Essays on Marketing and Online Social Networks

Inauguraldissertation zur Erlangung des akademischen Grades eines Doktors der Wirtschaftswissenschaften der Universität Mannheim

vorgelegt an der Fakultät für Betriebswirtschaftslehre der Universität Mannheim

Andreas U. Lanz, M.A.

Mannheim im März 2017

Dekan: Prof. Dr. Dieter Truxius

Erster Gutachter: Prof. Dr. Florian Stahl

Zweiter Gutachter: Prof. Dr. Florian Kraus

Tag der mündlichen Prüfung: 2. Juni 2017

"In God we trust; everyone else must bring data."

–

William Edwards Deming

Contents

List of Figures

List of Tables

1 General Introduction

Standing on the brink of the fourth industrial revolution, the whole business landscape is rapidly transforming, and companies are trying to adapt to a new state of the world in which value chains are becoming progressively digitized. This not only allows agile companies to realize competitive advantages against digital laggards, but also brings into being a whole new range of business opportunities. The unprecedented speed of current innovations is fueled by the increasing connectedness of markets, declining technology prices, and the ubiquity of data – where especially the latter has opened up vast new research opportunities in the social sciences.

Marketing in particular benefits from an increasing volume and variety of data, to a large extent based on the emergence of social networking platforms that map social systems all around the world and, by their construction, reveal behavior and preferences on an individual level. In fact, individuals use such platforms to serve a multitude of purposes in their daily lives. On the one hand, as consumers, they compare, purchase, and then review products online. On the other hand, social networking platforms allow individuals to act as producers in marketing and distributing their own products, services, content, or ideas. These contrasting functions generate granular and encompassing data about individuals acting as both consumers as well as producers, which enables marketing researchers to study behaviors, test predictions, and derive better-informed managerial insights. Thus far, however, predictions do not play a dominant role in the marketing domain, as most research focuses on disentangling specific effects without putting them in relationship. Research should not only be geared towards understanding individuals and aggregate behavior, but it also should aim at predicting the respective phenomenon. Even more importantly, the generalizability of the developed methods and generated insights imply on managerial relevance in addition to the reputation of marketing within social sciences, notably economics, information systems, and psychology. Accordingly, the two scientific essays embedded in this dissertation represent an attempt to live up to these postulations, which are also highlighted in the current Marketing

Science Institute's (MSI) *Research Priorities* summary (2016-2018).

Nowadays, most companies are struggling to cope with the different aspects of data – in particular, its increasing volume, variety, and velocity. To capitalize on such data and to create sustainable competitive advantages, companies need to heavily invest in the latest technologies, an efficient infrastructure, and in highly skilled human resources (Skiera 2016). Outstanding business intelligence capabilities also require marketers to adapt to the changing landscape, namely by embracing complexity and acquiring in-depth knowledge about theories and frameworks, constantly improving and re-evaluating the applied methods and metrics, further developing integrative skills, and thinking holistically as well as out of the box (Lemon 2016). Quantitative marketers and researchers alike mostly use experiments and observational studies for analytics. The former uncovers causal relationships, whereas the latter depends strongly on theory to derive valuable insights. In the context of social networking platforms, a per se highly endogenous environment, establishing causal relationships is challenging since the connectivity among users can make an experimental control group prone to contamination (for an overview about randomized trial designs, see Walker and Muchnik 2014). Moreover, experiments in the context of social networking platforms have been less frequently published due to the extremely difficult data acquisition, although enjoying increasing popularity amongst researchers (e.g., Aral and Walker 2012, 2014; Muchnik, Aral, and Taylor 2013). However, decades of building and empirically validating (social) network theory serves as a solid foundation for observational studies to derive insights, and has thus led to extensive research on underlying drivers of information dissemination (e.g., Ansari et al. 2016; Ansari, Koenigsberg, and Stahl 2011), along with optimal seeding policies (e.g., Aral, Muchnik, and Sundararajan 2009; Hinz et al. 2011). Despite the huge buzz around *Big Data*, larger datasets are not necessarily better, above all with respect to representativeness. Instead, marketing researchers predominantly rely on rich representative samples of data for analysis.

The focus of this dissertation is on user-generated content networks, which constitute a unique type of social networking platform (e.g., Goldenberg, Oestreicher-Singer, and Reich-

man 2012; Mayzlin and Yoganarasimhan 2012; Trusov, Bodapati, and Bucklin 2010). On user-generated content networks – the most prominent ones being Youtube (videos), SoundCloud (music), Instagram (pictures), and Twitter (tweets) – users upload their content (e.g., videos, music, pictures, and tweets) that in turn can be consumed by other users of the network. Therefore, user-generated content networks are essentially platforms for creators of content and their fan communities. These platforms offer various functionalities to exchange information between users, which allows creators to actively engage with fans and to acquire new ones. Some creators become social media celebrities like Casey Neistat, who over the course of around 700 daily vlogs received more than a billion views on Youtube, earning a considerable amount of money through paid-for social media endorsements (Bond 2016). On SoundCloud, there are similar success stories, where certain artists quickly accumulated followers as a result of their chosen seeding policy and uploaded content (Lilly 2014). Based on unique datasets of SoundCloud, this dissertation sets out (1) to investigate optimal seeding policies (e.g., as a creator of music) and (2) to create a managerial framework for early-stage investments in creators (e.g., as a record label). Each of the two essays embedded in this dissertation covers one of these topics, and as such they are two sides of the same coin. More precisely, the former takes the perspective of the creator – i.e., how to quickly accumulate followers after signing up to a user-generated content network – and the latter takes the perspective of the investor – i.e., how to invest in these recent sign-ups. This dissertation is based on unique datasets provided within the scope of a research collaboration with SoundCloud, a platform that shares the same basic functionalities as Youtube, Instagram, and Twitter. Hence, the derived insights can potentially be generalized to other user-generated content networks.

The first essay, co-authored with Jacob Goldenberg, Daniel Shapira, and Florian Stahl, is entitled *"Climb or Jump – Status-Based Seeding in User-Generated Content Networks"* and challenges the role of individuals with a high network status. This generally belongs to influencer marketing, a topic currently attracting a great deal of attention in business practice (Maheshwari 2016). The academic literature on seeding shares the common view

that the higher the network status is of the target, the more effective the seeding or buzz program (Gladwell 2000; Iyengar, Van den Bulte, and Valente 2011; Katz and Lazarsfeld 1955; Rogers 1995; Valente 1995; Van den Bulte and Joshi 2007). Moreover, the predominant view on dissemination dynamics is built on a strong assumption that the responsiveness of individuals with a high network status, i.e., the probability of responding to targeted promotional actions, is equal or similar to any other individual on the social networking platform (e.g., Hinz et al. 2011; Yoganarasimhan 2012). In user-generated content networks, it is not clear whether this assumption holds: Why should someone with high status treat endorsement requests by individuals with high and low status equally? Using data from SoundCloud, we study creators of music who seek to build and increase their follower base by directing promotional actions to other users of the networking platform. Along these lines, the two focal research questions are as follows:

(1) Considering an unknown creator of content who seeks to build and increase his or her follower base on a user-generated content network, what measures should be taken to reach this goal? (2) What is the optimal policy to attract followers and, thus, who on the social networking platform should a creator of content target in order to propagate relevant information or content into the network?

Focusing on the network status of both creator and seeding targets, we find that, in particular, unknown creators of music do not benefit from seeding high-status users or influencers. In fact, it appears that unknown creators should ignore predominant seeding policies and slowly "climb" across status levels of seeding targets, rather than attempt to "jump" towards those with the highest status. Our research extends the existing seeding literature by introducing the concept of risk to dissemination dynamics in online communications. We show evidence that unknown creators of music do not seed specific status levels, but rather choose a portfolio of seeding targets while solving risk versus return trade-offs. From this, we derive managerial implications for information dissemination and optimal seeding in user-generated content networks.

The second essay, also co-authored with Jacob Goldenberg, Daniel Shapira, and Florian Stahl, is entitled *"Allocation of Marketing Budget When Success Is a Rare Event"* and addresses the managerial decision problem of allocating marketing resources to a set of creators (e.g., creators of music) who recently signed up on a user-generated content network (e.g., SoundCloud). Considering that the success of a creator is a rare event, only few creators ultimately become best-selling. Consequently, managers (e.g., of a record label) face a complex allocation problem of how to invest and allocate their marketing budgets among multiple creators. If they decide to invest in all creators, each future best-selling creator received initial support. However, from an economic perspective, this policy is associated with considerable wasted investments due to the extensive resources allocated to the many creators that ultimately turn out to be failures. At the other extreme, managers could decide to invest in a single creator. In this case, they face high lost returns since it is almost certain that such a focused investment of marketing resources will not capture all the future best-selling creators. This trade-off leads to the following question:

In how many and, more precisely, in which creators of content should a manager invest in an early stage of the life cycle?

We propose a novel managerial framework for maximizing profits when deciding on how many as well as which creators to select for investment in an early stage of their life cycle, by solving the economic trade-off between wasted investments and lost returns. We test our proposed framework for creators of music and generate investment recommendations for record labels on the SoundCloud platform. Our analyses demonstrate that managers' investment decisions in the context of rare events are mostly inefficient, as the profit realization of our proposed framework is up to four times greater than currently used selection policies.

The remainder of this dissertation is organized as follows. Chapters 2 and 3 present scientific working papers and contain the above-introduced essays, followed by a general conclusion in Chapter 4 that summarizes and discusses the insights and implications of this dissertation.

2 Climb or Jump – Status-Based Seeding in User-Generated Content Networks

2.1 Introduction

In the last decade, user-generated content networks and social networking platforms like YouTube, SoundCloud, and Instagram have become ubiquitous and now capture a substantial part of the social media sphere. On these platforms, the content is generated and offered by individuals, small groups, and firms that are interested in promoting their own creations as well as their own network status, and in some cases, their career (e.g., Goldenberg, Oestreicher-Singer, and Reichman 2012; Mayzlin and Yoganarasimhan 2012; Netzer et al. 2012; Trusov, Bodapati, and Bucklin 2010). A well-known example is the Dutch electronic music artist San Holo, who focused all his self-promotion efforts on SoundCloud, a user-generated content network in the music domain with 175 million users (Pierce 2016). His efforts paid off, resulting in more than 2,000,000 plays and growth in his follower base from 4,000 to over 40,000 SoundCloud users (Voogt 2015).

Considering an unknown creator of content (e.g., San Halo) who seeks to build and increase his or her follower base on a user-generated content network, what measures should be taken to reach this goal? What is the optimal policy to attract followers and, thus, who on the social networking platform should a creator of content target in order to propagate relevant information or content into the network?

This problem generally belongs to influencer marketing, a topic currently attracting a great deal of attention in business practice (Maheshwari 2016). Influencer marketing and the academic literature on seeding share the common view that the higher the network status is of the target, the more effective the seeding or buzz program (Gladwell 2000; Iyengar, Van den Bulte, and Valente 2011; Katz and Lazarsfeld 1955; Rogers 1995; Valente 1995; Van den Bulte and Joshi 2007). Hence, a creator of content who seeks to build and increase his or

her follower base on a user-generated content platform should direct promotional actions to individuals with a high indegree, a common yet basic operationalization of network status (Ball et al. 2001; Hu and Van den Bulte 2014; Sauder, Lynn, and Podolny 2012). However, in this paper, we show that high-indegree seeding is in fact often inefficient, and we demonstrate and recommend a more effective policy, which is somewhat counterintuitive: Optimal seeding in user-generated content networks is achieved by approaching users with low status.

The predominant view on dissemination dynamics is built on a strong assumption that the responsiveness of individuals with a high network status, i.e., the probability of responding to targeted promotional actions, is equal or similar to any other individual on the social networking platform (e.g., Hinz et al. 2011; Yoganarasimhan 2012). In user-generated content platforms, it is not clear whether this assumption holds: Why should someone with high status equally treat endorsement requests by individuals with high and low status? If we take into account that the probability of responding to an endorsement request is dependent on the network status, optimal seeding policies in user-generated content networks may completely change. Indeed, we show that (1) in almost all cases, responsiveness is a function of the status differences between the sender and receiver of endorsement requests, and (2) due to this difference, targeting high-status individuals is a significantly inferior seeding policy compared to targeting low-status individuals.

In this paper, we study commonly used endorsement requests in user-generated content networks – *promotional actions* including *follows*, *private messages*, *reposts*, *comments*, and *likes*. In the context of SoundCloud, we first analyze the responsiveness of seeding targets, namely, when creators of music send promotional actions to other users of SoundCloud in order to get *follow-backs*. We find that the higher the difference of network status between the creator and the seeding target, the lower is the a priori probability of a response. Furthermore, we analyze the creator's return on a targeted promotional action and find that the return scheme is composed of two sources: the *direct return* (the follow-back from the seeding target) and the *indirect return* (the number of followers from the seeding target's follower base). Moreover, we find that the higher the network status of the seeding target,

the higher the indirect return. Hence, the return on high-status individuals is higher than on low-status individuals, which is in line with common social network literature in marketing. However, the return scheme changes once we consider returns based on different levels of responsiveness. For unknown creators who seek to build and increase their follower base, high-status individuals are therefore associated with *very low responsiveness* but potentially high return, whereas low-status individuals are associated with *much higher responsiveness* but relatively low return.

In addition, we analyze how creators of music on SoundCloud distribute promotional actions over seeding targets with different network status by taking into account that the individual "budget" of promotional actions is constrained (first, the time during which creators consider and exert promotional actions to build and increase their follower base is limited; second, creators face search costs to find seeding targets as well as anti-spam policies, which limit their self-promotion efforts). In this sense, our research further extends existing seeding literature. Contrary to a common assumption that the limit on seeding resources is not confined, we account for this constraint and show how budget considerations affect the selection of seeding targets. Drawing on the von Neumann–Morgenstern utility function (e.g., Archak, Ghose, and Ipeirotis 2011; Eliashberg 1980; Von Neumann and Morgenstern 1947), we show evidence that creators of music on SoundCloud make their seeding decisions while taking into account a risk versus return trade-off. Considering their portfolios of seeding targets in a given time period, we find that a fraction is "invested" in SoundCloud users with a high status difference compared to the creator of music under consideration, which is associated with *high risks* (due to low responsiveness) but potentially high returns. The remaining fraction of the portfolio is "invested" in SoundCloud users with a lower status difference, which is associated with *lower risks* (due to higher responsiveness) but relatively low returns.

Finally, we analyze the creators' aversion to risk, which affects the individual portfolio choice of seeding targets. We find that creators' portfolio choices of seeding targets are predominantly risky: Instead of slowly "climbing" the ladder of status levels of seeding

targets, they attempt to "jump" towards those with the highest status – and keep failing to effectively accumulate followers to build and increase their fan communities.

The remainder of the paper is organized as follows. An overview of the relevant literature is presented in the subsequent section followed by the data description, theoretical reasoning, and empirical findings, as well as a simulation study about optimal seeding policies. We conclude our paper with a discussion of our findings and implications for marketing and online communications practice.

2.2 Background

This paper is related to three broad streams of literature. One stream focuses on information dissemination within social networking platforms. Another studies optimal seeding policies and their implications on viral marketing, and a third stream of literature concentrates on the domain of social psychology, which investigates status differences and the resulting inter- and intragroup behaviors. We discuss the related literature with a focus on user-generated content networks, which constitute a unique type of social networking platform.

2.2.1 Information Dissemination

User-generated content networks offer companies as well as individuals new opportunities to increase their follower base and, hence, brand awareness. On one hand, they can build and increase their follower base by means of paid advertisements. On the other hand, promotional actions in the form of follows, messages, comments, and likes allow companies and individuals to directly engage with targeted users in order to get a follow-back, i.e., to form a reciprocal tie (Wasserman and Faust 1994). A follow-back can further trigger cascades (Chae et al. 2016; Kozinets et al. 2010), which result in additional followers. Information dissemination and, therefore, the success of seeding decisions are dependent on the structure of the respective social networking platform, e.g., indegree distribution and density (for an overview see Jackson (2010)). It is possible to infer from the indegree distribution and, more

specifically, the prevalence of users with a large follower base not only the speed but also the spread of the information dissemination (Goldenberg et al. 2009). The same applies to the density or degree of connectedness among users of a social networking platform (Katona, Zubcsek, and Sarvary 2011; Stephen and Toubia 2010).

