Predicting SaaS virality: A diffusion model

Daphne Simmonds
Metropolitan University of Denver

Ken McDonald
Metropolitan University of Denver

Katia Campbell
Metropolitan University of Denver

ABSTRACT

This research was conducted with two goals: 1) to understand how to measure virality – in particular of SaaS products – to understand what factors impact virality and also how these factors may be addressed in adjusting an existing base model and record the lessons learned from the experience in attempting to do so; and 2) to develop and validate a comprehensive model that can predict accurately virality. The data used was collected by a startup that had developed and implemented a SaaS product. The findings showed that no single model can be expected to accurately predict virality of even a single product as the virality of changes over time, and that model adjustments have to be made for different products and possibly, for different periods. This study has several implications for research and practice. These came by way of the several challenges met as sought to first fit the data to an existing model and then adjust the model to account for lack of fit. These served as the lessons that can guide practitioners seeking to develop models to predict virality of products being contemplated or being marketed. For practitioners, these are presented these as challenges and solutions which are described in detail.

Keywords**:** Software Diffusion, Electronic Word of Mouth (e-WoM), Virality model development, quantitative case research.

INTRODUCTION

The Internet and social media have resulted in an increasing shift of advertising spending from traditional modes – TV, radio, print -- to what is now known as viral marketing (Jurvetson and Draper 1997). Viral marketing involves attracting recommendations through electronic word-of-mouth (e-WoM) for products or services from social networks – not necessarily social media networks. The idea is that marketers would distribute the product, or product details to an initial group of persons – the seed (Ho and Dempsey 2010). The seed group, after trying/viewing the content, would then, if sufficiently impressed or incentivized, recommend and/or forward the content to those in their network(s). The [recommendations], hopefully seen as validation of the goodness of said products or services, are expected to spur successive groups of recipients to forward to others within their own networks. As the content is forwarded from one set of receivers to others, there is the potential to quickly reach a large group of persons at an exponential rate (Palka and Pousttchi 2008). In such a case, diffusion is considered to be viral (Hoang and Lim 2012). This phenomenon has been likened to the infection cycle witnessed in major epidemics, and is now "affectionately" known as viral growth (Kaplan and Haenlein 2011, Cao et al. 2009).

In industry, the extent of diffusion the product experiences – its virality -- is measured by what is known as the viral coefficient. Let us consider, for example, that a company provides a product to the market. Consider the case where each product user automatically and involuntarily (by virtue of using the product) invites 15 persons to use the product. Say there is a 3% probability that each of those 15 people will use the product and further invite 15 people to use the product, and the cycle continues as such. The product virality, denoted by K, its viral coefficient, equals 0.45 (3%*15). True viral diffusion is achieved when K>1 (Jarvis 2017a, Skok 2009). Only in that case can the company claim that there truly is viral growth and, possibly more important, only then can marketers determine that they can cut spending on marketing.

Viral marketing has been said to compound the benefits of a first-mover advantage and is considered critical for swift and large-scale adoption/growth (Jurvetson and Draper 1997). The need for products and services to experience viral adoption is of great interest to business owners and marketers. Therefore, both the campaigns that are launched and the subsequent adoption of the content are of interest to business managers and owners. Research efforts in academia have not been commensurate to the seeming importance of the topic in industry; nevertheless, some research has been done. For example, the literature includes a number of studies that explore the key success factors for viral marketing campaigns. Many of these studies have focused on either the impact of factors related to individuals to whom viral content has been forwarded and/or their audiences (Liu 2012, Jain et al. 2014, Wiedemann and Haunstetter… 2008). Others have focused on factors related to the network through which the content is expected to be diffused (Hildebrand, Hofstetter, and Herrmann 2012, Pousttchi and Wiedemann 2007), and a third set on factors related to the viral content itself (Deza and Parikh 2015, Guerini, Strapparava, and Ozbal 2011). In a few cases, researchers have focused on two or more of these factors (Hoang and Lim 2012). In addition to the above, there have also been studies of the effectiveness of viral marketing campaigns in which researchers have sought to categorize these campaigns using a range of typologies. Finally, there have been studies in which researchers have sought to develop prediction models. However, these have been few, and yet fewer still are those that have sought to predict virality using the viral coefficient – the industry measure of viral diffusion. This study seeks to contribute in this area – prediction modeling – by presenting a model that is focused specifically on diffusion of software.

Past research on software diffusion was focused on how pervasive a system was across organizations, that is to say, the focus was on the extent to which an implemented software was used throughout an organization and provided benefits to the users within said organization (Delone and McLean 2003). With migration to the Cloud, and software presenting as business ventures, as with SaaS, the focus of software diffusion is no longer within, but rather across, organizations. Viral diffusion is a relevant measure of software success – it indicates clear benefits as recommendations of, and the product itself, spread within and across social networks resulting in wide-scale adoption in multiple organizations. Viral diffusion can impact the success of SaaS ventures (Domingos 2005), and for investment purposes, it is considered key for evaluating SaaS and other startups (Jurvetson and Draper 1997).

