things equal). So if I have many nodes in my network and if these nodes, in turn, also have many nodes then my messages are more likely to be propagated through the network than messages from other members with fewer nodes (all things equal). So my 1st and 2nd degree metric calculation is a powerful measure of the influence that I am likely to exert in a network.

This metric uses _*actual*_ observable social network connections in the influence calculations. The rationale is that regardless of the number of theoretical "views" of a post, it is the actual ***engagement*** that counts most: therefore only users who actively participate in a conversation are considered in the influence calculations.

In summary, the four components of influence that are included in the calculation are as follows:

- Direct messages sent to members of a user's network
- Direct responses received from members of a user's network

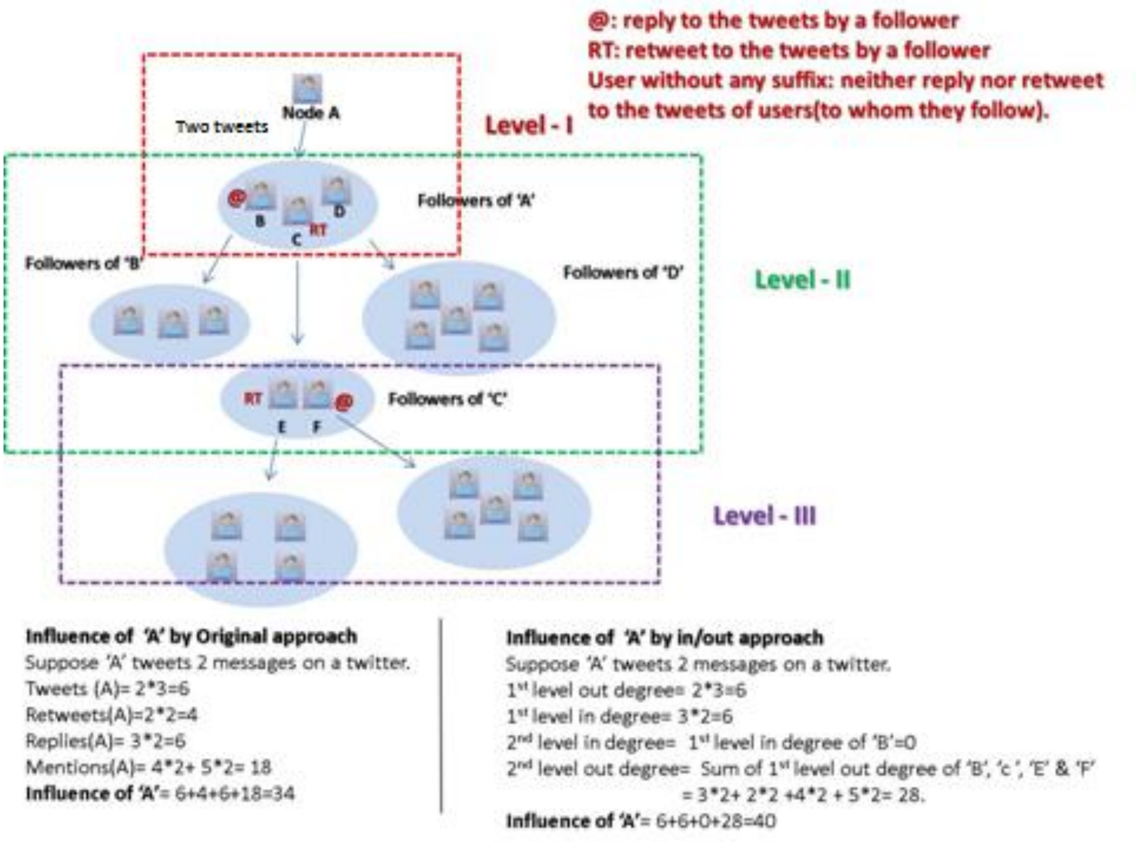
- Indirect (2nd degree or once-removed) outbound messages sent; that is, posts or comments made by members of the user's network
- Indirect (2nd degree or once-removed) inbound messages received; that is, comments received by members of the user's network.