We consider each creator in user-generated content networks to be a separate brand, also referred to as a *human brand* (Thomson 2006). Such human brands vary heavily in their network status, ranging from unknown to extremely popular. In social network literature, the level of status (Katz 1953; Moreno 1934) is also referred to as rank, deference, or popularity (Wasserman and Faust 1994), with the most basic operationalization being indegree (e.g., Van den Bulte and Wuyts 2007), that is, the number of social ties a node has (Shaw 1954). Hence, some researchers use the terms status and indegree interchangeably (e.g., Iyengar, Van den Bulte, and Valente 2011). Status can be further measured by means of closeness and betweenness centrality (Freeman1978), but unlike indegree centrality, these are not clearly visible to all nodes in online social networks. The limited availability of information in user-generated content networks forces unknown creators to assess the status of seeding targets only by their indegrees. The closeness and betweenness centrality, on the other hand, can be observed exclusively by the social networking platform itself. This stands in contrast to various studies in which it is assumed that all network information is accessible (e.g., Katona, Zubcsek, and Sarvary 2011; Stephen and Toubia 2010). Nodes that are extremely popular and have a high indegree are referred to as hubs (Goldenberg et al. 2009), opinion leaders (e.g., Iyengar, Van den Bulte, and Valente 2011; Weimann 1991), connectors (Gladwell 2000), and influentials (e.g., Lionberger 1953; Merton 1968; Watts and Dodds 2007), with each term differing slightly in meaning.

2.2.2 Optimal Seeding

A large body of literature exists on high-status individuals, or influencers (Kirby and Marsden 2006; Rosen 2010), and it is widely agreed that they play a pivotal role due to their

ability to either accelerate or block the dissemination process. The two-step flow model described by Katz and Lazarsfeld (1955) characterizes opinion leaders as disseminators of information and, therefore, as the link between mass media and the public. Hence, influencers are regarded as powerful seeding targets. Apart from high-indegree seeding, there are two other policies based on sociometric measures: low-indegree seeding (or fringes) and high-betweenness seeding (or bridges) (Granovetter 1973). Social network literature suggests almost entirely that self-promotion efforts should seek to seed individuals with a high indegree (e.g., Easley and Kleinberg 2010; Hanaki et al. 2007; Hinz et al. 2011; Iyengar, Van den Bulte, and Valente 2011; Van den Bulte and Joshi 2007; Yoganarasimhan 2012). On the contrary, in a computer simulation, Watts and Dodds (2007) find that for most cases, high-indegree seeding does not have a major impact on cascades of influence, which agrees with studies showing that high-status individuals are not influential per se (Aral and Walker 2012; Trusov, Bodapati, and Bucklin 2010).

Since marketing managers decide upon a set of seeding targets for viral marketing campaigns (e.g., Bampo et al. 2008; Libai, Muller, and Peres 2005), the assessment regarding the value of such influencers is essential. In this context, Haenlein and Libai (2013) suggest to shift the focus toward the customer lifetime value of seeding targets. However, marketing managers along with creators of content who seek to build and increase their follower base usually make seeding decisions with very little information, the key factors being the status and indegree. Hinz et al. (2011) compare three different policies based on sociometric data. They show that high-indegree seeding outperforms two other policies focusing on fringes and bridges, partially because well-connected nodes capitalize on their greater reach and not entirely due to the fact that they exhibit a higher influence than others. In another study based on sociometric measures, Yoganarasimhan (2012) investigates the seed's follower base and its effect on macro-level dissemination using Youtube data. The conclusion that emerges from this study corresponds to Hinz et al. (2011) because high-indegree seeding resulted in far more clicks on videos in comparison to random seeding.

2.2.3 The Role of Status

While research usually places the focus on dissemination processes of products, we move the emphasis to the dissemination of unknown creators of content, i.e., human brands. Since unknown creators of content seek to build and increase their follower base, all their efforts are considered to be promotional actions. In this context, we argue that the effectiveness of these promotional actions and, therefore, the return opportunities depend jointly on the status of the sender and the receiver, i.e., the creator of content and the seeding target. In a recent paper, Hu and Van den Bulte (2014) show that status matters not only in terms of one's susceptibility to adopt and the adoption time, but also in the way that one's behavior affects others in terms of adoption. The authors conclude that for commercial kits used in genetic engineering, optimal seeding targets are individuals with high and middle status due to inverse-U patterns regarding adoption susceptibility and time. However, a gap remains in our knowledge since the focus of Hu and Van den Bulte (2014) is mainly on the status of the seeding targets. We, on the other hand, take into account the status of both the seeding targets and the creators of content who seek to build and increase their follower base. Furthermore, we consider the confronted trade-offs in terms of risk versus return.

Risk (and hence the trade-off) occurs as a result of different responsiveness levels associated with different status levels of seeding targets. Status differences and the corresponding inter- and intragroup behaviors are investigated in the social psychology literature. In this domain, the social identity theory (Tajfel and Turner 1979, 1986) reveals that high-status individuals are characterized by self-focused and self-serving behavior because they exhibit stronger in-group identification as well as favoritism (Bettencourt et al. 2001), and aim to preserve group boundaries and members (Ellemers et al. 1992; Terry, Carey, and Callan 2001; Van Knippenberg and Ellemers 1993). In contrast, low-status individuals want to disconnect from the low-status category (Ellemers et al. 1988; Snyder, Lassegard, and Ford 1986) and aim to be associated with the high-status one (Tajfel 1974, 1975; Tajfel and Turner 1979). In fact, individuals aim to form ties with high-status individuals due to the status transfer,

which has been studied in the context of researchers (Goode 1978; Latour 1987; Merton 1973) and can be interpreted as a form of endorsement (Stuart, Hoang, and Hybels 1999). Although low-status individuals benefit from such endorsement, high-status individuals exhibit a weaker attachment to low-status individuals than vice versa (Gould 2002) and risk devaluing their high status (Podolny 2001, 2005). This becomes apparent in the context of online dating, where individuals take into consideration the status (or "market worth") of others, as well as their own (Heino, Ellison, and Gibbs 2010; Taylor et al. 2011). Studies reveal that individuals with a high value or physical attractiveness level are commonly targeted (Buss and Barnes 1986; Feingold 1990; Lee et al. 2008; Walster et al. 1966), whereas individuals with high physical attractiveness favor strong in-group preference (Buston and Emlen 2003; Kowner 1995; Little et al. 2001; Todd et al. 2007). These phenomena that (1) everyone tends to reach out to high-status individuals, and that (2) they, in turn, respond preferably only to their own sort serve as a starting point for our paper.

We consider creators of music who seek to build and increase their follower base utilizing user-generated content networks. Each creator exerts a number of *promotional actions*: follows, messages, comments, and likes. Promotional actions are directed to *seeding targets* – users who are not part of the creator's follower base at the time of seeding. Hence, we do not take retention efforts into consideration, that is, promotional actions directed to the follower base. Prior research shows that high-status individuals are bombarded with a large number of incoming actions (Buss and Barnes 1986; Feingold 1990; Lee et al. 2008; Walster et al. 1966), whereas low-status individuals are not exposed to such competition for their attention. In addition, high-status individuals tend to respond only to other high-status individuals (Buston and Emlen 2003; Kowner 1995; Little et al. 2001; Todd et al. 2007). Therefore, we expect that *the higher the status difference*[1] *is between the creator of content and the seeding target, the lower the responsiveness of the seeding target will be.*

[1]We define status difference as the indegree of the creator minus the indegree of the seeding target.

2.3 Data

In our empirical analysis, we use data from SoundCloud, the world's leading user-generated content network in the domain of music. SoundCloud is home to 12 million creators of music and attracts 170 million monthly listeners (Pierce 2016). The social networking platform consists of two types of user profiles: creators of music and fans. Creators are individuals who have uploaded at least one song and use the network for self-promotion purposes. They engage with their fans and seek to expand their follower base. Fans, on the other hand, receive updates from their favorite creators and listen to their songs, as well as connect with their peers on the network. Hence, the social networking platform is composed of creators of music and their fan communities and offers users different possibilities to interact with each other. As with Twitter or Instagram, users of SoundCloud can follow each other without being followed back. Thus, the social networking platform at hand is directed. Users can further listen to songs uploaded on the artists' profiles; they can like these songs and leave comments about them. If a user reposts a song, all followers of this user receive a notification in their news feed. As a result, song reposts have a considerable impact on the popularity of songs and, for this reason, on the creators of music. Finally, users can contact each other by sending private messages. Consistent with Saboo, Kumar, and Ramani (2015), we study creators of music who upload songs on their profiles that can be listened to by other network users. In this context, unknown creators[2] who seek to build and increase their follower base, i.e., their brand communities, can reach out to users of the social networking platform by following them, sending them private messages, reposting their songs, commenting on their songs, or liking their songs.[3]

Our first data sample consists of 35,956 users (24,020 creators of music and 11,936 fans) and their egocentric networks. These users represent all sign-ups in the first quarter of 2009. This dataset contains all information about the formation of the users' egocentric networks

[2]We define unknown creators of music as all users who have uploaded at least one song and whose indegree has not crossed two orders of magnitude, i.e., a fan community of 100 followers.

[3]Regarding all creators sign-ups in the first quarter of 2009 with at least one follower, 95% did not cross the mark of 100 followers at the end of the year. This statistic drops to 83% in the consecutive year.

over a period of five years (January 2009–March 2014), as well as all data on all incoming and outgoing activities of each user including follows, messages, plays, comments and likes over the entire period. Moreover, we collected this information about their first degree alters, i.e., their followers. In late 2012, SoundCloud introduced the function and possibility for users to repost songs, which appears in followers' news feeds. As our first data sample does not include returns on song reposts (indirect returns), in our empirical analysis we incorporate a second data sample that consists of 35,000 users (4,978 creators of music and 30,022 fans) who signed up in the first week of March 2013, and we tracked them over a period of two years (until July 2015). The descriptive statistics of both data samples used in our empirical analysis are provided in Table 8.

— Insert Table 8 about here —

To sum up, our two longitudinal datasets consist of 70,956 users along with their alters (a total of 11,203,205 users). These datasets include complete information on (1) follows, (2) messages, (3) song plays, (4) song reposts, (5) song comments, and (6) song likes.

2.4 Theoretical Reasoning and Empirical Findings

2.4.1 Status Difference Matters

To study the dynamics of reciprocity with the aim of investigating the responsiveness of seeding targets, we zoom in to the dyadic level and analyze all 4,964,174 promotional actions sent to users of SoundCloud who were not part of the creators' follower base at the time of sending. These promotional actions consisting of follows, messages, song comments, and song likes were sent by 18,005 creators of music over 1,959 days.[4] We focus on the responsiveness of seeding targets, a binary measure denoted as 1 if the seeding target followed the creator back

[4]Each promotional action is equally weighted, and we do not consider the content of the message or song comment (and thus its effect on virality; Berger and Milkman 2012).

within a week, and 0 otherwise. We further take into consideration the difference in indegree, the most basic operationalization of network status (e.g., Van den Bulte and Wuyts 2007), between the sender and receiver of promotional actions. The period in which we consider reactions in the form of follow-backs was set to one week because this corresponds to the average login frequency of users of SoundCloud. Figure 1 exhibits the a priori response (follow-back) probabilities, given the order of magnitude of the seeding target to creator status ratio. Each bar captures 2.5% of the distribution whereby, for example, the bar with values between 2.5 and 2.9 includes each seeding target whose status is 2.5 to 2.9 times higher in order of magnitude compared to the status of the creator (the status of the seeding target is between $10^{2.5} \approx 300$ and $10^{2.9} \approx 800$ times greater than the status of the creator). We do not measure the direct effect of status difference on the a priori response probabilities nor do we claim that there are no other mediating factors. Hence, Figure 1 exhibits a model-free representation of the probabilities. From the monotonicity of the curve, we conclude that the higher the status difference is between the creator and the seeding target, the lower the a priori probability of a response. When extending the reaction period to two or three weeks, the a priori probabilities increase by 13% and 20% on average, respectively. Yet, the monotonicity of the curves remains.

— Insert Figure 1 about here —

To further investigate this phenomenon for low- and high-status creators, we segregate these two groups and define the former as creators with status less than two orders of magnitude, i.e., 100 followers, and the latter as creators with status more than three orders of magnitude, i.e., 1,000 followers. Figure 2 exhibits the a priori response probabilities as a result of promotional actions from low- and high-status creators. Since there is a clear right shift of a priori response probabilities from low- to high-status creators, status matters when directing promotional actions to seeding targets. Put differently, high-status creators have a higher a priori response probability in comparison to low-status creators for any status of

a seeding target. The monotonicity of both curves is identical to Figure 1. This is also the case when allowing for longer reaction periods.

— Insert Figure 2 about here —

Both Figures 1 and 2 provide evidence that the higher the status difference between the creator and the seeding target, the lower is responsiveness. Whereas Figure 2 distinguishes between high- and low-status creators, Figure 1 generally exhibits the a priori response probability, given the status difference between the creator and the seeding target.

In a context where the probability to follow back varies, the optimal seed is not necessarily the influencer. However, our analyses do not reveal which seeding targets a creator *should* choose since different levels of responsiveness also correspond to different levels of return (an influencer can create a higher exposure relative to an ordinary individual). In the next subsection, we describe the individual portfolio of seeding targets, as well as associated returns.

2.4.2 Risk Versus Return Trade-Offs

Topical research assumes that seeding a target in online social networks is just a matter of choice and does not involve any risk, either in the form of time constraints, search costs to find seeding targets, anti-spam policies, or differences in responsiveness (e.g., Hinz et al. 2011; Yoganarasimhan 2012). Put differently, in current research, a chosen seeding target subsequently promotes the advertised product or service with certainty. We relax this assumption and consider the risk of getting a return when seeding a specific target. The previous subsection reveals that there is a difference in responsiveness when seeding an unconnected node compared to a highly connected one. Therefore, we again zoom into the relational mechanisms on a dyadic level (Rivera, Soderstrom, and Uzzi 2010) and associate different levels of responsiveness with different levels of *returns*. The return scheme, which results from a seeding target who responds to a promotional action, is composed of two sources: (1) a

direct return – the follow-back from the seeding target, which depends on the responsiveness; and (2) an *indirect return* – one that results from the seeding target's follower base, which depends on whether the seeding target further reposts songs from the creator or not. Song reposts trigger additional follows as they disseminate into the seeding target's egocentric network (e.g., Everett and Borgatti 2005; Wasserman and Faust 1994). Therefore, creators make seeding decisions while accounting for the influence of status difference on the a priori probability of response, along with the associated potential returns.

Creators vary in their expenditure of time for self-promotion and seeding efforts. Along these lines, we consider the number of promotional actions sent by a creator within a time period as *budget*. On average, low-status creators (those with less than 100 followers) have a weekly budget of 2.3 promotional actions. Due to the large size of user-generated content networks such as SoundCloud, creators cannot reach out to all users, neither at once nor over the entire lifetime. Moreover, due to time constraints, search costs to find seeding targets, and anti-spam policies, the individual budget of promotional actions is not unlimited, as is often (mostly implicitly) assumed. In line with Facebook and Twitter, SoundCloud's Community Guidelines do not allow users to "post identical or almost identical comments or messages in large volumes; repeatedly follow large volumes of accounts in a short period of time; repeatedly contribute your tracks to large volumes of groups in a short period of time; repeatedly unfollow and refollow the same accounts, in order to draw attention to your own profile; repeatedly repost or like tracks that you have reposted or liked in the past" (SoundCloud 2016). Therefore, creators are forced to decide upon a set of seeding targets and, thus, have to create a *portfolio* of individuals, namely SoundCloud users they want to target with promotional actions to receive follow-backs.

According to common social network literature in marketing (e.g., Goldenberg et al. 2009; Libai, Muller, and Peres 2013), the optimal portfolio choice of seeding targets is a corner solution: to direct all promotional actions to individuals with a high indegree, because in the case of response, the increase in the creator's follower base is higher compared to a response by a person with a low indegree. However, our analysis reveals that the probability of re-

sponse by a target with a high indegree depends on the status of the creator. For unknown creators who seek to build and increase their follower base, targeting influencers is associated with high risk (due to their low responsiveness), while targeting ordinary individuals is associated with lower risk (due to their higher responsiveness). The optimal composition of the portfolio of seeding targets with different indegrees depends on the creator's aversion to risk. Consequently, we define the creator's seeding problem for the purpose of self-promotion in user-generated content networks as *a risk versus return trade-off, depending on the individual aversion to risk*.