The study focused on building a predictive model for software diffusion. The academic literature has developed diffusion models using various techniques including agent-based modeling and those equations developed by Henry Bass (Bass 1969); however, while the viral diffusion models in the literature provided several virality measures and prediction methods, a challenge facing marketers is how to use data from actual business scenarios to predict virality.

A stated importance of prediction is to enhance the ability of marketers to know, based on prediction, how much money they will need to spend in order to attain the necessary effectiveness of the viral campaign. For example, consider the company discussed in the introduction above. This company provides a SaaS product which experiences a viral coefficient, K=0.45. Recall that virality is achieved when K>1, in which case marketers would have no need to spend money if such growth occurred through involuntary recommendations, forwarding and adoption of the product. However, the viral coefficient, K at 0.45 necessitates the company spending on marketing. Necessarily, while spending, the goal will be to continually increase K so as to eventually spend less or no money at all. From this example, it is clear that the ability to predict the effectiveness of viral growth is thus important to plan, among other important considerations, the marketing budget and schedule that will be necessary to achieve the desired viral growth.

Following the review of the literature, two particular gaps that were thought necessary to be addressed were identified. The first was the need to develop a comprehensive, validated model that can be used to predict virality of SaaS products. The second was the need to understand the challenges that practitioners may face as they seek to customize a generic virality model to where it becomes relevant to their particular practice, and to document a set of practices that can help to address these challenges. To achieve these goals, a case study was conducted that used actual SaaS diffusion data to customize and refine an industry-based model -- the Jarvis (2017) model. The goal was to understand what other factors impact virality and how these may be addressed in model development. In the next section is a description of the existing model (Jarvis 2017a) built on. Next are details of the steps taken to develop the prediction model to contribute to this important research and industry gap. The results showed that a single model does not accurately predict virality and that adjustments have to be made for different periods and possibly, for different products. The study makes several contributions to the software diffusion literature and the viral marketing literatures. The study also contribute to practice by offering guidance to practitioners on refining prediction models, particularly those in organizations that are pursuing SaaS ventures.

The rest of this paper proceeds as follows. In the next section is a description of the background to the study including the background on viral diffusion, viral growth models and software diffusion. The method is described next, including the features of a model adopted, and the refinements made based on the existing limitations. This is followed by a discussion

of the model performance of as well as its limitations. Implications for research and practice are discussed at the end.

BACKGROUND - VIRAL DIFFUSION

2.1. Diffusion Definitions and Measures

Virality is often used in the viral marketing literature, and in industry, to capture the extent of diffusion. Virality has been defined and measured in several ways in the literature. It has, for example, been defined as the number of times content has been forwarded (Berger and Milkman 2010, Deza and Parikh 2015); the extent to which content has been endorsed (Cao et al. 2009); the probability that an item is purchased (Domingos 2005); the number of persons accessing content in given interval (Guerini, Strapparava, and Ozbal 2011); the speed of diffusion of content in and/or across social network(s) (Hildebrand, Hofstetter, and Herrmann 2012); the popularity of content (Hoang and Lim 2012, Jamali and Ester 2010, Scholz et al. 2017, Weng, Menczer, and Ahn 2013); the rate of viewership of video content (Jain et al. 2014, Liu 2012); the number of new customers that each existing customer is able to successfully convert to become a customer -- the viral coefficient (Jarvis 2017b, Hoang and Lim 2012); and exponential transmission of content or growth of a company (Pousttchi and Wiedemann 2007, Trusov, Bucklin, and Pauwels 2009, Wiedemann and Palka… 2008); among others.

Similar among the above is the idea that products or content designed to viral should experience large exposure and influence. The major differences are in the proxies used to measure virality in the studies – the factors that are measured to determine the extent of success achieved by viral marketing efforts.

Success Factors

The viral marketing literature includes studies that sought to identify the factors that lead to high diffusion rates for content designed for virality. For example, Poustchi and Wiedemann (2007) identified eight success factors, a few of which have been identified in the IS literature. The eight are: perceived usefulness by recipient; perceived ease of use; reward for communicator; free mobile viral content; initial contacts; first mover's advantage; critical mass; and scalability. Kaplan (2011) suggested that the attending message be edgy and memorable. Other research found that content must be designed to be forwarded via SMS or MMS -- at no cost to forwarders (Wiedemann and Haunstetter… 2008); that viral content should be easily opened and recipients assured of security (Wiedemann and Palka… 2008); that the more positive the content, the more viral it was likely to be, and that the relationship between emotion and social transmission is more complex than just valence -- arousal also played a role (Berger and Milkman 2010); that virality increased when certain topics were dominant - Animal, Synthetically Generated, Not Beautiful, Explicit and Sexual and that image content, rather than associated textual content (such as a title), was the prime signal in human perception of image virality (Deza and Parikh 2015); that virality is strictly determined by the nature of content rather than influencers who spread it and has many facets that only partially overlap (Guerini, Strapparava, and Ozbal 2011); and that virality was dependent on the structure of viral information dispersion and, contrary to the suggestions of Guerini et al (2011), the social, interpersonal influence of forwarders (Jain et al. 2014).