The inclusion of the 2nd degree, once-removed metric, introduces an element of likely impact of any given message given the size, density and interconnectedness of the social network of a user. This is based on observable and ~~--~~ empirical~~--~~ measures of network activity.

IMPLEMENTATION AND RESULTS

As shown above, the 2nd degree metric calculation is in some respects an evolution of our earliest notions of breaking influence ~~up~~ into "content" and "conversation. In our earlier example, "conversation" was measured by replies, re-tweets and mentions. In the new framework, "conversation" continues to be measured by summing up replies, re-tweets and mentions. In addition, we overlay the notion of collecting conversation potential by including 2nd degree calculations in the construction of the metric.

The difference between the two methodologies can be illustrated by means of an example, shown here in Figure 3. When the original approach is compared to the inbound versus outbound network analytic approach, we can first of all see that the computed metric is similar in both approaches (34 in the original approach versus 40 in the network analytics approach). In this scenario, the author that we are considering is shown in the top of the diagram as "Node A". This author sends two tweets. Node A has 3 followers. So the "Tweet" score in the original approach is shown as 2 * 3 → 6. This score is replicated in the alternative network analytic calculations by the 1st level out degree. Here the number of outbound messages are shown as 2 tweets, each of which goes to 3 followers. This results in the same score as in the original approach; that is, 2 * 3 → 6.


@: reply to the tweets by a follower
RT: retweet to the tweets by a follower
User without any suffix: neither reply nor retweet to the tweets of users(to whom they follow).
Node A
Two tweets
Level - I
@
B
C
RT
D
Followers of 'A'
Followers of 'B'
Followers of 'D'
Level - II
RT
E
F
@
Followers of 'C'
Level - III

Influence of 'A' by Original approach
Suppose 'A' tweets 2 messages on a twitter.
Tweets (A)= 2*3=6
Retweets(A)=2*2=4
Replies(A)= 3*2=6
Mentions(A)= 4*2+ 5*2= 18
Influence of 'A'= 6+4+6+18=34

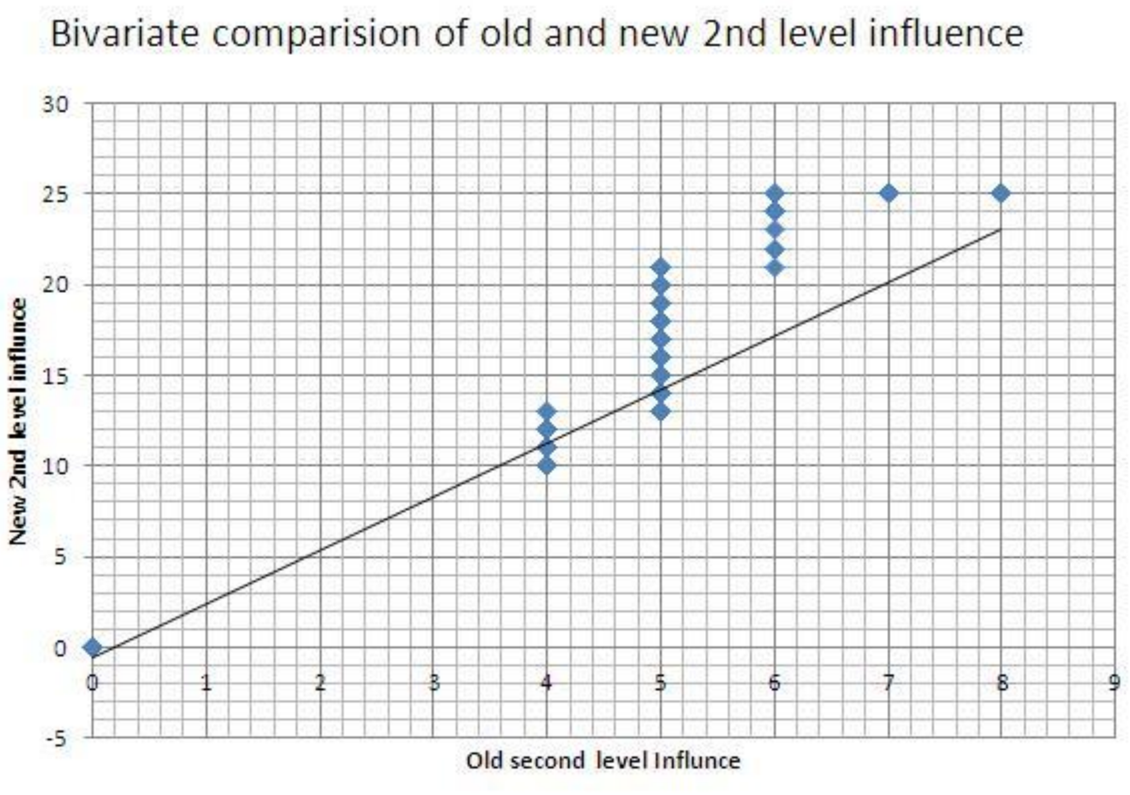
Influence of 'A' by in/out approach
Suppose 'A' tweets 2 messages on a twitter.
1st level out degree= 2*3=6
1st level in degree= 3*2=6
2nd level in degree= 1st level in degree of 'B'=0
2nd level out degree= Sum of 1st level out degree of 'B', 'c ', 'E' & 'F'
= 3*2+ 2*2 +4*2 + 5*2= 28.
Influence of 'A'= 6+6+0+28=40