We define an optimal portfolio of seeding targets by drawing on the von Neumann–Morgenstern utility framework (see Archak, Ghose, and Ipeirotis 2011; Eliashberg 1980; Hauser and Urban 1977; Hauser 1978; Hauser and Urban 1979), which implies that the creator selects a portfolio that maximizes expected utility. We assume that the creator can choose between two investment tracks, namely influencers (individuals with a high status and indegree) and ordinary individuals. We further assume that the creator gains utility from the returns on a portfolio, where the utility is an increasing function of returns. Endowed with a budget B, the creator is confronted with a portfolio choice under uncertainty and has to invest a fraction X of budget B in low-status individuals and a fraction Y of budget B in high-status individuals, where $X + Y = B$. Investments in targets with a low status (ordinary individuals) are successful with probability p_L and, subsequently, yield a low return L. Investments in targets with a high status (influencers) are successful with probability p_H and yield a high return H. Following the risk versus return trade-off, the return on ordinary individuals is lower than on influencers, i.e., $L < H$; however, the probability of success when investing in low-status individuals is higher, i.e., $p_L > p_H$.[5]

We consider a static model with one time period and incorporate in the magnitude of H

[5]Although the response (follow-back) probability of high-status individuals is very low and the probability that they repost a song is marginal, it is possible that a series of follows from high-status individuals is triggered if one of them responds. In this case, the indirect return of the high-status individual that received the promotional action, i.e., H, would just be rescaled. Moreover, as the global clustering coefficient of SoundCloud is very low, low-status individuals are usually not interrelated. Based on the above, we thus assume independency of returns.

all expected returns resulting from cascades initiated by a response from an influencer, as well as the status increase of the creator due to the additional number of followers. Formally, a creator directs a promotional action to seeding target R, where Z_R is the return that the creator gains as a result. Thus an investment in a low-status individual yields

$$Z_R = \begin{cases} L & \text{with probability } \ p_L \ , \\ 0 & \text{with probability } \ (1 - p_L) \ , \end{cases} \tag{1}$$

whereas an investment in a high-status individual yields

$$Z_R = \begin{cases} H & \text{with probability } \ p_H \ , \\ 0 & \text{with probability } \ (1 - p_H) \ . \end{cases} \tag{2}$$

Based on the above return scheme, the creator invests his or her budget in low- and high-status individuals, i.e., ordinary individuals and influencers. The expected utility of this portfolio choice is given by

$$EU = \sum_{x=0}^{X} \sum_{y=0}^{Y} \binom{X}{x} \binom{Y}{y} p_L^x \, p_H^y \, (1 - p_L)^{X-x} \, (1 - p_H)^{Y-y} \ U(xL + yH) \ . \tag{3}$$

While deciding upon a set of seeding targets, the creator solves the following optimization problem:

$$\underset{X,Y}{\text{Max}} \ EU(Z(X, Y)) \quad \text{s.t.} \quad X + Y = B \ , \tag{4}$$

where $Z(X, Y)$ is a random variable that expresses the return, given an investment X and Y in low- and high-status individuals, respectively, such that $Z(X, Y) = xL + yH$ with probability $\binom{X}{x}\binom{Y}{y} p_L^x \, p_H^y \, (1 - p_L)^{X-x} \, (1 - p_H)^{Y-y}$.

If we assume that the creator is risk neutral, then uncertainty does not influence the portfolio choice. Therefore, the creator solely chooses the investment track that yields the highest expected return.[6] In case the expected return on high-status individuals is higher

[6]The expected return on each seeding target is the expected value of the return given the distribution of returns.

than on low-status individuals, the creator invests the whole budget in high-status individuals. Otherwise, the whole budget is invested in low-status individuals. Consequently, if the creator invests the budget in both investment tracks, then he or she cannot be risk neutral.

Proposition 1. *If a creator of content is risk neutral, then the whole budget is invested in high-status individuals, i.e.,* $Y^* = B$*, in case the expected return on high-status individuals is higher than on low-status individuals, i.e.,* $p_H H > p_L L$*. Otherwise, the whole budget is invested in low-status individuals, i.e.,* $X^* = B$ *(see Appendix A.1).*

Corollary 1. *If a creator of content invests in both low- and high-status individuals, i.e.* $X^* \neq 0$ *and* $Y^* \neq 0$*, then they cannot be risk neutral (see Appendix A.1).*

From Corollary 1, it follows that a creator of music on SoundCloud cannot be risk neutral if his or her individual portfolio choice of seeding targets includes both low- and high-status individuals. If we assume that the status difference between the creator and the seeding target does not influence the response (follow-back) probability of seeding targets, that is, if the creator believes that the response probability is the same across all status levels of seeding targets, then the optimal choice are high-status individuals and influencers, respectively.

Proposition 2. *If the response probability of high-status individuals equals the response probability of low-status individuals, i.e.,* $p = p_L = p_H$*, then the whole budget is invested in high-status individuals, i.e.,* $Y^* = B$ *(see Appendix A.1).*

From Proposition 2, it follows that if a creator of music on SoundCloud does not take different levels of responsiveness into account, then their individual portfolio choice of seeding targets includes only high-status individuals, as these yield the highest return. However, if we assume that high-status individuals are extremely unresponsive, namely, that their a priori response probability is close to zero, then the optimal choice of seeding targets would be low-status individuals. In this context, sending promotional actions to high-status individuals and influencers, respectively is a waste of resources.

Proposition 3. *If the response probability of high-status individuals is extremely low, then the whole budget is invested in low-status individuals, independent of the aversion to risk. More precisely, for any utility function U there exists a large number K such that for any $p_H < \frac{1}{K}$ the creator of content directs all promotional actions to low-status individuals, i.e. $X^* = B$ (see Appendix A.1).*

From Proposition 3, it follows that if a creator of music on SoundCloud takes into account different levels of responsiveness and assuming that high-status individuals are extremely unresponsive, then the creator's portfolio choice of seeding targets includes only ordinary (low-status) individuals. Propositions 2 and 3 represent two extreme choices of seeding targets and, hence, set the limits for all possible portfolio choices. In Proposition 2, the difference between the responsiveness of low- and high-status individuals is marginal. Therefore, the creator's tendency is to invest in high-status individuals and influencers, respectively. In proposition 3, however, the difference in responsiveness between the status levels is extremely high and, as a result, the creator tends to invest in ordinary (low-status) individuals. All portfolios with investments in both status levels exhibit a risk versus return trade-off. Moreover, if the creator directs promotional actions to low- and high-status individuals, then from Proposition 2, it follows that the creator cannot be risk neutral.

Along these lines, we assume that the higher the status of the creator, the higher the a priori response probabilities. Furthermore, with increasing status of the creator, the relative improvement of the a priori response probability regarding promotional actions directed to high-status individuals is higher compared to low-status individuals (see Tables 2 and 3). Thus, the higher the status of the creator, the more promotional actions are directed to seeding targets with higher status.

Proposition 4. *If the status of a creator of content, i.e. S, increases, then the more is invested in high-status individuals, i.e., $\frac{dX^*}{dS} < 0$ and $\frac{dY^*}{dS} > 0$ (see Appendix A.1).*

From Proposition 4, it follows that if a creator of music on SoundCloud gains followers and, hence, his or her status increases, then the creator reallocates promotional actions from

low- to high-status individuals. The allocation of promotional actions is influenced not only by the creator's status, but also by the number of promotional actions sent within a time period, i.e., the budget size. If a creator is endowed with a low budget and few promotional actions, respectively, then with increasing budget, he or she will not necessarily reallocate promotional actions to seeding targets with higher status, as creators want to reach at least a certain baseline growth of follow-backs. If this baseline growth is achieved, then the creator is able to take higher risks and, consequently, can direct promotional actions to seeding targets with higher status, i.e., "first bread then butter".

Proposition 5. *Assuming that the response probability of high-status individuals is extremely low, i.e.,* $p_H \ll \frac{1}{B}$*, and the resulting return (in case the high-status individual responds) is extremely high, i.e.,* $BL \ll H$*, then the larger the budget of a creator of content, the more is invested in high-status individuals, i.e.,* $\frac{dY^*}{dB} > 0$ *(see Appendix A.1).*

From Proposition 5, it follows that if a creator of music on SoundCloud increases the budget size, then the creator starts to reallocate promotional actions to seeding targets with higher status. Propositions 1 to 5 lay the foundation to examine if and how creators of music on SoundCloud are solving risk versus return trade-offs. For this purpose, we cluster SoundCloud users by their status, i.e., indegree, and separate them according to order of magnitude. This classification is appropriate as it follows a logarithmic scale and, thereby, lives up to the dispersion of indegrees in well-established online social networks. As a result, we consider four groups of users: *type 1* users have fewer than or equal to 100 followers; *type 2* users have more than 100 but fewer than or equal to 1,000 followers; *type 3* users have more than 1,000 but fewer than or equal to 10,000 followers; and *type 4* users have more than 10,000 followers. In Figure 3, we contrast the choices of seeding targets of creator types 1, 2, 3, and 4. Along these lines, we define unknown creators as type 1 – all users who uploaded at least one song and whose indegree has not crossed two orders of magnitude, i.e., a fan community of 100 followers. Figure 3 illustrates that all creators of a specific type do not direct all promotional actions to a specific type of seeding target; a corner solution does not appear. In fact, they spread their budget of promotional actions over several orders of

magnitude in terms of status of seeding targets. Moreover, in accordance with the insights provided by Propositions 2 and 3, creators choose a portfolio of seeding targets because they do not direct promotional actions just to low- or high-status individuals. These different portfolios, in the form of four bell-shaped distributions over several orders of magnitude in terms of status, are illustrated in Figure 3. Based on the insights from Proposition 1 and Corollary 1, creators are thus not risk neutral. They further consider the influence of status difference on the a priori probability of response; otherwise, they would, according to Proposition 2, simply direct promotional actions to individuals with high status, i.e., influencers with a high indegree. Therefore, we find supplementary evidence in support of our expectation regarding the effect of status difference on the a priori response probabilities. Furthermore, Figure 3 shows that the four bell-shaped distributions shift more to the right as the status of creators increases, providing evidence for Proposition 4, which states that the higher the creator's status, the higher the status of their seeding targets.

— Insert Figure 3 about here —

Could this allocation result from a random selection of targets? We analyze and compare our findings with the seeding policy in which the creator of music on SoundCloud randomly directs promotional actions to seeding targets. This random seeding policy is reflected by the indegree distribution of SoundCloud and serves as a benchmark. For this reason, we assess the status of all available users of SoundCloud, which amount to 394,262 creators of music and fans. As we expect a right shift of the indegree distribution over time due to SoundCloud's growth, we calculate it at the end of the observation period – in the first week of 2014. Since the different portfolios of seeding targets in the form of the four distribution curves are different from the indegree distribution, we conclude that creators do not randomly direct promotional actions to seeding targets. Moreover, Figure 3 reveals an increased tendency to direct promotional actions to seeding targets with higher status because the distribution curves lie on the right side of the indegree distribution, i.e., the random seeding policy. More specifically, in addition to direct returns, creators aim for indirect returns to get follow-

backs from the followers of the seeding targets. To sum up, Figure 3 provides evidence for our expectation that unknown creators who seek to build and increase their follower base solve a risk versus return trade-off.

With Proposition 5 in mind, we study the effect of the number of promotional actions sent within a time period (considering the creator's budget size) on the choice of status levels of seeding targets. More precisely, there is heterogeneity among creators of the same type regarding their budget size: Some creators of music spend a larger fraction of their time on SoundCloud and, as a result, exert more promotional actions within the time period. Recall that our analyses focus on unknown creators who seek to build and increase their follower base. Figure 4 exhibits, given their weekly budget, a type-1 creator's relative investment in seeding targets of types 1, 2, 3, and 4. The weekly budget size is split into one to ten actions and more. For example, a type-1 creator with a budget of one promotional action per week allots 38% to type 1, 30% to type 2, 22% to type 3, and 10% to type 4. On average, a type-1 creator has a weekly budget of 2.3 promotional actions. Figure 4 suggests that these creators become less risk-averse with increasing budget. The probability to invest in type 1 seeding targets decreases with increasing budget, from 38% to 28%, which is found to be highly statistically significant, i.e., the Pearson's chi-squared test with Yates' continuity correction gives a p-value of < 0.0001.

— Insert Figure 4 about here —

So far, we have found evidence for both our expectations regarding the effect of status difference on the a priori response (follow-back) probabilities, as well as the risk versus return trade-offs. Due to the monotonicity of the curve in Figure 1, we conclude that the higher the status difference between the creator and seeding target, the lower the responsiveness. Furthermore, Figure 2 shows that for any status of a seeding target, high-status creators have a higher a priori probability of getting a response or direct return in comparison to low-status creators. In the absence of indirect returns – namely, follow-backs from the followers of the

seeding targets – unknown (low-status) creators who seek to build and increase their follower base would only consider response probabilities. In this case, they would not include high-status individuals and influencers, respectively in their portfolio of seeding targets. However, Figure 3 shows that creators spread their budgets of promotional actions over several orders of magnitude in terms of status, meaning they choose a portfolio and do not send only to a certain status. More specifically, the higher their own status, the more promotional actions are sent to seeding targets with higher status. When comparing the portfolios with the indegree distribution of SoundCloud, it becomes apparent that creators of music also consider indirect returns in addition to direct returns, as they send promotional actions to high-status seeding targets too. These findings support our expectation regarding the risk versus return trade-offs. Thus, we conclude that unknown creators who seek to build and increase their follower base solve a risk versus return trade-off when choosing their portfolio. Furthermore, with increasing budget, they have a lower tendency to allocate promotional actions to low-status seeding targets, as Figure 4 shows.

Summing up, our empirical analysis of creators' revealed preferences on SoundCloud indicates that unknown creators of music are choosing a portfolio of seeding targets while solving a risk versus return trade-off. Hence, they do not reach out exclusively to those individuals with the highest status in the network in terms of centrality, as suggested by current social network literature in marketing. In the following section, we investigate expected total returns on seeding targets retrieved from the data and discuss these implications on the individual aversion to risk.

2.4.3 Aversion to Risk

The creator's choice of seeding a target with a specific status, i.e., the creator's allocation of budget to high- and low-status individuals, depends on the creator's aversion to risk. To develop deeper insights into creators' seeding policies and how these are influenced by their aversion to risk, we compute the expected total return on each seeding target (the expected

value of the returns given their distribution), which accounts for (1) the expected direct returns and (2) the expected (indirect) returns associated with a song repost.

We first measure the a priori response (follow-back) and song repost probabilities before assessing the indirect return – the number of follows from the seeding target's follower base in the case of a song repost. Note that the direct return is a binary measure (equals 1 if the seeding target responds to the promotional action from the creator, or 0 otherwise). In our analysis, we classify users (creators and seeding targets) again by their indegree and separate them according to the order of magnitude, resulting in four groups of creators and seeding targets, respectively (same definition of types as before). Table 2 shows the a priori response probabilities for all combinations of creator and seeding target types, and whether the seeding target follows the creator back within a week after receiving a promotional action. Recall that our analyses focus on low-status (unknown) creators, type 1, who seek to build and increase their follower base. In the case of a type-1 creator, the a priori response probabilities range between 7.41% (type-1 seeding target) and 0.03% (type 4-seeding target). For a type-4 creator (with a fan community of more than 100,000 followers), these a priori response probabilities range as high as 15.03% (type-1 seeding target) and 0.75% (type-4 seeding target).

— Insert Table 2 about here —

To compute the expected (indirect) returns associated with a song repost, for all combinations of creator and seeding target types, we further analyze the a priori song repost probabilities – primarily, whether the seeding target reposts a song of the creator within a week after receiving a promotional action from the creator. Compared to the a priori response probabilities, as shown in Table 2, the a priori song repost probabilities are significantly lower. Table 3 shows that in the case of a type-1 creator, the a priori song repost probabilities range between 0.109% (type-1 seeding target) and 0.001% (type-4 seeding target). For a type-4 creator, these a priori song repost probabilities increase to 0.372% (type-1

seeding target) and 0.023% (type-4 seeding target).

— Insert Table 3 about here —

After measuring both the a priori response and song repost probabilities, we further compute the expected (indirect) return associated with a song repost to finally calculate the expected total return of a creator's promotional action and self-promotion effort. In our analysis, we consider 1,501,051 song reposts from 9,402 creators on SoundCloud over 424 days. For all combinations of creator and seeding target types, Table 4 exhibits the expected follows and song plays realized within a week after a song repost. The results reveal that the expected indirect return for an unknown creator within a week, given a song repost from a type-4 user, amounts to 7.6 follows and 281.6 plays on average. The reaction period was set to one week, as this time span corresponds to the average login frequency of users of SoundCloud and accounts for the short lifespan of published information in the user's news feed in which the song repost is shown. Doubling the reaction period to two weeks results in only a marginal increase: on average in a total of 8.1 new follows and 304.4 song plays, respectively. Both the conversion of plays to follows and the general level of the figures are low. Qualitatively, our results show that the a priori song repost probabilities, as shown in Table 3, as well as the expected indirect returns given a song repost, as shown in Table 4, are extremely low for an unknown creator who sends promotional actions to high-status individuals.