Virality Prediction Models

With the impact of e-WOM and social media marketing, a number of studies have

begun to focus on developing virality prediction models (Deza and Parikh 2015, Jarvis 2017b, Yu and Wang 2014, Mochalova and Nanopoulos 2015). Multiple methods have been used in these models; many of them seeking to predict diffusion rates, and many reporting a measure of success in their efforts. For example, Deza et al (2015) developed a prediction model that sought to identify the virality of content based on the images placed in the content. They tested their model using human subjects and machine experiments. They report that their findings showed good prediction accuracy; however greater with trained machines (68:10%) than with humans (60:12%). In addition to the nature of the content, they also tested the impact of key visual attributes such as context, neighboring images, recently viewed images and image title/caption.

Hoang et al (2012) developed a diffusion model using a predictive algorithm based on mutual dependency. They sought to measure the simultaneous impact of user virality, item virality, and user susceptibility on the quantity of retweets. Their model also considered the mutual dependencies. They claimed to measure both popularity -- the number of adopters, and viral coefficient -- the rate of adoption. In their measure of virality, they used a rate greater than one as an indication of virality. Validation of their model was one using both synthetic and real datasets. Their experiments showed that their model performed well for predicting those hashtags that have higher retweet likelihood.

Jain et al (2014)developed a predictive model using CrowdCast. CrowdCast applies online machine learning to map natural language. Their model included real-time responsiveness as a first-order requirement. They used the model to observe Twitter tweets and predict which YouTube videos were most likely to "go viral" in the near future. Their independent variable was the perceived "influence" of the sender. The spread of each video was predicted through a sociological model, derived from the emerging structure of the graph over which the video-related tweets are (still) spreading. Combining metrics of influence and live structure, CrowdCast output sets of candidate videos, identified as likely to become viral in the next few hours. They monitored Twitter for more than 30 days and found that CrowdCast's real-time predictions demonstrated encouraging correlation with the actual YouTube viewership they observed in the then-near future.

Liu (2012) developed a predictive model that tested content virality by tracking viewership of videos for 60 days based on historical evidence they found that stated that that many successful viral videos gained popularity in a very short period of time. To check the robustness of the model, the author compared the predicted versus actual cumulative views of the holdout sample during different time periods. Their findings show a high correlation between predicted and observed views, ranging from 0.76 to 0.97 for the holdout videos, with a mean correlation coefficient of 0.91.

Mochalova et al (2015) developed a model to test the impact of various non-intrusive seeding approaches. The idea was that, instead of activating seeds directly, marketers would do so by employing the power of user-to-user interactions. The authors compared three seed selection methods: degree centrality; betweenness centrality; and percolation centrality. Percolation centrality was found to be superior to the other methods. They found that users with high percolation centrality will act as bridges for further propagation. However, they note that care must be taken to ensure that the campaign does not become intrusive with increasing #waves of transmissions. more seeds are used to start a viral campaign, (this may be better used in propagation rather than prediction measurement.

Weng et al (2013) developed a model to predict the popularity of memes based on which memes would produce more tweets or become adopted by more users than a certain percentile threshold (70, 80, 90) of memes. Their results indicated that the future popularity of a meme can be predicted by quantifying its early spreading pattern in terms of community concentration such that, the more communities, the more viral.

Yu et al (2014) developed a prediction model to evaluate viral marketing efficiency within a given deadline. Specifically, the authors proposed two methods to generate deadline graphs -- Shortest-Distance method and Time-Iteration method, based on which, a Reverse Tree method is exploited to predict the probability that users would actually buy marketed products/services on Twitter, Friendster and Random. They used experiments to test their model. Their findings showed that their method -- deadline graph -- is key for evaluating viral marketing propagating efficiency within a given deadline, and it provides overwhelming advantages over traditional prediction methods.