Figure 3: Example Influence Calculations for the Original Approach Compared to the Network Analytic

Approach

Subsequent calculations are shown for the two approaches in Figure 3. In both cases, 4 separate metrics are computed. These 4 metrics are combined and normalized so that the sum is presented as a number that varies from 0 – 100. As we can see, the two sets of calculations are very similar: 34 versus 40.

To see the comparison of the two metrics beyond one case, Figure 4 shows a graphic comparison of the respective calculations across a representative set of Twitter observations. In this example, there is a strong correlation between the two sets of metrics. The favorable conclusion indicates that the two sets of metrics track one another, they are functionally interchangeable



CONCLUSION … NEXT STEPS

Our original intention in the construction of the Twitter influence metric was to capture aspects of both “content” and “conversation”. This directly addressed the well-known phenomenon that number of followers alone does not provide a full picture of influence. As we expanded the calculation of influence to other data sources, we saw the advantage of moving to the well-established network analytic framework of using in degree and out degree metrics based on 1st and 2nd levels of association. This framework is particularly appropriate to calculate influence in Facebook fan pages. This framework also positions us to take advantage of a wide range of other network analytic calculations – such as cohesiveness, centrality, and will also support the calculation of “Reach”; that is, the number of links in a network that a given message ***could*** travel given the connectedness of the network. The metrics of “Reach”, “Centrality” and “Cohesiveness” are all metrics that are available given the inbound and outbound message counts that we use in Facebook. These measures correspond with current standards in social network analytics.

We are pleased that our original conceptualization of influence calculation in Twitter – which was informed by a compelling interest to capture conversation compared to simple content – is so faithfully reproduced by the standard network analytic computations. This provides confirmation that most – if not all -- of the original intention in the construction of the influence metric is preserved in the new influence formulation. It also provides confidence in the implementation of network-analytic approaches across other sources of social media data.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Don Hatcher
SAS Campus Drive
SAS Institute Inc.
E-mail: Don.Hatcher@sas.com

Gurpreet Singh Bawa
SAS Campus Drive
SAS Institute Inc.
E-mail: Gurpreet.Bawa@sas.com

Barry de Ville
SAS Campus Drive
SAS Institute Inc.
E-mail: Barry.deVille@sas.com

REFERENCES

M. Cha, H. Haddadi, F. Benevenuto, K.P. Gummadi, **MEASURING USER INFLUENCE IN TWITTER: THE MILLION FOLLOWER FALLACY,** paper presented at the 4th Int'l AAAI Conference on Weblogs and Social Media,

May 23-26, 2010, George Washington University, Washington, DC.

Raghuram Iyengar Sangman, H.S. Gupta, **DO FRIENDS INFLUENCE PURCHASES IN A SOCIAL NETWORK?**, Harvard Business School, Working Paper 09-123, February 26, 2009.