— Insert Table 4 about here —

Taking into account the a priori response (follow-back) and song repost probabilities, as well as the expected indirect returns associated with a song repost, we are able to analyze and compute the expected total return on a promotional action directed to a seeding target. The expected total return on a seeding target consists of the expected direct return, which is

determined by the a priori response probabilities, and the expected indirect return, which is determined by the a priori song repost probabilities, as well as the expected indirect returns given a song repost (see Appendix A.2 for a detailed description).

Analyzing 4,964,174 promotional actions that include follows, messages, song comments, and song likes of 18,005 creators of music over 1,959 days, Table 5 shows the expected total returns for all combinations of creator and seeding target types. Surprisingly, the expected total return on high-status individuals is *lower* than the expected total return on low-status individuals. In particular, an unknown creator who directs a promotional action to a type-1 seeding target gets on average .0741 follow-backs with a standard deviation of .2619. In contrast, directing a promotional action to a type-4 seeding target yields almost no return with certainty because the expected total return is as low as .0003 follow-backs, and the corresponding standard deviation amounts to .0446. In other words, the lower the target's status, the higher the expected total return, which implies that a creator of music on SoundCloud should not direct promotional actions to high-status individuals and influencers, respectively.

— Insert Table 5 about here —

Our analysis shows that for unknown creators who seek to build and increase their follower base, influencers are associated not only with surprisingly low return but also with high risks (due to their low responsiveness), while ordinary individuals are associated with relatively lower return, but also lower risk (due to their higher responsiveness). However, we also find that creators spread their budgets of promotional actions over several orders of magnitude in terms of status, targeting even high-status individuals who feature a lower expected total return than ordinary (low-status) individuals. This indicates that utility-maximizing creators of music on SoundCloud are not completely risk-averse.

Proposition 6. *If the return on high-status individuals is much higher than on low-status individuals such that* $H > BL$ *but the expected return on high-status individuals is lower*

than on low-status individuals, i.e., $p_H H < p_L L$, and there is an optimal solution such that a creator invests in high-status individuals, i.e., $Y^* \neq 0$, then the creator cannot be risk averse (and the individual utility function U is not concave) (see Appendix A.1).

If it is common knowledge among creators of music that the value of high-status individuals on SoundCloud in terms of expected total returns is questionable, it follows from Proposition 6 that high-indegree seeding indicates risk-seeking behavior: Risk-averse, utility-maximizing creators of music on SoundCloud would never direct promotional actions to high-status individuals based on this return scheme.

To conclude, our empirical analyses reveal that high-status individuals on SoundCloud are associated with surprisingly low returns, apart from the high risks due to their low responsiveness. Given that this return scheme is common knowledge, creators reveal a behavior in the context of SoundCloud that is associated with risk seeking because they spread their budgets of promotional actions over several orders of magnitude in terms of status. In the subsequent section, we investigate different seeding policies. In particular, we study the consequences of the observed risk-seeking policy of creators of music on SoundCloud.

2.5 Comparing the Effectiveness of Seeding Policies

The previous sections suggest that if an unknown (low-status) creator of content seeks to build and increase his or her follower base, then he or she should ignore predominant seeding policies and slowly "climb" the ladder of status levels of seeding targets rather than attempt to "jump" towards those with the highest status. More specifically, by directing promotional actions to seeding targets with the lowest status, an unknown creator can generate the highest possible continuous growth of his or her follower base, in addition to natural baseline follows. The accumulation of follows (baseline plus return on seeding targets) increases the creator's status, which goes hand in hand with higher a priori probabilities and, thus, expected total returns on each seeding target.

By means of a randomized dissemination process using the example of a creator who has just signed up on SoundCloud, we contrast three seeding policies. In the first, we simulate unknown creators, initially with zero followers, who invest their budgets of promotional actions in line with the *status quo*. More precisely, for each status we retrieve the average portfolio choice from the data and simulate unknown creators who exert promotional actions accordingly, where the return in the form of additional number of followers is drawn in correspondence with the probabilities observed in the data. In the second policy, we simulate unknown creators who invest in line with common social network literature in marketing by exclusively seeding targets with *the highest status*. In the third policy, the simulation takes into account unknown creators following the seeding policy suggested in this paper who invest only in seeding targets with *the lowest status*.

2.5.1 Method

For each of the three seeding policies, we compute the median growth of a creator's follower base over a 24-month time period. We focus on creators of music who have just signed up on SoundCloud and, thus, have zero followers in the beginning. During each of the 24 months, the simulated creator invests 40 promotional actions, which corresponds to a weekly budget of 10 and an overall budget of 960 promotional actions, respectively. The simulated creator invests according to one of the three policies over the whole time period, for a total of 1,000 iterations. The mechanism is the same for any chosen policy: In each month, the creator's increase in follower base is determined by the status-dependent probability of a non-zero return on a seeding target and, further, on the status-dependent probability of either a direct or indirect return.

More specifically, if the return in a given month for a given seeding target is non-zero, then there are two different scenarios. On one hand, the investment in this seeding target can yield both a direct and indirect return – a follow-back from the seeding target and follows from subsequent song reposts. We consider the average number of song reposts from

a seeding target over a year to account for the long-term indirect return on a follow-back. On the other hand, the investment in this seeding target can yield only an indirect return, i.e., follows from a (single) song repost. The monthly accumulation of baseline follows as well as returns on seeding targets increases the creator's status, which go hand in hand with higher a priori probabilities and, thus, expected total returns on each seeding target. Both the natural baseline follows and the a priori probabilities including the expected returns on each seeding target are updated in multiples of 25 followers with regard to the growing follower base, after a reaching a community size of ≥ 25 followers, ≥ 50 followers, and so forth. To sum up, the randomized dissemination process first takes into account the status-dependent probability of a non-zero return on a seeding target and, subsequently, considers whether the non-zero return is realized directly or indirectly (see Appendix A.3 for a detailed description of the simulation study).

2.5.2 Results

Figure 5 exhibits the median growth of a follower base of a creator endowed with zero followers when signing up and reveals that different seeding policies vary greatly in their outcomes. We find that investments in line with the current social network literature in marketing – high-indegree seeding – amount to a median of 22 followers over 24 months.[7] Such investments are not worthwhile, since an unknown creator who directs promotional actions to high-status individuals faces extremely low a priori song repost probabilities, and also very low expected indirect returns given a song repost. Even more striking, the expected total return on individuals with a high indegree is lower than the expected total return on individuals with a low indegree, as exhibited in Table 5. Hence, investments in high-status individuals result in the accumulation of natural baseline follows.

Furthermore, we find that investments according to the actual portfolios observed in the data result in a median of 72 followers.[8] This seeding policy, which reflects the (status-

[7] After 24 months, the middle 50% have between 19 and 26 followers.
[8] After 24 months, the middle 50% have between 46 and 81 followers.

dependent) status quo seeding policy of creators of music on SoundCloud, outperforms constant investments in high-status individuals by more than threefold. Therefore, the choice of seeding targets by creators of music on SoundCloud is more effective than the one suggested by current social network literature in marketing.

Finally, we find that investing only in seeding targets with the lowest status results in a median of 127 followers within 24 months.[9] As shown in Table 5, the expected total return on high-status individuals is lower than the expected total return on low-status individuals; therefore, low-indegree seeding manages to accumulate followers more effectively. Specifically, by directing promotional actions to seeding targets with the lowest status, an unknown creator generates the highest possible continuous growth of his or her follower base, in addition to the natural baseline follows. This, in turn, increases the creator's status, which goes hand in hand with higher a priori probabilities and expected total returns on each seeding target, hence the slightly convex curve. Our results show that investments only in seeding targets with lowest status clearly dominate the other two policies. The seeding policy suggested in this paper not only outperforms the chosen seeding policy of creators of music on SoundCloud, it is superior to the one suggested by the common social network literature, by close to sixfold after not more than two years.

— Insert Figure 5 about here —

In summary, patience pays off very well: Unknown creators who seek to build and increase their follower base should ignore predominant seeding policies and slowly "climb" across status levels of seeding targets rather than attempting to "jump" towards those with the highest status. Hence, unknown creators should invest in seeding targets with the lowest status, instead of chasing indirect returns by directing promotional actions to high-status individuals.

[9] After 24 months, the middle 50% have between 116 and 139 followers.

2.6 Discussion

Topical research assumes that seeding a target on online social networking platforms is just a matter of choice and does not involve risk in the form of time constraints, search costs to find seeding targets, anti-spam policies, or differences in responsiveness (e.g., Goldenberg et al. 2009; Hinz et al. 2011; Libai, Muller, and Peres 2013; Yoganarasimhan 2012). In the context of user-generated content networks, which constitute a unique type of social networking platform (e.g., Goldenberg, Oestreicher-Singer, and Reichman 2012; Mayzlin and Yoganarasimhan 2012; Netzer et al. 2012; Trusov, Bodapati, and Bucklin 2010), we relax this assumption and consider the risk of getting a return when seeding a specific target. Our research extends existing seeding literature by taking into consideration that on user-generated content networks (1) the difference of network status between the creator and the seeding target matters, (2) the creator's budget of promotional actions is constrained, and (3) since different levels of returns are associated with different levels of responsiveness, creators of content solve a risk versus return trade-off when choosing their portfolios of seeding targets.

Our analyses reveal that creators of music on SoundCloud, the world's leading user-generated content network in the domain of music, do not direct promotional actions only to influencers, i.e., users with a high indegree. In fact, they spread their budgets of promotional actions and create a portfolio of seeding targets over several orders of magnitude in terms of their network status. Moreover, the higher the creator's status, the greater are the number of promotional actions sent to seeding targets with higher status. When comparing the portfolios with the indegree distribution of SoundCloud, it becomes apparent that creators of music also consider indirect returns in addition to direct returns, as they send promotional actions to high-status seeding targets too. Our analyses show that unknown creators who seek to build and increase their follower base solve a risk versus return trade-off when deciding upon a set of seeding targets: A fraction is "invested" in SoundCloud users with a high status difference compared to the creator of music under consideration, which is associated with

high risk (due to low responsiveness) but potentially high return. The remaining fraction of the portfolio is "invested" in SoundCloud users with a lower status difference, which is associated with lower risk (due to higher responsiveness) but relatively low return.

We analyze data from a user-generated content network, which might limit our results to this type of platform. But let us revisit the assumption that the probability of an individual to endorse a person, small group, or firm that requested such an action is constant. This assumption is a very strong one and calls for new examination. It might be that the same monotonicity we discovered in this paper exists in other cases or platforms. In fact, any small- or medium-sized business faces a similar risk versus return trade-off when reaching out to seeding targets. Trying to activate high-status individuals might be the appropriate policy for large corporations with the financial power to compensate influencers who promote their products and services. However, this seeding policy may not apply for small- and medium-sized businesses. With increasing size and thus status of such businesses, the probability of response as well as the associated expected total return when reaching out to influencers improves continuously. The insights for effective seeding policies may be of high importance because, according to the recent analyses of federal statistical offices, most businesses are small- and medium-sized (e.g., 99.7% in the U.S. and 99.3% in Germany), and they engage a large proportion of labor (e.g., 48.4% in the U.S. and 60% in Germany).

In the context of SoundCloud, our empirical analyses reveal that the expected total return on individuals with a high indegree is lower than the expected total return on individuals with a low indegree. Put differently, high-status individuals are associated not only with low responsiveness but also with surprisingly low return. As a result, an unknown creator of content who seeks to build and increase his or her follower base should ignore predominant seeding policies and slowly "climb" in the ladder of status levels of seeding targets rather than attempt to "jump" towards those with the highest status. By directing promotional actions to seeding targets with lowest status, an unknown creator can generate the highest possible continuous growth of the follower base. Future research should investigate this phenomenon in other networks, e.g., in the context of telecommunication, and disentangle the underlying

psychological processes, especially the individual aversion to risk, which is beyond the scope of this paper. We hope this paper encourages work in these and other related directions.

3 Allocation of Marketing Budget When Success Is a Rare Event

3.1 Introduction

For eight consecutive years after its inception in 2002, *American Idol* was the most popular television show in the US, attracting over 30 million viewers during its most successful episode (Koblin 2015). The American Idol format revolves around the early discovery of talent and the promise of substantial economic value. One of the reasons for the show's popularity is that a panel of judges as well as the audience make selection decisions based on the contestants' stage performances, and eventually determine who will receive the record deal with a major label. Some marketing practices are similar, in principle, to the American Idol format: Managers select product and service ideas based on (limited) preliminary information and then invest in some development and later allocate marketing resources, while considering the profit potential of each new product or service (e.g., Fischer et al. 2011). One of the puzzles managers have in each stage of the process is the focal question: *Will it ever fly?* (Golder and Tellis 1997). This fundamental question has not changed for decades, nor has the fact that success is a *rare event*. In effect, most products or services will never fly: For example, 97% of new consumer packaged goods fail to breach sales thresholds to be considered a big success (Schneider and Hall 2011).

Marketing researchers have developed a wide range of prediction models in the context of innovation growth (e.g., Garber et al. 2004; Markovitch and Golder 2008; Peres, Muller, and Mahajan 2010; Foster, Golder, and Tellis 2004; Van Everdingen, Fok, and Stremersch 2009), many of which are used by managers to make better-informed investment decisions. In companies with broad and deep product pipelines, managers face an even more complex

allocation problem of how to invest and allocate their marketing budgets among multiple products. If they decide to further invest in all products, each future best-selling product received initial support. However, from an economic perspective, this policy is associated with considerable *wasted investments* due to the extensive resources allocated to the many products that ultimately turn out to be failures. At the other extreme, managers could decide to invest in a single product. In this case, managers face high *lost returns*, since it is almost certain that such a focused investment of marketing resources will not capture all the company's future best-selling products. This trade-off leads to the following question: *In how many and, more precisely, in which products should a manager invest in an early stage of the life cycle?*

If managers have a perfect prediction model at hand, then the selection size (the number of products selected for further investment) would simply equal the prevalence of success, with no wasted investments or lost returns. However, it is impossible to perfectly predict product success. Various types of discrete choice models (e.g., logit or probit models) are used to calculate the a posteriori probability of success for each product, given the available information, but they are far from being perfect predictors, and a real success, even these days, is still a rare event. The common procedure with probability models is to apply the maximum a posteriori probability principle (MAP; for an overview see Gallager 2013). However, if managers follow this naïve approach, then they should, in principle, reject all products because the a posteriori probabilities likely indicate non-success for each product. In effect, popular statistical procedures can generate heavy underestimations of the probability of rare events (King and Zeng 2001b).

We propose a novel managerial framework for maximizing profits when deciding on how many as well as which products to select. More precisely, our proposed framework recommends a selection along decreasing a posteriori probabilities by solving the economic trade-off between wasted investments and lost returns. Please note that our proposed framework does not depend on a specific prediction model. Instead, it uses the model that is shown to have the best predictive power in terms of probability of detection and probability of false discov-

ery. These two measures are inspired by *signal detection theory* and allow us to distinguish between information-bearing patterns and noise (see for details Schonhoff and Giordano 2006; Wickens 2001; Peterson, Birdsall, and Fox 1954). The prediction model's probability of detection and probability of false discovery serve as a starting point, which we link to wasted investments and lost returns, and subsequently include in the profit maximization. Given the level of return on investment (ROI) of the relevant product category, the profit maximization determines the optimal selection size along decreasing a posteriori probabilities of success, independent of the MAP. Hence, the profit maximization draws on the available information as well as the chosen prediction model and only requires the product category ROI as an input.

We test our proposed framework in the context of user-generated content networks (e.g., Goldenberg, Oestreicher-Singer, and Reichman 2012; Mayzlin and Yoganarasimhan 2012; Netzer et al. 2012; Trusov, Bodapati, and Bucklin 2010). In our empirical application, we use a unique dataset from SoundCloud, the largest user-generated content network in the music domain with more than 175 million users (Pierce 2016). SoundCloud provides a platform used by music artists around the world to market their songs. User-generated content networks such as SoundCloud generate rich individual-level data that record labels can use, for example, to better assess a music artist's prospects and thus to increase the effectiveness of their marketing investments. Furthermore, user-generated content networks also reveal that most artists will remain unknown (following the power-law degree distribution that is common in social networks, e.g., Barabási 2003; Goldenberg et al. 2009; Rivera, Soderstrom, and Uzzi 2010), and shared content on such networks mainly triggers only short cascades (Watts 2002; Watts and Dodds 2007). The richness of the SoundCloud dataset we use allows us to mimic and analyze the decision problem of record labels that evaluate in how many and, more precisely, in which music artists from a new wave of sign-ups to invest, often by observing them over several months after sign-up. Drawing on artist-specific characteristics, we identify a set of effective early predictors of success (defined in our application as whether a music artist will cross a certain threshold of song-plays in a period of several years after

sign-up). An important finding is that the managers' investment decisions in the context of rare events are mostly inefficient, as the profit realization of our proposed framework is up to four times greater than their current selection policies.