Software Diffusion Models

While many of the models above could potentially be used in the prediction of software systems, the information systems (IS) literature includes various studies that have sought to explain software diffusion. One of the most popular theories used in the IS literature is the diffusion of innovations theory (DIT) (Rogers 2010) which explains that the rate of diffusion of an innovation, represented by a characteristic S curve, is influenced by five factors: relative advantage, compatibility, trialability, observability, and complexity (Rogers, 1995). Of these, the first four factors are generally positively correlated with rate of diffusion while the last, complexity, is negatively correlated. The actual rate of diffusion is argued to be governed by both the early and later spreads. Low cost innovations are expected to have a rapid take-off while innovations whose value increases with widespread adoption (network effects) may have faster late stage diffusion. Innovation adoption rates can, however, be impacted by other phenomena, for example a competing innovation can potentially inferior lock technologies in place. The theory also explains diffusion in terms of the nature of the individuals adopting throughout the product diffusion in terms of the different degrees of willingness to adopt that they possess. This is expected to make them approximately normally distributed over time, a normal distribution which divides them into adopters that display five levels of innovativeness (from earliest to latest adopters). These levels include the: innovators - venturesome, educated persons with multiple information sources; early adopters - social leaders, popular, educated; early majority- deliberate adopters who have many informal social contacts; late majority - skeptical, traditional individuals from the lower socio-economic status; and laggards – individuals whose neighbors and friends are their main information sources and who tentatively adopt, ruled by a fear of debt. A number of IS researchers have conducted studies using DIT, many of them modifying the theory to explain its applicability to technology innovations in particular. In many cases, these researchers have consistently agreed on the impact on diffusion of three DIT factors – compatibility, complexity and relative advantage (Cooper and Zmud 1990, Crum, Premkumar, and Ramamurthy 1996, Agarwal and Prasad 1998).

Another theory used widely to explain adoption and also diffusion of software is TAM – the technology acceptance model (Davis 1989). TAM posits that the rate of acceptance (and by extension, diffusion) of technology is determined by two factors related to a technology – the *perceived ease of use* -- "the degree to which a person believes that using a particular system would be free from effort," and the *perceived usefulness* of the innovation -- "the degree to which a person believes that using a particular system would enhance his or her job performance" (Davis 1989). These factors concur with those of the DIT, but overall, the theory offers less explanation of the diffusion phenomenon, and has actually been thought to account for only 40% of the spread (Legris, Ingham, and Collerette 2003). Another factor that has been shown to be critical in the literature is the extent to which the innovation “fits” user tasks (Goodhue and Thompson 1995, Serrano and Karahanna 2016).

Other models that have been more directly focused on measuring diffusion is the Bass

diffusion model (Bass 1969). The Bass model, which contributed some mathematical ideas to the concepts proposed by the DIT, is derived from a simple premise that the conditional probability of adoption by a randomly chosen consumer at time T given that adoption has not yet occurred, is a linear function of the number of previous adopters. The model consists of a simple equation: $\frac{f(t)}{1-f(t)} = p + qF(t)$, where:

- $f(t)$ represents the rate of change of the installed base fraction;
- $F(t)$ represents the installed base fraction;
- p represents the coefficient of innovation; and
- q represents the coefficient of imitation.

The equation essentially describes the spread of innovation within a population, and the relationships between current and new adopters. Similar to (though not as detailed as) DIT, the Bass Model classifies adopters as innovators and imitators (later adopters) shown by their speed of adoption (their level of innovativeness). This model has been particularly useful in predicting the sales/diffusion of consumer products including software technology innovations (Jiang and Sarkar 2009, Lee and Tan 2013, Robertson and Gatignon 1986).

While the Bass model remains popular in the academic literature, a number of key factors are noted as missing from an industry practice perspective. These factors are included in the model proposed by Jarvis (2017b) discusses a number of factors related to viral growth in an industry-facing blog. Jarvis built on previous models presented in seminal blogs by Skok (2009) and Chen (2008), using as the measure of virality, the viral coefficient, K discussed in many of the models above. The Jarvis-Skok model involves a set of equations for the measurement of virality. The model factors include the number of invitations sent by the seed crowd, the conversion rate for invitations, cycle time, churn, and market size. The model is discussed below in greater detail in the methods section.

RESEARCH METHOD

The model built was built on the core growth model proposed by Chen (2008) and Skok (2009) and later enhanced by Jarvis (2017). The method involved fitting the data to the model to see where limitations in the existing set of equations lie concerning the viral diffusion of the SaaS product used for the study. Then, having observed the limitations, the challenges met were recorded and variants tried that would address those challenges in an effort to adjust the existing model to remove the limitations found.

Research Data

Data used are related to an industry case of a software as a service (SaaS) app to which access had been given. The company is referred to as CompanyX.

Research Model

The study started by considering the Chen/Skok model. In this model, a single equation representing viral growth/diffusion is based on the simple formula below:

1. $c^t = c^{t-1} * K + c^{t-1}$

Where:

- **c** represents the number of customers
- **K** represents the viral coefficient
- **t** represents the number of viral cycles that have been completed.