The proposed framework can be built on the model with the best predictive power in terms of probability of detection and probability of false discovery. Our framework recommends an optimal selection size of products or, in our empirical application music artists, independent of whether the MAP indicates success or non-success. More precisely, the optimal selection size represents the number of artists worthy of investment along decreasing a posteriori probabilities of success, which resolves the economic trade-off between wasted investments and lost returns.

The remainder of the paper is organized as follows. In the subsequent section we present an overview of related literature followed by a description of our proposed framework and an empirical application as well as a simulation study. We conclude the paper with a discussion of our findings and implications for marketing practice.

3.2 Background

A large body of literature about the early stages of product growth focuses on the takeoff and its timing, which is driven by multiple factors including price (Golder and Tellis 1997; Foster, Golder, and Tellis 2004) and network externalities (Goldenberg, Libai, and Muller 2010). In the context of online social networks, other drivers of dissemination have been researched, such as word of mouth (Libai, Muller, and Peres 2013; Trusov, Bucklin, and Pauwels 2009), connectedness (Katona, Zubcsek, and Sarvary 2011; Stephen and Toubia 2010), embeddedness (Aral and Walker 2014; Aral and Van Alstyne 2011), and the number of adopters with a high indegree (Goldenberg et al. 2009; Hinz et al. 2011; Yoganarasimhan 2012), which is, however, a point of contention (Aral and Walker 2012; Trusov, Bodapati, and Bucklin 2010; Watts and Dodds 2007). Other research has focused on forecasting and predictions based on these insights (Dover, Goldenberg, and Shapira 2012; Garber et al.

2004; Goldenberg et al. 2009; Goldenberg, Lowengart, and Shapira 2009).

In marketing, forecasting and predictions are not only made with respect to dissemination, but to a host of outcomes that range from predicting a person's product or brand choice (e.g., Chung and Rao 2003; Guadagni and Little 1983; Jacobs, Donkers, and Fok 2016) to predicting their churn decision (e.g., Bolton, Kannan, and Bramlett 2000; Ganesh, Arnold, and Reynolds 2000). In these predictions, the prevalence of all possible future outcomes is typically of the same order of magnitude. However, this is not the case when outcomes include rare events, such as when managers aim to predict the future success of recently launched products. By definition, rare events occur at a very low probability and are therefore heavily underrepresented among all possible future outcomes. As a result, applying a logistic regression or binary probit to predict future successful products leads to an underestimation of the probability of occurrence. King and Zeng (2001b) show that this underestimation is particularly severe if both the mean of the binary variable (the relative frequency of events in the data) and the number of observations is small, and consequently biases the constant term of the regression. Another source of the underestimated probability is the inflexible nature of parametric link functions used for logistic regressions and binary probits (Blattberg, Kim, and Neslin 2010; Kamakura et al. 2005).

Firth (1993) addresses these concerns – especially the associated biased maximum likelihood estimates – and suggests a correction of the logistic likelihood by means of Jeffreys invariant prior, specifically during the maximization procedure. Another and widely used approach, predominantly for the binary logit model, is to oversample the rare event in the training set and then to correct the resulting overestimation with respect to the rare event (e.g., Cosslett 1993; Imbens and Lancaster 1996; Scott and Wild 1997). King and Zeng (2001a,b) suggest applying either a weighting or intercept correction approach when using the logit model. The former refers to maximizing the weighted log-likelihood, a more elegant technique with similar performance compared to the "weighted exogenous sampling maximum-likelihood" method (Manski and Lerman 1977; Singh 2005), which corrects the estimates of the maximum likelihood estimation with the known probability of the rare event.

The effectiveness is highest for sample sizes with several thousands of observations, in which the rare event occurs with a probability less than 5%. King and Zeng (2001a,b) assume that this probability is common knowledge, which is a fair assumption in their application using international conflict data. A third approach is to improve the flexibility of parametric link functions, which still has to assert itself for rare events (Bult and Wansbeek 1995; Naik and Tsai 2004). By applying the bias correction approach to the logit models after balancing the training set (King and Zeng 2001a,b), Donkers, Franses, and Verhoef (2003) achieve better results when predicting defection rates in the insurance domain. This is also true for extensive data sets, as Lemmens and Croux (2006) show in the context of bagging and boosting classification techniques when predicting churn in telecommunications.

In this paper, we are indifferent to the specific prediction model and introduce a managerial framework designed to help managers who make investment decisions in the early stage of product life cycles involving rare success events. In effect, our framework fits any prediction model that is preferred or more suitable for the case at hand. Based on the ranking procedure of Morrison (1969), we propose to rank-order all products according to their a posteriori probabilities of success and then to decide in which ones to invest. Unlike Chatterjee, Hoffman, and Novak (2003) who build on the procedure of Morrison (1969) and propose to select an amount in the prediction set that equals the prevalence of success known from historic data, we acknowledge that any selection size comes at a cost (creating a trade-off between wasted investments and lost returns). Moreover, our proposed managerial framework does not depend on a specific prediction model but can be built on the model with the best predictive power, or any model that the firm uses.

To assess different prediction models as well as predictors, signal detection theory offers a concept to measure the resulting prediction accuracy. Radar researchers initially developed signal detection theory as a tool to distinguish meaningful information from noise (Peterson, Birdsall, and Fox 1954), but the concept quickly gained traction in psychology, where detection theory was combined with statistical decision theory (Green and Swets 1966; Krantz 1969; Swets 1961; Tanner Jr. and Swets 1954; Wickens 2001). Signal detection theory's pop-

ularity in the field of psychophysics (for an overview see Macmillan and Creelman 2004) and memory recognition (e.g., Yonelinas and Parks 2007; Wixted 2007) has inspired applications in marketing that investigate advertising recognition and response (e.g., Cradit, Tashchian, and Hofacker 1994; Mercurio and Forehand 2011; Nielsen, Shapiro, and Mason 2010). The popularity of this theory can be attributed to its potential use to separately measure detectability and response bias. More specifically, it allows for the comparison of respondents with similar detection rates of stimuli presented during an experiment by taking into consideration their false alarms. Similarly, we aim to differentiate between future best-selling products (signal) and failures (noise) based on early-stage data. To this end, we try to find a threshold rule such that the noise is removed. By setting this threshold rule, we consider the probability of detection and the probability of false discovery (for an overview see Kay 2013) before linking these measures to the economic trade-off introduced in this paper. The former probability refers to the ratio of detected future successful products to the number of all market launches, whereas the latter probability refers to the ratio of detected future successful products that turn out to be failures to the number of detections by the threshold rule.

If false discoveries entail high costs – such as in the context of rare events of success following extensive investments – the evaluation of prediction models as well as predictors should specifically include considerations concerning the probability of false discovery (McMorrow 2009; Pinker 2007). This is the case in resource allocation problems of, for example, multinational companies, venture capital funds, and record labels when deciding on how many as well as which products, start-ups, or music artists to select in an early stage of the life cycle: Each selection is associated with a high investment and thus a failed investment carries a high monetary consequence. Therefore, in this paper, we assess the prediction accuracy of models in terms of probability of detection as well as false discovery, and subsequently link these measures to the trade-off between wasted investments and lost returns, which we describe in greater detail below.

3.3 Balancing Lost Returns and Wasted Investments

3.3.1 Probability of Detection and False Discovery

This paper addresses the manager's allocation problem when deciding in how many as well as in which recently launched products to further invest marketing resources. Although success is a rare event, and only few products will eventually become best-sellers in the long term, in most cases a manager can access information about similar products that were launched on the market in the recent past, along with information on their performance over time. This historic data can be used as the training set to identify the model with the best predictive power in terms of probability of detection and probability of false discovery, among a set of alternative models. A moving window, for example of two years, allows the manager to monitor and re-evaluate the chosen prediction model. When faced with an investment allocation decision, the manager can then build on this prediction model and apply it to early-stage sales data (which serves as the prediction set) to select S products for further investment. Based on the training set, the fraction of successful products r is common knowledge. Hence, it is also common knowledge that the approximated number of future successful products among N recent market launches in the prediction set is $n = rN$. Since success in such a context is a rare event, we assume that $n \ll N$. Naturally, use of the prediction model should assist the manager in generating significantly better detections of these n products in the prediction set, compared to a random selection.[10]

The model's predictive power – as a function of the selection size – is assessed in the training set based on simulated out-of-sample predictions, e.g., using historic data in a moving two-year window. In each iteration of the simulation, we use two measures to assess the out-of-sample predictions: First, we set the number of successful products in the respective selection, n_S, in relation to the total number of successful products in the training set, n. In line with signal detection theory (e.g., Peterson, Birdsall, and Fox 1954), we term this

[10]The proportion of successful products in a random selection would be approximately $\frac{n}{N}$.

goodness measure *probability of detection* (PD), also known as the true positive rate:

$$PD(S) = \frac{n_S}{n}\,, \tag{5}$$

where the manager's prediction model should, of course, outperform the random selection, $PD(S) \gg PD_{Random}(S)$[11], assuming a small selection size, $S \ll N$. Increasing the selection size results in a higher probability of detection, which eventually converges to one, $\lim_{S \to N} PD(S) = 1$.

Second, the predictive power of the model is further assessed in terms of the proportion of unsuccessful products, $S - n_S$, in the respective selection, S. In line with signal detection theory (e.g., Peterson, Birdsall, and Fox 1954), we term this goodness measure *probability of false discovery* (PF), also known as false discovery rate:

$$PF(S) = 1 - \frac{n_S}{S}\,, \tag{6}$$

where the prediction model should outperform the random selection, $PF(S) \ll PF(S)_{random}$. Increasing the selection size results in a lower probability of false discovery, which eventually converges to the proportion of unsuccessful products in the product set, $\lim_{S \to N} PD(S) = 1 - \frac{n}{N}$. Since $n_S = n \cdot PD(S)$ (see Equation 5), Equation 6 can be rewritten as

$$PF(S) = 1 - \frac{n}{S} \cdot PD(S)\,. \tag{7}$$

Please note that since $n = rN$, where r and N is common knowledge for any market launch, Equation 7 holds for any selection size S given N products in a recent market launch – based on the probability of detection and probability of false discovery of the chosen prediction model (calculated in the training set).

3.3.2 Wasted Investments Versus Lost Returns

We assume that a manager has a budget B and selects products from a large set of recently launched products, each of which associated with an investment C. The manager allocates

[11] $PD(S)_{Random} = \sum_{n_S=0}^{n} \frac{n_S}{n} Prob_{Random}(n_S|S)$, where $Prob_{Random}(n_S|S)$ is the probability to get n_S successful products when making a random selection of S products, i.e., $\frac{\binom{n}{n_S}\binom{N-n}{S-n_S}}{\binom{N}{S}}$.

his or her budget among the selected products: Hence, the manager can invest in $\frac{B}{C}$ products at most. If a selected product turns out to be a success, then we assume that the manager receives an expected revenue R, otherwise the investment C is lost and no expected revenue is generated, $R = 0$. Recall that based on the training set, the fraction of successful products r is common knowledge. Therefore, it is also common knowledge that the approximated number of future successful products among N recent market launches in the prediction set is $n = rN$. As new product success is a rare event, $n \ll N$, and since the manager's budget is limited, $B \ll NC$, the manager is unable to invest in all products from the recent market launches, $\frac{B}{C} \ll N$, and thus must decide on a selection of products when making an investment.

As outlined above, a manager can apply the chosen prediction model – which is trained and evaluated on historic data (the training set) – to early-stage dissemination data (the prediction set) to select S products from a large set of N recently launched products, where $S \leq \frac{B}{C}$. For the sake of brevity we also assume that $B \geq n \cdot C$, namely that the manager's budget is sufficiently large. Every selection size involves a trade-off. On the one hand, investments in an early stage of the product life cycle are associated with the significant risk that many of the selected products will be failures, which means that certain investments are lost. We term the proportion of the allocated budget that does not generate any revenue as *wasted investments* (WI). The expected value of wasted investments if S products are selected from a large set of recently launched products is given by

$$WI(S) = PF(S) \cdot S \cdot C\,. \tag{8}$$

On the other hand, not selecting a future successful product detracts from the entire return potential, i.e., the manager fails to skim $n \cdot (R - C)$. We term this foregone return potential as *lost returns* (LR). The expected value of such lost returns is given by

$$LR(S) = (1 - PD(S)) \cdot n \cdot (R - C)\,. \tag{9}$$

The decision problem of selecting S products by means of an imperfect prediction model involves considerations about profits $\Pi(S)$. In fact, the manager resolves a trade-off between

wasted investments and lost returns in order to maximize the profit potential of new products, which is given by

$$\Pi(S) = \overline{\Pi} - LR(S) - WI(S) \quad \text{s.t.} \quad S \geq 0\,, \tag{10}$$

where $\overline{\Pi}$ is the maximum profit potential, $\overline{\Pi} = n \cdot (R - C)$. In case the selection size is zero, $S = 0$, then profits are also zero, $\Pi(S = 0) = 0$ (for details on the derivation and further elaborations see Appendix A.4).

Equation 10 implies that wasted investments and lost returns both reduce profits. In addition, the more new products a manager invests in, the higher the wasted investments and the lower the lost returns. Hence, the optimal selection size S^* resolves the trade-off between these two measures. Moreover, Equation 10 reveals that a perfect model would skim the entire profit potential, $\Pi = \overline{\Pi}$, since in a perfect model wasted investments and lost returns are zero, $WI(S) = 0$ and $LR(S) = 0$. In fact, if wasted investments are zero, then the probability of false discovery is also zero, $PF(S) = 0$, since in this case, the manager allocates his or her entire budget to future successful products, $S^* \leq n$. Furthermore, if lost returns are zero, then the probability of detection is one, $PD(S) = 1$, as all future successful products are detected, $S^* \geq n$. Hence, no wasted investments or lost returns implies that the selection size equals the prevalence of future successful products, $S^* = n = n_S$.

To guarantee that S is non-negative, we apply the Kuhn and Tucker (1951) condition and maximize profits by applying the method of Lagrange multipliers. The optimal selection size S^* that resolves the following first order condition (FOC) equation (for the derivation and further elaborations see Appendix A.4) is:

$$\left| \frac{\mathrm{d}LR(S)}{\mathrm{d}S} \right| = \frac{\mathrm{d}WI(S)}{\mathrm{d}S}\,, \tag{11}$$

or equivalently

$$\frac{\partial PF(S)}{\partial S} \cdot S + PF(S) = \frac{ROI}{1 + ROI}\,, \tag{12}$$

where the return on investment (ROI) is the proportion of profits generated by an investment in a successful product, $ROI = \frac{R-C}{C}$. The FOC equation holds – and hence there exists an

internal solution – if and only if $\Pi(S^*) > 0$, namely

$$PF(S^*) < \frac{ROI}{ROI + 1} \,. \tag{13}$$

There exists an internal solution to Equation 13 if the ROI in a given product category is sufficiently high. Otherwise, there exists a corner solution, where the selection size is zero, $S^* = 0$, as it is not worthwhile for a manager to make any investment. Note that $PF(S)$ is known from the training set based on out-of-sample predictions, for example using historic data in a moving two-year window. Therefore, S^* can be derived numerically by resolving Equation 12, which takes into account the constraint given by Equation 13. If the ROI gives a non-zero selection size, $S^* > 0$, then we propose to rank-order all products according to their a posteriori probabilities of success in the prediction set (Chatterjee, Hoffman, and Novak 2003; Morrison 1969) and to subsequently select S^* products along these decreasing probabilities, which then determines (1) in how many and, more precisely, (2) in which products a manager should invest in an early stage of their life cycle.

In contrast, most popular statistical procedures lead to a selection size that is close to or even zero, a direct consequence of applying the maximum a posteriori probability principle (MAP) in the context of rare events, which are by definition not likely to occur. Unlike existing methodological approaches for addressing rare events (e.g., Firth 1993; King and Zeng 2001a,b), our proposed managerial framework does not depend on a specific prediction model but can simply be built on the model with the best predictive power in terms of probability of detection and probability of false discovery. Moreover, our proposed framework does not consider the MAP and suggests an optimal selection size along decreasing a posteriori probabilities of success while taking into account the economic trade-off between wasted investments and lost returns – whether or not these probabilities indicate success or non-success according to the MAP.

In the following section, we report two studies conducted to test our proposed framework. We first introduce the data and the early-stage predictors of success before describing the modeling approach. We then conduct a preliminary study to assess several model specifica-

tions according to their predictive power, measured in terms of probability of detection and probability of false discovery. Based on the specification with the best predictive power, we conduct a second study to test how efficient managers' current selection policies are, and further explore the benefits that our proposed framework offers.

3.4 Empirical Test and Application

3.4.1 Data

We empirically test and illustrate our proposed managerial framework in the context of music artists on SoundCloud, the largest user-generated content network in the domain of music with more than 175 million music artists and fans (Pierce 2016). Hence, the products to which we apply our managerial framework are human brands (Thomson 2006). Users on SoundCloud have various objectives: Music artists market their music by developing their follower base to generate more song-plays, whereas fans follow their favorite artists, listen to their songs, and connect with other fans. We consider 534 artists that signed up to SoundCloud in March 2013 and track them over 123 weeks until July 2015. In empirically testing and illustrating our proposed managerial framework, our objective is to find predictors that effectively detect future successful music artists on SoundCloud several weeks after their sign-up. More specifically, our early-stage observation period for prediction is four, eight, and twelve weeks after an artist's sign-up, and we aim to predict his or her success two years later. We define an artist to be successful if he or she achieves to exceed more than three orders of magnitude in regard to his or her average received monthly song-plays (i.e., $> 1,000$). In the target time period for prediction (May to July 2015), 55 artists in our sample (10%) are successful according to this criterion, while 479 (90%) are not. Of this group, 135 have between 100 and 1,000 average monthly song-plays while 344 artists have fewer than 100 average monthly song-plays. For a detailed description of the data, see Appendix A.5.

To trigger follow-backs (and thus increase their follower base and generate more song-

plays), music artists can reach out to other users by following them, sending them private messages, reposting their songs, commenting on their songs, and liking their songs. In line with Ansari et al. (2016) and Saboo, Kumar, and Ramani (2015), we assume that promotional outgoing activities directly affect success (i.e., song-plays), and indirectly through the artist's egocentric network. The value embedded in an individual's egocentric network is known as social capital (e.g., Coleman 1988, 1990; Putnam 2001) and takes an inside-out view, focusing on the music artist (ego) as well as on the different layers of the artist's follower base (egocentric network). In line with social capital measures (Borgatti, Jones, and Everett 1998; Burt 1983), these layers are most commonly characterized by the structure or density of the egocentric network and, furthermore, by centrality measures (degree, betweenness, and closeness centrality) that represent the artist's position within the entire network (Freeman 1978; Wasserman and Faust 1994). Hence, as predictors for success we consider (1) artists' activities and (2) artists' social capital on SoundCloud. For a detailed description of the theoretical foundation, see Appendix A.5.

3.4.2 Assessing the Predictive Power of Model Specifications – A Preliminary Study

The purpose of this preliminary study is to assess various model specifications according to their predictive power, which we measure in terms of probability of detection and probability of false discovery. More precisely, we make out-of-sample predictions over 1,000 iterations, and in each iteration randomly divide our dataset of 534 unknown music artists into two halves, where one half is treated as the training set and the second half is treated as the prediction set. This allows for the calculation of the probability of detection and the probability of false discovery as a function of the selection size (see Equation 5 and Equation 6), which we use to evaluate the predictive power of the alternative model specifications. To incorporate the different orders of magnitude of average monthly song-plays, we use an ordered-logit model with the following early-stage predictors of success: the average monthly song-plays, the average monthly promotional activities, the total number of uploaded songs, and one

of the following four social capital measures: (1) first-degree followers, (2) second-degree followers, (3) first-degree clustering, and (4) reciprocity. For a detailed description of the modeling approach, see Appendix A.5.

For each of the four specifications of the ordered-logit model along with the three early-stage observation periods, Table 6 exhibits the average probability of detection and probability of false discovery, where averaging is performed over selection sizes of 10, 20, and 30 music artists.[12] Table 6 indicates that even the naïve model, which incorporates only the average monthly song-plays in the early-stage observation period, has some predictive power. However, especially in weeks four and eight, the predictive power of the full model is greater by several percentage points. Specifically, using the model that includes reciprocity increases the average probability of detection by around 5 percentage points and decreases the average probability of false discovery by around 10 percentage points. In this model including reciprocity, Table 6 further shows that the later the prediction is made, the higher the average probability of detection (probability increases from 10.2% in week four to 19.8% in week twelve). Moreover, the later the prediction, the lower the average probability of false discovery (probability declines from 72.1% in week four to 45.9% in week twelve). Nonetheless, even at the end of the twelve-week prediction period, the music artist is still at a very early stage of establishing his or her follower base. Although a cohesive follower base is pivotal for long-term success (Ansari et al. 2016), first-degree clustering presumably plays a stronger role as the artist's career progresses: First, the artist triggers follow-backs by reaching out to other SoundCloud users through promotional activities. By strengthening ties to users that followed back and establishing bi-directional connections (increasing reciprocity), an artist can build trust (Coleman 1988). Only then does the connectivity within the follower base potentially come into play, when an artist's followers start building a cohesive fan community by following each other.

[12] These selection sizes are equivalent to around 33%, 66%, and 100% of the prevalence of future successful products in the prediction set. Note that there are 27 future successful artists in the prediction set on average.

— Insert Table 6 about here —

Focusing now solely on the model specification that includes reciprocity, we compare the ordered-logit model to three benchmarks models that are widely used in the context of rare events: (1) *logit*, (2) *Firth* (Firth 1993; Heinze 2006; Heinze and Schemper 2002), and (3) *ReLogit* (Imai, King, and Lau 2008; King and Zeng 2001a,b). For each of the three early-stage observation periods, we compare the models' predictive power by calculating the average probability of detection and average probability of false discovery following the same simulated out-of-sample predictions as for the above evaluation of the best early-stage predictors of success.

Table 7 shows that in week 4, the benchmark models collapse as a result of the insufficient information contained in the data due to the lack of variation when only a binary success variable is employed. In terms of the average probability of false discovery, the ordered-logit model dominates the benchmark models by more than 15 percentage points. However, the logit and Firth benchmark models' average probabilities of detection in week eight and week twelve outperform the ordered-logit model by around 2 percentage points, while the ReLogit model generates lower values for the average probability of detection. Yet, the economic impact of a higher probability of false discovery is more severe, especially if the single investment in a music artist is high. Hence, every investment that turns out to be a failure has significant monetary consequences, strongly reducing profits. Along these lines, Equation A.97 shows that given an ROI and a selection size, expected profits decline as the probability of false discovery increases. We therefore conclude that the ordered-logit model outperforms the benchmark models.

— Insert Table 7 about here —

Figure 6 presents for the ordered-logit model – and at a higher resolution – the probability of detection and probability of false discovery (as a function of the selection size)

for predictions four (panels A and B), eight (panels C and D), and twelve weeks (panels E and F) after the music artists' sign-up. Congruent to the findings in Table 6 and Table 7, Figure 6 reveals that the probabilities of detection increase and the probabilities of false discovery decline the later the prediction is made. Furthermore, Figure 6 again reveals that the full model outperforms the naïve model, which only incorporates the average monthly song-plays. Naturally, the results of both models remain within the boundaries given by the perfect model and the random selection (see discussion of Equation 10).

— Insert Figure 6 about here —

3.4.3 The Effectiveness of Current Selection Policies: Assessing Managers' Performance

After assessing the predictive power of several model specifications in terms of probability of detection and probability of false discovery, it remains to explore whether the proposed managerial framework outperforms managers' current selection policies. The purpose of this study is to test the effectiveness of conventional managerial selection policies and demonstrate the benefits of the proposed framework. Therefore, we compare the monetary outcome of the selection of music artists recommended by our proposed framework against selections based on the following popular policies currently used by managers: (1) the *naïve selection*, namely to choose the amount of music artists along decreasing a posteriori probabilities that equals the prevalence of success known from historic data[13] (Chatterjee, Hoffman, and Novak 2003; Morrison 1969), (2) the *MAP*, namely to select all artists that feature a posteriori probabilities indicating success rather than non-success (Gallager 2013), and (3) the *random selection*, where the selection size equals the prevalence of success as in the rare event selection.

[13]Note that the fraction of successful music artists r is common knowledge based on the training set. Hence, it is also common knowledge that the approximated number of future successful artists among N recent sign-ups in the prediction set is $n = r \cdot N$.

We conduct a simulation of the steps a manager at a record label would have to undertake in practice when following our proposed managerial framework and the three commonly used policies noted above: the naïve selection, the MAP, and the random selection. Hence, we distinguish between historic data and early-stage data. According to our proposed framework, a manager uses historic data to identify the model with the best predictive power in terms of probability of detection and probability of false discovery, and calculates the optimal selection size. Then, the manager can build on the trained and evaluated prediction model and apply it to early-stage dissemination data. Therefore, in each iteration of our simulation study, we randomly divide the sample into thirds, where two thirds are treated as historic data and one third is treated as early-stage dissemination data up to four, eight, and twelve weeks. Specifically, we calculate profit realizations in 1,000 iterations, where the mechanism *in each iteration* is based on the following five sequential steps.

First, the mechanism randomly divides the sample of 534 unknown music artists into thirds, where two thirds are treated as historic data (which serves as the training set) and one third is treated as early-stage dissemination data up to four, eight, and twelve weeks (which serves as the prediction set). Crucial for the profit calculation, the prediction set also contains information on the final outcome (whether or not the artist's average monthly song-plays exceeded three orders of magnitude). Note that in our simulation we use the final outcome only for the sake of validation.

Second, in each of the 1,000 iterations, 1,000 out-of-sample predictions are made based solely on the training set (determined in step 1) by randomly dividing this training set into two halves. This allows for assessment of the predictive power of the chosen model (in terms of probability of detection and probability of false discovery) as a function of the selection size (see Equation 5 and Equation 6), which are averaged over the 1,000 out-of-sample predictions. These two goodness measures thus become common knowledge on the basis of the training set.

Third, based on the resulting average probability of false discovery, the optimal selec-

tion size is calculated to maximize expected profits given the ROI (see Equation 10 and Equation A.97). More specifically, the mechanism resolves the optimal selection size S^* numerically since the probability of false discovery $PF(S)$ allows for the direct calculation of *expected profits*,

$$\frac{\Pi}{C} = S \cdot (ROI - PF(S) - ROI \cdot PF(S)) \,, \tag{14}$$

where the units are given by C, which denotes the investment associated with selecting one music artist from a new wave of sign-ups, and $ROI = \frac{R-C}{C}$. The optimal selection size $S*$ is given by the selection size that maximizes expected profits, $\frac{\Pi}{C}(S^*|ROI)$, such that S^* is non-negative and thus the expected profits are non-negative. Hence, the training set generates the average probability of detection and average probability of false discovery, along with the optimal selection size.

Fourth, the mechanism uses the entire training set (determined in step 1), calibrates the model parameters, and applies the model to the prediction set (determined in step 1) to generate a posteriori probabilities of success for predictions four, eight, and twelve weeks after the artists' sign-up.

Fifth, all music artists are rank-ordered according to their a posteriori probabilities of success in the prediction set, and the optimal selection size, which depends on the respective ROI, is calculated. In each iteration, the profit realization is calculated for each selection size S^*, to allow a comparison of our proposed managerial framework against the managers' commonly used selection policies. In the naïve as well as the random selection policy, the selection size equals the prevalence of success in the training set, where the former suggests choosing these music artists not randomly but along decreasing a posteriori probabilities of success. In comparison to the naïve selection, our proposed framework takes into account the economic trade-off between wasted investments and lost returns along decreasing a posteriori probabilities of success and suggests a selection size S^* that maximizes expected profits. The selection policy based on the MAP only suggests selecting music artists whose a posteriori probabilities indicate success, namely if they are greater than 50%. The profit realization of

a selection is the ratio between the total realized profits, $\Pi = n_S R - S^* C$, and the maximum profit potential, $\overline{\Pi} = n \cdot (R - C)$, and is calculated as follows:

$$\frac{\Pi}{\overline{\Pi}} = \frac{n_S \cdot (ROI + 1) - S^*}{n \cdot ROI} \,, \tag{15}$$

where n_S is the proportion of eventually successful music artists in the selection S^*, R is the expected revenue if a selected artist turns out to be a success, C is the associated investment in a selected artist, and $ROI = \frac{R-C}{C}$. Equation 10 reveals that a perfect prediction model would skim the entire profit potential, $\Pi = \overline{\Pi}$, since in a perfect model, wasted investments and lost returns are zero and the selection size equals the prevalence of future successful products, $S^* = n = n_S$. Therefore, it follows that in a perfect prediction model, the profit realization is maximal, $\frac{\Pi}{\overline{\Pi}} = 1$.

Based on the above-described simulation study mimicking the steps a manager at a record label would have to undertake in practice, Figure 7 exhibits the average profit realization – as a function of the ROI – for predictions four (panel A), eight (panel B), and twelve weeks (panel C) after the music artists' sign-up. Interestingly, the random selection policy outperforms the MAP, especially for higher levels of ROI. While the selection size is close to or even zero for the MAP, the future successful artists detected by the random selection by chance alone compensate, at higher levels of ROI, for the monetary consequences of the loss of investments in unsuccessful artists. The later the prediction is made and thus the weaker the noise compared to the signal incorporated in the model, the greater the chance that the MAP-based selection indicates success, which then results in a non-zero selection size. Therefore, the later the prediction is made, the greater the profit realization generated by the MAP, although it increases only marginally from 0% (week 4) to 7.6% (week 12). On the other hand, the naïve selection policy suggests choosing the same number of artists as the random selection policy, but does so along decreasing a posteriori probabilities of success. Indeed, the realized profits triple, but this selection fails to take into account the trade-off between wasted investments and lost returns. Our proposed managerial framework addresses this trade-off and, based on expected profits, recommends a selection size for each ROI, whereas the benchmark approaches choose the same selection size independent of the

ROI. Hence, in all early-stage observation periods for prediction and especially at higher levels of ROI, profit realization of our proposed framework is up to four times greater than the benchmark approaches (see Figure 7).

— Insert Figure 7 about here —

The incorporation of the economic trade-off along decreasing a posteriori probabilities enhances decision-making, and highlights the gross inefficiency of managers' current selection policies in the context of rare events. Based on the probability of detection and probability of false discovery as a function of selection size (see Figure 6), which we link to wasted investments and lost returns, our proposed framework calculates the optimal selection size S^* as a function of ROI. The later the prediction is made, the richer the available information, and the better the prediction model, the lower is the threshold for ROI to select at least one music artist among a new wave of sign-ups. In other words, there is a threshold below which it is not worthwhile to make any investment. Figure 8 exhibits the optimal selection size S^* as a function of ROI and reveals that this minimum threshold for a non-zero selection size is 184%, 77%, and 13% for predictions four, eight, and twelve weeks after the music artists' sign-up. Hence, an investment of \$100,000 dollars in a successful music artist has to result in at least a total return of \$284,000, \$177,000, and \$113,000 dollars. If the ROI is above the minimum threshold and amounts to for example 500%, then the optimal selection size increases to 16, 27, and 35 music artists four, eight, and twelve weeks after sign-up.

— Insert Figure 8 about here —

3.5 Discussion

Our paper offers managers a quantitative framework for allocating marketing resources to products shortly after their market launch. The commonly used procedure to choose an

optimal selection of products for further investment is to employ a probability model, which predicts success based on the MAP. Hence, a product is selected if it is more likely to be a success than a non-success. However, this approach is not effective in the context of rare events, because success is determined by various factors – most of them may be even unobserved – where in practical models many of these factors are considered as noise. A typical feature of early-stage prediction is that this noise is much stronger than the signal incorporated by the model. Hence, applying the MAP when predicting success of products results in an optimal selection size of close to or even zero because in reality most if not all products are more likely to be a failure.

When investing in only a single new product, managers face enormous lost returns since it is almost certain that at least several future successful products are excluded from investments of marketing resources. At the other extreme, when investing in all new products, managers will waste extensive resources on the many products that later turn out to be failures. Along these lines, we propose and test a framework to extend current literature by (1) evaluating the predictive power of models in terms of probability of detection and probability of false discovery (Peterson, Birdsall, and Fox 1954), (2) rank-ordering all products according to their a posteriori probabilities of success (Chatterjee, Hoffman, and Novak 2003; Morrison 1969), and (3) resolving the economic trade-off between wasted investments and lost returns along decreasing a posteriori probabilities of success, which subsequently determines in how many and, more precisely, in which products a manager should invest in an early stage of their life cycle.