The diffusion computed using this basic formula is expected to be repeated each viral cycle,

for however long the viral cycle measured. For example, if the viral cycle is 10 days, based on the simple formula above, the expectation is 36.5 viral cycles each year with growth each 10 days measured as per the formula. The formula was later adjusted by Jarvis (2017a) to account for other critical factors that impact the spread of innovations. The equations involved in the revised model are below:

1. $\mathbf{AR} = \mathbf{R} - [\mathbf{R}\,\mathbf{x}\,(\mathbf{CD}/\mathbf{M})]$: Adjusted conversion rate (AR)
2. $\mathbf{K} = \mathbf{I}\,\mathbf{x}\,\mathbf{AR}$: Viral coefficient (K) - Simplistic calculation of the viral coefficient. Assumes the % customers that make invitations is set
3. $\mathbf{VIF} = \mathbf{U}(\mathbf{t}+\mathbf{1}) = [\mathbf{U}(\mathbf{t}) * (\mathbf{K} * (\mathbf{1}-\mathbf{L}))]^{[((\mathbf{t}\,/\,\mathbf{ct})+\mathbf{1})-\mathbf{1}]\,/\,[(\mathbf{K}\,\mathbf{x}\,(\mathbf{1}-\mathbf{L}))-\mathbf{1}]}$: Viral invitation factor (VIF) – The number of new users that come from a cohort adjusted for churn
4. $\mathbf{EVGF} = \mathbf{K}\,\mathbf{x}\,(\mathbf{1}-\mathbf{L})$: Effective V-Growth Factor (EVGF) - Viral coefficient less churn rate, on the first iteration, where:
    - **M** represents the market size
    - **K** represents the viral coefficient
    - **I** represents the number of invitations sent by adopters
    - **R** represents the rate at which adopters are converted
    - **AR** represents the conversion rate adjusted for churn
    - **ct** represents the Viral cycle time -- the time it takes for referral to turn into a customer
    - **t** represents the number of viral cycles that have been completed
    - **L** represents churn

Model Limitations

One challenge met in attempting to apply the industry data to the model is that measuring viral cycle time is not that easy. In the above formula, the assumption is that cycle time is a discrete number and users all "infect" new users right at the cycle time (e.g., at each 10th day in the example above), or at least, that there is a standard bell curve distribution around each 10 day mark. Contrary to such an assumption, two primary and two secondary issues arose in the data. First, viral cycle time can stretch on for weeks, months or even years. In industry, one sees that if you have a six month cycle time and your model assumes there that is no virality until the six month mark, it is going to be very inaccurate until you hit the six month mark as you will be experiencing virality before the model says you should be.

The second issue found was that the distribution curve is often not a "tight" curve right around the viral cycle time. Again, that data show that there was a lot of infections almost immediately and then a long tail of minimal infections. For example, you may therefore have a six month cycle time and yet you get a big spike on day one followed by a long tail of a small amount of infections, so that, ultimately, your median infection time may be six months, but your actual business results, especially as measured on a month by month basis, will vary significantly from the standard model.

In addition to these issues, a few secondary issues were observed. First, K often was not observed to be constant over time. For example, at CompanyX the percentage of users who created a team in their first year on the platform increased by 240% (you had 2.4 times the previous number of users) during the first five years the company was incorporated. This number then fell by 10%.

The hypothesis is that in the early days of the system, the software was a little more rudimentary, so users were less enthusiastic about inviting other potential users to the platform. During this time, CompanyX started to see wide adoption of the mobile platform which for many users was key to increasing value of the platform.

The hypothesis was that the more recent decline in virality (i.e., team creation percentage) is due to saturation in key geographic markets, particularly within top sports. In other words, virality burns out in those areas as everyone is "infected." Barring any changes in the business strategy or business model, this trend would be expected to continue.

Another secondary issue is that K may vary during the year due to seasonal factors. CompanyX's business is highly season. Certain sports are more active at certain times of years and each sport has its unique characteristics. In addition, there are certain "off season" months where teams are often are prepping for the season but are less active. Examining the data from 2010 to 2017, one could see a clear pattern where users created in months that typically were associated with the off season (May through July and November through December) are less likely to create their own teams.

RESULTS

Below are the results of the study, illustrating the success met with respect to meeting the research goals.

Model Accuracy

Actual virality was 1.41 times the model in 2017 and 1.20 times -- 120% more accurate than -- the model in 2018. The variance to actuals occurred in two places. First, the model underpredicted virality. The reason for this is that in the existing model, the probability that someone would create a team (and invite others) within their first 36 months on the platform. The problem is that this only captures 80.4% of the virality (measured so far). In other words, 19.6% of time when a user creates a new team, it is outside the 36-month window. These "missing" 19.6% in turn causes the model to underpredict the amount of actual virality.

Model Adjustments

To address these issues, several modifications were made to the models described by Jarvis (2017b) and Skok (2009). First, the viral cycle time was broken down into sub-segments. For example, a multi month cycle time was broken down in 36 sub periods. Next is a look at how many users were infected in sub period 1, sub period 2, and etcetera. The decision to break the cycle time down into 36 periods was driven purely by business reporting interests. 36-Month long periods were examined. However, one could take the cycle time and break it down into as many sub-segments as the data allowed. The key is having the underlying data to understand infection rates for each sub period. By breaking the infections down into all these sub periods, the method effectively accounted for distribution curves that are irregular and not highly centered around the viral cycle time.

A separate version of the model was also created to take into account changes to the viral coefficient over time and monthly seasonality. Jarvis' (2017a) model allows for a changing viral coefficient but provided no guidance on how to effect the required changes, and other authors, for example Skok (2009), failed to touch on this issue. In the model, historical data for K was used to extrapolate what K might look like into the future. Finally, a seasonal component was added. To do so, the method looked at how the viral coefficient changed month by month using seasonal patterns and applied this on a "going forward" basis. That said, changing virality over time and adding seasonality did not increase accuracy. Instead it raised additional issues for further study as will be discussed.

DISCUSSION

This study had two research goals: 1) to develop a comprehensive, validated model that can predict accurately virality – in particular of SaaS products; and 2) to understand how to refine an existing virality model -- what factors impact virality and also how these factors may be addressed in adjusting the base model -- and to record the lessons learned from the experience had. The study has several implications for research and practice. These came by way of the several challenges met in seeking to first fit the data to the existing model and then adjust the model to account for lack of fit. These served as the lessons learned concerning model development and refinement which is thought can serve as guidance for practitioners seeking to develop models to predict virality of products being contemplated or being marketed. Below, these implications are presented. For practitioners, these are presented as challenges which are first described and followed with a discussion of how they may be overcome in a subsection called "*Solutions*." The paper ends with a discussion of the study limitations and some guidance for future research.

Implications for Research

The research contributes to the literature by developing and validating a model that can be used to more accurately predict virality than others so far seen in the literature. This was done by introducing factors that before had not been considered. One such factor was seasonality, the concept of sub-segments for the viral cycle time and changing virality over time. These may also be particularly helpful in customizing the model to the business for practitioners who would engage in measuring and predicting virality. Further, the study showed how a dataset can be used to facilitate the customizations mentioned.

Implications for Practice

In industry, virality is very important in marketing budgeting and many other areas of a business. The study contributes to practice both by making further adjustments to Jarvis' (2017) model and by recording the challenges faced in development and offering solutions to those challenges that practitioners may heed.

Some of the model refinement challenges encountered include: 1) determining the data collection period; 2) dealing with the time impact on product virality; and 3) capturing specificities regarding factors that may impact the particular product -- for example, seasonality -- discussed in greater detail below.

With respect to the data collection period, in the specific case, the study used data from 2009 to 2016 as inputs to determine the virality. In other words, the study used that data to compute virality – what was the likelihood that a user would create their own team and invite additional people to CompanyX. The revised model was then used the along with the computed virality to predict results for 2017 and 2018. The challenge from a modeling standpoint is two-fold. First, it takes a very long time to collect this data. Waiting more than 36 months at an Internet company is a lifetime. Second, the tail just keeps growing and growing. For example, CompanyX has been around for 10 years and some users have created their first team after waiting 10 years.

With respect to the time impact on virality, while the study could extend the model beyond 36 months, it would then face another challenge. That is, the issue that rs cited that virality changes over time. Take for example, a look at virality over say 60 months would capture 96.6% of the virality. (Note that this percentage will decrease as the cohorts age and more people create teams outside of the 60-month window.) The issue is that if virality is changing over time, the virality data from much older cohorts that may not be predictive of

the future. While CompanyX's virality cycle is particularly long, other companies do have similar elements. A piece of content can be shared well after a user first sees it. Similarly, it can take a while for a recipient of that shared content to act on the received content.

Finally, with respect to other factors that may impact the viral diffusion of a software product, the other challenge that came up is seasonality – the findings show that most of the variance to the model is at certain times of the year. When the study included the standard model -- the one not adjusted for seasonality -- the biggest variance is consistently during several months of the year. In Figure 1, one can see that there are certain months where the model overpredicted virality both in 2017 and 2018 and other months where it underpredicted. To address seasonality concerns, a variant of the model that had a seasonal element to try to capture it was created. Specifically, the virality was adjusted for when the user was created. Users who are created at certain times of year create teams on a different schedule than users created at other times of year. However, it was quickly realized that this was too simplistic a model. Seasonality, at least in the case of CompanyX, is based on both when a user is created and the month involved. Said another way, certain months are peak months for people to create new teams (usually peak sports months). As well, the time at which a user was created during the year can have an impact on their predicted schedule for creating a team in their first 36 months.

While CompanyX's seasonality is unusual, it is thought that most sites would have a similar element. For example, the time of day or day of the week when you view content may impact your likelihood of sharing that content with others. For example, if you view it late at night, you might be less inclined to share that content for fear of disturbing the recipient.