We empirically test and illustrate our proposed framework in the context of user-generated content networks. We use a dataset of SoundCloud, the largest user-generated content network in the music domain, and recommend in how many as well as which music artists from a new wave of sign-ups to invest, a common decision problem of record labels. We find that the managers' investment decisions in the context of rare events are mostly inefficient, as the profit realization of our proposed framework is up to four times greater than their current selection policies.

Our framework is not only geared towards improving resource allocation at companies with broad and deep product pipelines (e.g., Nestlé) or social networking platforms with millions of users (e.g., SoundCloud). On the contrary, our proposed framework can also assist venture capitalists to make better-informed investment decisions when evaluating start-ups. Based on historic data, the venture capitalist would first identify the model with the best predictive power in terms of probability of detection and probability of false discovery, before applying it to early-stage dissemination data of new investment opportunities, such as sales figures, market size projections, or qualitative information about the founding team. Although managers may be tempted to eschew quantitative decision-making methods in early stages of the start-ups' life cycles, when signals are weak and seem to have limited value, our proposed framework offers a practical tool to handle the available information in such noisy environments. We even argue that our framework applies to instances where the sample is balanced – unlike in the case of rare events – as the major contribution of this paper lies on the trade-off between wasted investments and lost returns that is associated with any selection size. Our proposed framework does not consider the MAP and suggests an optimal selection size along decreasing a posteriori probabilities of success while taking into account this economic trade-off – whether or not these probabilities indicate success or non-success according to the MAP.

The universal applicability of our proposed framework may be affected by several limitations: First, it relies exclusively on quantitative measures and does not integrate qualitative ones, which, in some cases, such as the assessment of start-ups, may be significant in predicting future success. Second, our proposed framework depends on historic data covering successful cases as well as failures when identifying the model with the best predictive power. Absence of such data can limit the effectiveness of this approach. Third, our proposed framework also depends on early-stage dissemination data and thus focuses on post-launch resource allocation. Future research should therefore be geared towards the empirical validation of our framework's generalizability to other industries, other stages in the product life cycle, and the incorporation of qualitative measures. We hope this paper encourages work in these

and related directions.

4 General Conclusion

This dissertation sets out (1) to investigate optimal seeding policies (e.g., as a creator of music), and (2) to create a managerial framework for early-stage investments in creators (e.g., as a record label). In the context of user-generated content networks, which constitute a unique type of social networking platform (e.g., Goldenberg, Oestreicher-Singer, and Reichman 2012; Mayzlin and Yoganarasimhan 2012; Trusov, Bodapati, and Bucklin 2010), each of the two embedded essays covers one of these topics. More precisely, the former takes the perspective of the creator – i.e., how to quickly accumulate followers after signing up to a user-generated content network – and the latter takes the perspective of the investor – i.e., how to invest in such recent sign-ups.

The first essay, co-authored with Jacob Goldenberg, Daniel Shapira, and Florian Stahl, entitled *"Climb or Jump – Status-Based Seeding in User-Generated Content Networks"* challenges the role of influencers or high status individuals. Based on unique datasets of SoundCloud and by means of empirical as well as analytical analyses, we study creators of music who seek to build and increase their follower base by directing promotional actions to other users of the networking platform. Current social network literature in marketing suggests that such creators should reach out exclusively to individuals with the highest status in the network in terms of centrality (e.g., Goldenberg et al. 2009; Hinz et al. 2011; Libai, Muller, and Peres 2013; Yoganarasimhan 2012). In fact, topical research treats seeding as a matter of choice, which does not involve risk in the form of time constraints, search costs to find seeding targets, anti-spam policies, or differences in responsiveness. However, if the difference of network status between the creator and the seeding target matters, the creator's budget of promotional actions is constrained, and since different levels of returns are associated with different levels of responsiveness, it follows that creators must solve a risk versus return trade-off when choosing their portfolios of seeding targets. Hence, the optimal

seed is not necessarily the individual with the highest status, especially if the probability to follow back varies. As a consequence, unknown creators who seek to build and increase their follower base, targeting high-status individuals is associated with high risk (due to their low responsiveness), while targeting low-status individuals is associated with lower risk (due to their higher responsiveness). Indeed, our empirical analysis of creators' revealed preferences on SoundCloud indicates that the higher the status difference is between the creator and the seeding target, the lower the a priori probability of a follow back. Further analyses also reveal that creators of music spread their budgets of promotional actions over several orders of magnitude in terms of status, meaning they choose a portfolio and do not send only to a certain status – and definitely not exclusively to individuals with the highest status in the network. Finally, our simulations show that unknown creators who seek to build and increase their follower base should ignore predominant seeding policies and slowly "climb" across status levels of seeding targets rather than attempting to "jump" towards those with the highest status. Hence, unknown creators should invest in seeding targets with the lowest status, instead of chasing indirect returns by directing promotional actions to high-status individuals.

The second essay, with the same co-authors, entitled *"Allocation of Marketing Budget When Success Is a Rare Event"* addresses the managerial decision problem of allocating marketing resources to a set of unknown creators of content, recently launched products, or start-ups. We propose and test a framework to extend the current literature by evaluating the predictive power of models in terms of probability of detection and probability of false discovery (Peterson, Birdsall, and Fox 1954), rank-ordering all creators, products, and start-ups according to their a posteriori probabilities of success (Chatterjee, Hoffman, and Novak 2003; Morrison 1969), and resolving the economic trade-off between wasted investments and lost returns along decreasing a posteriori probabilities of success, which subsequently determines in how many and, more precisely, in which creators, products, and start-ups a manager should invest in an early stage of their life cycle. We empirically test and illustrate our proposed framework in the context of SoundCloud and recommend in how many as

well as which music artists from a new wave of sign-ups to invest in – a common decision problem of record labels. We find that the managers' investment decisions in the context of rare events are mostly inefficient, as the profit realization of our proposed framework is up to four times greater than their current selection policies.

Bibliography

Ansari, Asim, Oded Koenigsberg, and Florian Stahl (2011), "Modeling Multiple Relationships in Social Networks," *Journal of Marketing Research*, 48 (4), 713–728.

Ansari, Asim, Florian Stahl, Mark Heitmann, and Lucas Bremer (2016), "Building a Social Network for Success," *working paper*.

Aral, Sinan, Lev Muchnik, and Arun Sundararajan (2009), "Distinguishing Influence-Based Contagion from Homophily-Driven Diffusion in Dynamic Networks," *Proceedings of the National Academy of Sciences*, 106 (51), 21544–21549.

Aral, Sinan and Marshall Van Alstyne (2011), "The Diversity-Bandwidth Trade-Off," *American Journal of Sociology*, 117 (1), 90–171.

Aral, Sinan and Dylan Walker (2012), "Identifying Influential and Susceptible Members of Social Networks," *Science*, 337 (6092), 337–341.

——— (2014), "Tie Strength, Embeddedness, and Social Influence: A Large-Scale Networked Experiment," *Management Science*, 60 (6), 1352–1370.

Archak, Nikolay, Anindya Ghose, and Panagiotis G. Ipeirotis (2011), "Deriving the Pricing Power of Product Features by Mining Consumer Reviews," *Management Science*, 57 (8), 1485–1509.

Ball, Sheryl, Catherine Eckel, Philip J. Grossman, and William Zame (2001), "Status in Markets," *Quarterly Journal of Economics*, 116 (1), 161–188.

Bampo, Mauro, Michael T. Ewing, Dineli R. Mather, David Stewart, and Mark Wallace (2008), "The Effects of the Social Structure of Digital Networks on Viral Marketing Performance," *Information Systems Research*, 19 (3), 273–290.

Barabási, Albert-László (2003), *Linked*, New York, NY: Plume.

Berger, Jonah and Katherine L. Milkman (2012), "What Makes Online Content Viral?" *Journal of Marketing Research*, 49 (2), 192–205.

Bettencourt, Ann, Kelly Charlton, Nancy Dorr, and Deborah L. Hume (2001), "Status Differences and In-Group Bias: A Meta-Analytic Examination of the Effects of Status Stability, Status Legitimacy, and Group Permeability," *Psychological Bulletin*, 127 (4), 520.

Blattberg, Robert C., Byung-Do Kim, and Scott A. Neslin (2010), *Database Marketing: Analyzing and Managing Customers*, New York, NY: Springer.

Blau, Peter M. (1963), *The Dynamics of Bureaucracy: Study of Interpersonal Relations in Two Government Agencies (rev. ed.*, Chicago, IL: University of Chicago Press.

Bolton, Ruth N., P.K. Kannan, and Matthew D. Bramlett (2000), "Implications of Loyalty Program Membership and Service Experiences for Customer Retention and Value," *Journal of the Academy of Marketing Science*, 28 (1), 95–108.

Bond, David (2016), "Social Media Celebrities Warned on Sponsorship Disclosure," (accessed March 8, 2017), [available at `https://www.ft.com/content/26d91166-6d3d-11e6-9ac1-1055824ca907`].

Borgatti, Stephen P., Candace Jones, and Martin G. Everett (1998), "Network Measures of Social Capital," *Connections*, 21 (2), 27–36.

Bult, Jan Roelf and Tom Wansbeek (1995), "Optimal Selection for Direct Mail," *Marketing Science*, 14 (4), 378–394.

Burt, Ronald S. (1983), "Range," in Ronald S. Burt and Michael J. Minor, eds., *Applied Network Analysis*, Beverly Hills, CA: Sage, 176–194.

Buss, David M. and Michael Barnes (1986), "Preferences in Human Mate Selection. Journal of Personality and Social Psychology," *Journal of Personality and Social Psychology*, 50 (3), 559–570.

Buston, Peter M. and Stephen T. Emlen (2003), "Cognitive Processes Underlying Human Mate Choice: The Relationship Between Self-Perception and Mate Preference in Western Society," *Proceedings of the National Academy of Sciences*, 100 (15), 8805–8810.

Chae, Inyoung, Andrew T. Stephen, Yakov Bart, and Yao Dai (2016), "Spillover Effects In Seeded Word-Of-Mouth Marketing Campaigns," *Marketing Science*, forthcoming.

Chatterjee, Patrali, Donna L. Hoffman, and Thomas P. Novak (2003), "Modeling the Clickstream: Implications for Web-Based Advertising Efforts," *Marketing Science*, 22 (4), 520–541.

Chung, Jaihak and Vithala R. Rao (2003), "A General Choice Model for Bundles with Multiple-Category Products: Application to Market Segmentation and Optimal Pricing for Bundles," *Journal of Marketing Research*, 40 (2), 115–130.

Coleman, James S. (1988), "Social Capital in the Creation of Human Capital," *American Journal of Sociology*, 94, 95–120.

——— (1990), *Foundations of Social Theory*, Cambridge, MA: Harvard University Press.

Cosslett, Stephen R. (1993), "Estimation From Endogenously Stratified Samples," *Handbook of Statistics*, 11, 1–43.

Cradit, J. Dennis, Armen Tashchian, and Charles F. Hofacker (1994), "Signal Detection Theory and Single Observation Designs: Methods and Indices for Advertising Recognition Testing," *Journal of Marketing Research*, 31 (1), 117–127.

Donkers, Bas, Philip Hans Franses, and Peter C. Verhoef (2003), "Selective Sampling for Binary Choice Models," *Journal of Marketing Research*, 40 (4), 492–497.

Dover, Yaniv, Jacob Goldenberg, and Daniel Shapira (2012), "Network Traces on Penetration: Uncovering Degree Distribution From Adoption Data," *Marketing Science*, 31 (4), 689–712.

Easley, David and Jon Kleinberg (2010), *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*, Cambridge, UK: Cambridge University Press.

Eliashberg, Jehoshua (1980), "Consumer Preference Judgments: An Exposition with Empirical Applications," *Management Science*, 26 (1), 60–77.

Ellemers, Naomi, Bertjan Doosje, Ad Van Knippenberg, and Henk Wilke (1992), "Status Protection in High Status Minority Groups," *European Journal of Social Psychology*, 22 (2), 123–140.

Ellemers, Naomi, Ad Van Knippenberg, Nanne De Vries, and Henk Wilke (1988), "Social Identification and Permeability of Group Boundaries," *European Journal of Social Psychology*, 18 (6), 497–513.

Everett, Martin G. and Stephen P. Borgatti (2005), "Ego Network Betweenness," *Social Networks*, 27 (1), 31–38.

Feingold, Alan (1990), "Gender Differences in Effects of Physical Attractiveness on Romantic Attraction: A Comparison Across Five Research Paradigms," *Journal of Personality and Social Psychology*, 59 (5), 981.

Firth, David (1993), "Bias Reduction of Maximum Likelihood Estimates," *Biometrika*, 80 (1), 27–38.

Fischer, Marc, Sönke Albers, Nils Wagner, and Monika Frie (2011), "Dynamic Marketing Budget Allocation Across Countries, Products, and Marketing Activities," *Marketing Science*, 30 (4), 568–585.

Foster, Joseph A., Peter N. Golder, and Gerard J. Tellis (2004), "Predicting Sales Takeoff for Whirlpool's New Personal Valet," *Marketing Science*, 23 (2), 182–185.

Freeman, Linton C. (1978), "Centrality in Social Networks: Conceptual Clarification," *Social Networks*, 1 (3), 215–239.

Gallager, Robert G. (2013), *Stochastic Processes: Theory for Applications*, New York, NY: Cambridge University Press.

Ganesh, Jaishankar, Mark J. Arnold, and Kristy E. Reynolds (2000), "Understanding the Customer Base of Service Providers: An Examination of the Differences Between Switchers and Stayers," *Journal of Marketing*, 64 (3), 65–87.

Garber, Tal, Jacob Goldenberg, Barak Libai, and Eitan Muller (2004), "From Density to Destiny: Using Spatial Dimension of Sales Data for Early Prediction of New Product Success," *Marketing Science*, 23 (3), 419–428.

Gladwell, Malcolm (2000), *The Tipping Point*, New York, NY: Little, Brown and Company.

Goldenberg, Jacob, Sangman Han, Donald R. Lehmann, and Jae Weon Hong (2009), "The Role of Hubs in the Adoption Process," *Journal of Marketing*, 73 (2), 1–13.

Goldenberg, Jacob, Barak Libai, and Eitan Muller (2010), "The Chilling Effects of Network Externalities," *International Journal of Research in Marketing*, 27 (1), 4–15.

Goldenberg, Jacob, Oded Lowengart, and Daniel Shapira (2009), "Zooming in: Self-Emergence of Movements in New Product Growth," *Marketing Science*, 28 (2), 274–292.

Goldenberg, Jacob, Gal Oestreicher-Singer, and Shachar Reichman (2012), "The Quest for Content: How User-Generated Links Can Facilitate Online Exploration," *Journal of Marketing Research*, 49 (4), 452–468.

Golder, Peter N. and Gerard J. Tellis (1997), "Will it Ever Fly? Modeling the Takeoff of Really New Consumer Durables," *Marketing Science*, 16 (3), 256–270.

Goode, William J. (1978), *The Celebration of Heroes: Prestige as a Control System*, Berkley, CA: University of California Press.

Gould, Roger V. (2002), "The Origins of Status Hierarchies: A Formal Theory and Empirical Test," *American Journal of Sociology*, 107 (5), 1143–1178.

Granovetter, Mark S. (1973), "The Strength of Weak Ties," *American Journal of Sociology*, 78 (6), 1360–1380.

Green, D.M. and John A. Swets (1966), *Signal Detection Theory and Psychophysics*, New York, NY: Wiley.

Guadagni, Peter M. and John D.C. Little (1983), "A Logit Model of Brand Choice Calibrated on Scanner Data," *Marketing Science*, 2 (3), 203–238.

Hadar, Josef and William R. Russell (1969), "Rules For Ordering Uncertain Prospects," *The American Economic Review*, 59 (1), 25–34.

Haenlein, Michael and Barak Libai (2013), "Targeting Revenue Leaders for a New Product," *Journal of Marketing*, 77 (3), 65–80.

Hanaki, Nobuyuki, Alexander Peterhansl, Peter S. Dodds, and Duncan J. Watts (2007), "Cooperation in Evolving Social Networks," *Management Science*, 53 (7), 1036–1050.

Hargadon, Andrew and Robert I. Sutton (1997), "Technology Brokering and Innovation in a Product Development Firm," *Administrative Science Quarterly*, 42 (4), 716–749.

Hauser, John R. (1978), "Consumer Preference Axioms: Behavioral Postulates for Describing and Predicting Stochastic Choice," *Management Science*, 24 (13), 1331–1341.

Hauser, John R. and Glen L. Urban (1977), "A Normative Methodology for Modeling Consumer Response to Innovation," *Operations Research*, 25 (4), 579–619.