In summary, in industry, virality is often described in simplistic terms. Your users "infect" other users and get them to join your platform, thereby driving your growth without you, and consequently reducing the amount you have to spend on paid marketing. Being able to understand how much virality you can expect and how to optimize that virality is core to this strategy.

Skok (2009), Jarvis (2017), and Jurvetson et al (1997) discuss how virality can drive your business. Skok provides high-level equations and Jarvis drives those equations into an Excel model. However, in practice, one realizes that some of the core variables in virality are quite complicated. For example, a viral coefficient can change over time and even change month to month. Furthermore, the distribution curve for how long it takes to infect other users can vary from business to business. This distribution curve also often has a long tail to it, making it hard to build a model while the business is still developing. Often what is being created is a model without enough years to really know how virality will change in the long run.

In terms of solutions, in order to accurately predict virality and address the challenges outlined above, the recommendation is that practitioners should ensure the following are done:

1. Accurately measure viral coefficient including whether it is changing over time. For example, is virality increasing due to an improved product or decreasing due to saturation?
2. Balance these historical trends in virality with the importance of using more fresh data. If you use years and years of historical virality data you may get a better picture of trends, but you may also not be reflecting what is happening in your business today.
3. Break the virality down into time periods within one viral cycle. For example, if the median time to infect is one week, a more accurate viral model will look at how many users are infected each day over a period of several weeks.
4. Look at what factors impact virality (e.g., seasonality) and do their best to capture those

factors in the model or at least note how where you are likely to see variances.

Limitations and Future Research

Based on the work done and from the results presented above, there are two main limitations observed. First, there is more work to be done to figure out how to better model team creation beyond the 36-month window while at the same time not making the virality data so stale as to be unusable. Second a better way to model seasonality needs to be tackled. Specifically, it would be important to capture all the inputs to build a more robust seasonality model – for example, the time at which a user was created.

The model development had other limitation. First, the data used were from a single company related to a single software and specific to a particular purpose. This limits the generalizability of the research findings and their application across models and in industry. Future research could seek to test and validate the prediction model with data related to other SaaS products. Another is the reliance on quantitative data to refine the model. Future research could involve qualitative data, particularly from adopters, to understand more of the challenges and refine the model to improve the prediction accuracy.

REFERENCES

Agarwal, Ritu, and Jayesh Prasad. 1998. "A conceptual and operational definition of personal innovativeness in the domain of information technology." *Information systems research* 9 (2):204-215.

Bass, Frank M. 1969. "A new product growth for model consumer durables." *Management science* 15 (5):215-227.

Berger, J, and KL Milkman. 2010. "Social transmission and viral culture."

Cao, J, T Knotts, J Xu, and M Chau. 2009. "Word of mouth marketing through online social networks."

Chen, A. 2008. "Facebook viral marketing: When and why do apps "jump the shark?"." https://andrewchen.co/facebook-viral-marketing-when-and-why-do-apps-jump-the-shark/.

Cooper, Randolph B., and Robert W. Zmud. 1990. "Information technology implementation research: a technological diffusion approach." *Management science* 36 (2):123-139.

Crum, Michael R, G Premkumar, and K Ramamurthy. 1996. "An assessment of motor carrier adoption, use, and satisfaction with EDI." *Transportation journal*:44-57.

Davis, Fred D. 1989. "Perceived usefulness, perceived ease of use, and user acceptance of information technology." *MIS quarterly*:319-340.

Delone, William H, and Ephraim R McLean. 2003. "The DeLone and McLean model of information systems success: a ten-year update." *Journal of management information systems* 19 (4):9-30.

Deza, A, and D Parikh. 2015. "Understanding image virality."

Domingos, P. 2005. "Mining social networks for viral marketing."

Goodhue, Dale L, and Ronald L Thompson. 1995. "Task-technology fit and individual performance." *MIS quarterly*:213-236.

Guerini, Marco, Carlo Strapparava, and Gozde Ozbal. 2011. "Exploring Text Virality in Social Networks." *https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/index*. doi: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2820.

Hildebrand, Christian, Reto Hofstetter, and Andreas Herrmann. 2012. "Modeling viral marketing dynamics in social networks–Findings from computational experiments with agent-based simulation models."

Ho, JYC, and M Dempsey. 2010. "Viral marketing: Motivations to forward online content."

Hoang, TA, and EP Lim. 2012. "Virality and susceptibility in information diffusions."

Jain, Puneet, Justin Manweiler, Arup Acharya, and Romit Roy Choudhury. 2014. "Scalable social analytics for live viral event prediction." Eighth International AAAI Conference on Weblogs and Social Media.

Jamali, Mohsen, and Martin Ester. 2010. "A matrix factorization technique with trust propagation for recommendation in social networks." Proceedings of the fourth ACM conference on Recommender systems.

Jarvis, A. 2017a. "How to model viral growth at your startup." https://www.alexanderjarvis.com/2017/11/03/model-viral-growth-startup/.

Jarvis, A. 2017b. "How to model viral growth at your startup -." *@adjblog*.

Jiang, Zhengrui, and Sumit Sarkar. 2009. "Speed matters: The role of free software offer in software diffusion." *Journal of Management Information Systems* 26 (3):207-240.

Jurvetson, S, and T Draper. 1997. "Viral marketing."

Kaplan, AM, and M Haenlein. 2011. "Two hearts in three-quarter time: How to waltz the social media/viral marketing dance."

Lee, Young-Jin, and Yong Tan. 2013. "Effects of different types of free trials and ratings in sampling of consumer software: An empirical study." *Journal of Management Information Systems* 30 (3):213-246.
Legris, Paul, John Ingham, and Pierre Collerette. 2003. "Why do people use information technology? A critical review of the technology acceptance model." *Information & management* 40 (3):191-204.
Liu, Y. 2012. "Seeding viral content: The role of message and network factors."
Mochalova, Anastasia, and Alexandros Nanopoulos. 2015. "Non-intrusive Viral Marketing Based on Percolation Centrality." ECIS.
Palka, W, and K Pousttchi. 2008. "Understanding the determinants of mobile viral effects-towards a grounded theory of mobile viral marketing."
Pousttchi, K, and DG Wiedemann. 2007. "Success factors in mobile viral marketing: A multi-case study approach."
Robertson, Thomas S, and Hubert Gatignon. 1986. "Competitive effects on technology diffusion." *Journal of Marketing* 50 (3):1-12.
Rogers, Everett M. 2010. *Diffusion of innovations*: Simon and Schuster.
Scholz, Christin, Elisa C Baek, Matthew Brook O'Donnell, Hyun Suk Kim, Joseph N Cappella, and Emily B Falk. 2017. "A neural model of valuation and information virality." *Proceedings of the National Academy of Sciences* 114 (11):2881-2886.
Serrano, Christina, and Elena Karahanna. 2016. "The Compensatory Interaction between User Capabilities and Technology Capabilities in Influencing Task Performance: An Empirical Assessment in Telemedicine Consultations." *Management Information Systems Quarterly* 40 (3):597-621.
Skok, D. 2009. "Lessons Learned – Viral Marketing."
Trusov, Michael, Randolph E Bucklin, and Koen Pauwels. 2009. "Estimating the dynamic effects of online word-of-mouth on member growth of a social network site." *Journal of Marketing* 73 (5):90-102.
Weng, Lilian, Filippo Menczer, and Yong-Yeol Ahn. 2013. "Virality prediction and community structure in social networks." *Scientific reports* 3:2522.
Wiedemann, DG, and T Haunstetter…. 2008. "Analyzing the basic elements of mobile viral marketing-an empirical study."
Wiedemann, DG, and W Palka…. 2008. "Understanding the determinants of mobile viral effects-towards a grounded theory of mobile viral marketing."
Yu, Li, and Nan Wang. 2014. "Predicting viral Marketing Propagating Efficiency within given Deadline." PACIS.

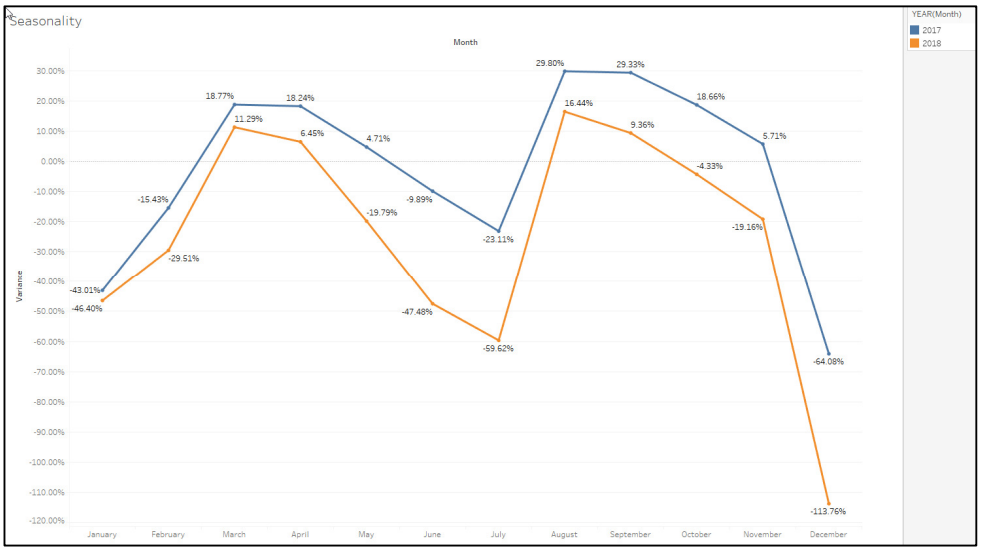

Table 1: Seasonality Impact