——— (1979), "Assessment of Attribute Importances and Consumer Utility Functions: Von Neumann–Morgenstern Theory Applied to Consumer Behavior," *Journal of Consumer Research*, 5 (4), 251–262.

Heino, Rebecca D., Nicole B. Ellison, and Jennifer L. Gibbs (2010), "Relationshopping: Investigating the Market Metaphor in Online Dating," *Journal of Social and Personal Relationships*, 27 (4), 427–447.

Heinze, Georg (2006), "A Comparative Investigation of Methods for Logistic Regression with Separated or Nearly Separated Data," *Statistics in Medicine*, 25 (24), 4216–4226.

Heinze, Georg and Michael Schemper (2002), "A Solution to the Problem of Separation in Logistic Regression," *Statistics in Medicine*, 21 (16), 2409–2419.

Hinz, Oliver, Bernd Skiera, Christian Barrot, and Jan U. Becker (2011), "Seeding Strategies for Viral Marketing: An Empirical Comparison," *Journal of Marketing*, 75 (6), 55–71.

Hu, Yansong and Christophe Van den Bulte (2014), "Nonmonotonic Status Effects in New Product Adoption," *Marketing Science*, 33 (4), 509–533.

Imai, Kosuke, Gary King, and Olivia Lau (2008), "Toward a Common Framework for Statistical Analysis and Development," *Journal of Computational and Graphical Statistics*, 17 (4), 892–913.

Imbens, Guido W. and Tony Lancaster (1996), "Efficient Estimation and Stratified Sampling," *Journal of Econometrics*, 74 (2), 289–318.

Iyengar, Raghuram, Christophe Van den Bulte, and Thomas W. Valente (2011), "Opinion Leadership and Social Contagion in New Product Diffusion," *Marketing Science*, 30 (2), 195–212.

Jackson, Matthew O. (2010), *Social and Economic Networks*, Princeton, NJ: Princeton University Press.

Jacobs, Bruno J.D., Bas Donkers, and Dennis Fok (2016), "Model-Based Purchase Predictions for Large Assortments," *Marketing Science*, 35 (3), 389–404.

Kamakura, Wagner, Carl F. Mela, Asim Ansari, Anand Bodapati, Pete Fader, Raghuram Iyengar, Prasad Naik, Scott Neslin, Baohong Sun, Peter C. Verhoef et al. (2005), "Choice Models and Customer Relationship Management," *Marketing Letters*, 16 (3), 279–291.

Katona, Zsolt, Peter P. Zubcsek, and Miklos Sarvary (2011), "Network Effects and Personal Influences: The Diffusion of an Online Social Network," *Journal of Marketing Research*, 48 (3), 425–443.

Katz, Elihu and Paul F. Lazarsfeld (1955), *Personal Influence: The Part Played by People in the Flow of Mass Communications*, Glencoe, IL: The Free Press.

Katz, Leo (1953), "A New Status Index Derived From Sociometric Analysis," *Psychometrika*, 18 (1), 39–43.

Kay, Steven M. (2013), *Fundamentals of Statistical Signal Processing: Practical Algorithm Development*, vol. 3, Upper Saddle River, NJ: Prentice-Hall.

King, Gary and Langche Zeng (2001a), "Explaining Rare Events in International Relations," *International Organization*, 55 (3), 693–715.

——— (2001b), "Logistic Regression in Rare Events Data," *Political Analysis*, 9 (2), 137–163.

Kirby, Justin and Paul Marsden (2006), *Connected Marketing: The Viral, Buzz and Word of Mouth Revolution*, London, UK: Routledge.

Koblin, John (2015), "American Idol Will End Its Run in 2016," (accessed October 11, 2016), [available at `http://nyti.ms/1F1FGm3`].

Kowner, Rotem (1995), "The Effect of Physical Attractiveness Comparison on Choice of Partners," *The Journal of Social Psychology*, 135 (2), 153–165.

Kozinets, Robert V., Kristine De Valck, Andrea C. Wojnicki, and Sarah J.S. Wilner (2010), "Networked Narratives: Understanding Word-of-Mouth Marketing in Online Communities," *Journal of Marketing*, 74 (2), 71–89.

Krantz, David H. (1969), "Threshold Theories of Signal Detection." *Psychological rReview*, 76 (3), 308.

Kuhn, Harold W. and Albert W. Tucker (1951), "Nonlinear Programming," in *Proceedings of the 2nd Berkley Symposium*, Monterey, CA: University of California Press, 481–492.

Latour, Bruno (1987), *Science in Action: How to Follow Scientists and Engineers Through Society*, Cambridge, MA: Harvard University Press.

Lee, Leonard, George Loewenstein, Dan Ariely, James Hong, and Jim Young (2008), "If I'm Not Hot, Are You Not or Not? Physical-Attractiveness Evaluations and Dating Preferences as a Function of One's Own Attractiveness," *Psychological Science*, 19 (7), 669–677.

Lemmens, Aurélie and Christophe Croux (2006), "Bagging and Boosting Classification Trees to Predict Churn," *Journal of Marketing Research*, 43 (2), 276–286.

Lemon, Katherine N. (2016), "The Art of Creating Attractive Consumer Experiences at the Right Time: Skills Marketers Will Need to Survive and Thrive," *GfK Marketing Intelligence Review*, 8 (2), 44–49.

Libai, Barak, Eitan Muller, and Renana Peres (2005), "The Role of Seeding in Multi-Market Entry," *International Journal of Research in Marketing*, 22 (4), 375–393.

——— (2013), "Decomposing the Value of Word-Of-Mouth Seeding Programs: Acceleration Versus Expansion," *Journal of Marketing Research*, 50 (2), 161–176.

Lilly, Amanda (2014), "Ultra Music President On the Next Big Trend In Electronic Music," (accessed March 8, 2017), [available at `http://blogs.wsj.com/speakeasy/2014/08/18/ultra-music-president-on-the-next-big-trend-in-electronic-music/`].

Lionberger, Herbert F. (1953), "Some Characteristics of Farm Operators Sought as Sources of Farm Information in a Missouri Community," *Rural Sociology*, 18 (4), 327–338.

Little, Anthony C., D. Michael Burt, Ian S. Penton-Voak, and David I. Perrett (2001), "Self-Perceived Attractiveness Influences Human Female Preferences for Sexual Dimorphism and

Symmetry in Male Faces," *Proceedings of the Royal Society of London B: Biological Sciences*, 268 (1462), 39–44.

Macmillan, Neil A. and C. Douglas Creelman (2004), *Detection Theory: A User's Guide*, Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Maheshwari, Sapna (2016), "Endorsed on Instagram by a Kardashian, but Is It Love or Just an Ad?" (accessed September 8, 2016), [available at `http://nyti.ms/2c1sVAp`].

Manski, Charles F. and Steven R. Lerman (1977), "The Estimation of Choice Probabilities from Choice Based Samples," *Econometrica*, 45 (8), 1977–1988.

Markovitch, Dmitri G. and Peter N. Golder (2008), "Findings-Using Stock Prices to Predict Market Events: Evidence on Sales Takeoff and Long-Term Firm Survival," *Marketing Science*, 27 (4), 717–729.

Mayzlin, Dina and Hema Yoganarasimhan (2012), "Link to Success: How Blogs Build an Audience by Promoting Rivals," *Management Science*, 58 (9), 1651–1668.

McMorrow, Dan (2009), *Rare Events*, McLean, VA: The MITRE Corporation.

Mercurio, Kathryn R. and Mark R. Forehand (2011), "An Interpretive Frame Model of Identity-Dependent Learning: The Moderating Role of Content-State Association," *Journal of Consumer Research*, 38 (3), 555–577.

Merton, Robert K. (1968), *Social Theory and Social Structure*, New York, NY: Simon & Schuster.

——— (1973), *The Sociology of Science: Theoretical and Empirical Investigations*, Chicago, IL: University of Chicago Press.

Moreno, Jacob L. (1934), *Who Shall Survive?: A New Approach to the Problem of Human Interrelations*, Washington, DC: Nervous and Mental Disease Publishing Co.

Morrison, Donald G. (1969), "On the Interpretation of Discriminant Analysis," *Journal of Marketing Research*, 156–163.

Muchnik, Lev, Sinan Aral, and Sean J. Taylor (2013), "Social Influence Bias: A Randomized Experiment," *Science*, 341 (6146), 647–651.

Naik, Prasad A. and Chih-Ling Tsai (2004), "Isotonic Single-Index Model for High-Dimensional Database Marketing," *Computational Statistics & Data Analysis*, 47 (4), 775–790.

Netzer, Oded, Ronen Feldman, Jacob Goldenberg, and Moshe Fresko (2012), "Mine your Own Business: Market-Structure Surveillance Through Text Mining," *Marketing Science*, 31 (3), 521–543.

Newman, Mark (2010), *Networks: An Introduction*, New York, NY: Oxford University Press.

Nielsen, Jesper H., Stewart A. Shapiro, and Charlotte H. Mason (2010), "Emotionality and Semantic Onsets: Exploring Orienting Attention Responses in Advertising," *Journal of Marketing Research*, 47 (6), 1138–1150.

Peres, Renana, Eitan Muller, and Vijay Mahajan (2010), "Innovation Diffusion and New Product Growth Models: A Critical Review and Research Directions," *International Journal of Research in Marketing*, 27 (2), 91–106.

Peterson, Wesley W., Theodore G. Birdsall, and We Fox (1954), "The Theory of Signal Detectability," *Transactions of the IRE Professional Group on Information Theory*, 4 (4), 171–212.

Pierce, David (2016), "SoundCloud Go: An Audacious Answer to Spotify That's Dying to Stand Out," (accessed August 10, 2016), [available at

http://www.wired.com/2016/03/soundclouds-new-venture-mixes-social-network-music-service/].

Pinker, Edieal J. (2007), "An Analysis of Short-Term Responses to Threats of Terrorism," *Management Science*, 53 (6), 865–880.

Podolny, Joel M. (2001), "Networks as the Pipes and Prisms of the Market," *American Journal of Sociology*, 107 (1), 33–60.

——— (2005), *Status Signals: A Sociological Study of Market Competition*, Princeton, NJ: Princeton University Press.

Putnam, Robert (2001), "Social Capital: Measurement and Consequences," *Canadian Journal of Policy Research*, 2 (1), 41–51.

Putnam, Robert D. (2000), *Bowling Alone: The Collapse and Revival of American Community*, New York, NY: Simon & Schuster.

Rivera, Mark T., Sara B. Soderstrom, and Brian Uzzi (2010), "Dynamics of Dyads in Social Networks: Assortative, Relational, and Proximity Mechanisms," *Annual Review of Sociology*, 36, 91–115.

Rogers, Everett M. (1995), *Diffusion of Innovations*, New York, NY: Free Press.

Rosen, Emanuel (2010), *Buzz: Real-Life Lessons in Word-of-Mouth Marketing*, London, UK: Profile Books.

Saboo, Alok R., V. Kumar, and Girish Ramani (2015), "Evaluating the Impact of Social Media Activities on Human Brand Sales," *International Journal of Research in Marketing*, 33 (3), 524–541.

Sauder, Michael, Freda Lynn, and Joel M. Podolny (2012), "Status: Insights from Organizational Sociology," *Annual Review of Sociology*, 38, 267–283.

Schneider, Joan and Julie Hall (2011), "Why Most Product Launches Fail," (accessed November 29, 2016), [available at https://hbr.org/2011/04/why-most-product-launches-fail].

Schonhoff, Thomas A. and Arthur Anthony Giordano (2006), *Detection and Estimation Theory and its Applications*, New Jersey, NJ: Pearson Education.

Scott, Alastair J. and Chris J. Wild (1997), "Fitting Regression Models to Case-Control Cata by Maximum Likelihood," *Biometrika*, 84 (1), 57–71.

Shaw, Marvin E. (1954), "Group Structure and the Behavior of Individuals in Small Groups," *The Journal of Psychology*, 38 (1), 139–149.

Singh, Jasjit (2005), "Collaborative Networks as Determinants of Knowledge Diffusion Patterns," *Management Science*, 51 (5), 756–770.

Skiera, Bernd (2016), "Data, Data and Even More Data: Harvesting Insights From the Data Jungle," *GfK Marketing Intelligence Review*, 8 (2), 10–17.

Snyder, C. Richard, Mary Anne Lassegard, and Carol E. Ford (1986), "Distancing After Group Success and Failure: Basking in Reflected Glory and Cutting Offreflected Failure," *Journal of Personality and Social Psychology*, 51 (2), 382–388.

SoundCloud (2016), "Community Guidelines," (accessed August 10, 2016), [available at https://soundcloud.com/community-guidelines].

Stephen, Andrew T. and Olivier Toubia (2010), "Deriving Value from Social Commerce Networks," *Journal of Marketing Research*, 47 (2), 215–228.

Stuart, Toby E. (1998), "Network Positions and Propensities to Collaborate: An Investigation of Strategic Alliance Formation in a High-Technology Industry," *Administrative Science Quarterly*, 43 (3), 668–698.

Stuart, Toby E., Ha Hoang, and Ralph C. Hybels (1999), "Interorganizational Endorsements and the Performance of Entrepreneurial Ventures," *Administrative Science Quarterly*, 44 (2), 315–349.

Swets, John A. (1961), "Is There a Sensory Threshold?" *Science*, 134 (3473), 168–177.

Tajfel, Henri (1974), "Social Identity and Intergroup Behaviour," *Social Science Information*, 13 (2), 65–93.

——— (1975), "The Exit of Social Mobility and the Voice of Social Change," *Social Science Information*, 14 (2), 101–118.

Tajfel, Henri and John C. Turner (1979), "An Integrative Theory of Intergroup Conflict," in William G. Austin and Stephen Worchel, eds., *The Social Psychology of Intergroup Relations*, Monterey, CA: Brooks-Cole, 33–47.

——— (1986), "The Social Identity Theory of Intergroup Behaviour," in Douglas Brownlie, Mike Saren, Robin Wensley, and Richard Whittington, eds., *Psychology of Intergroup Relations*, Chicago, IL: Nelson–Hall, 7–24.

Tanner Jr., Wilson P. and John A. Swets (1954), "A Decision-Making Theory of Visual Detection." *Psychological Review*, 61 (6), 401–409.

Taylor, Lindsay Shaw, Andrew T. Fiore, G.A. Mendelsohn, and Coye Cheshire (2011), "Out of My League: A Real-World Test of the Matching Hypothesis," *Personality and Social Psychology Bulletin*, 37 (7), 942–954.

Terry, Deborah J., Craig J. Carey, and Victor J. Callan (2001), "Employee Adjustment to an Organizational Merger: An Intergroup Perspective," *Personality and Social Psychology Bulletin*, 27 (3), 267–280.

Thomson, Matthew (2006), "Human Brands: Investigating Antecedents to Consumers' Strong Attachments to Celebrities," *Journal of Marketing*, 70 (3), 104–119.

Todd, Peter M., Lars Penke, Barbara Fasolo, and Alison P. Lenton (2007), "Different Cognitive Processes Underlie Human Mate Choices and Mate Preferences," *Proceedings of the National Academy of Sciences*, 104 (38), 15011–15016.

Trusov, Michael, Anand V. Bodapati, and Randolph E. Bucklin (2010), "Determining Influential Users in Internet Social Networks," *Journal of Marketing Research*, 47 (4), 643–658.

Trusov, Michael, Randolph E. Bucklin, and Keon Pauwels (2009), "Effects of Word-of-Mouth Versus Traditional Marketing: Findings from and Internet Social Networking Site," *Journal of Marketing Research*, 73 (5), 90–102.

Valente, Thomas W. (1995), *Network Models of the Diffusion of Innovations*, Cresskill, NJ: Hampton Press.

Van den Bulte, Christophe and Yogesh V. Joshi (2007), "New Product Diffusion with Influentials and Imitators," *Marketing Science*, 26 (3), 400–421.

Van den Bulte, Christophe and Stefan Wuyts (2007), *Social Networks and Marketing*, Cambridge, MA: Marketing Science Institute.

Van Everdingen, Yvonne, Dennis Fok, and Stefan Stremersch (2009), "Modeling Global Spillover of New Product Takeoff," *Journal of Marketing Research*, 46 (5), 637–652.

Van Knippenberg, Ad and Naomi Ellemers (1993), "Strategies In Intergroup Relations," in Michael A. Hogg and Dominic Ed Abrams, eds., *Group Motivation: Social Psychological Perspectives*, London, UK: Harvester Wheatsheaf, 17–24.

Von Neumann, John and Oskar Morgenstern (1947), *Theory of Games and Economic Behavior*, Princeton, NJ: Princeton University Press.

Voogt, Budi (2015), "Creating and Maintaining Momentum on SoundCloud," (accessed August 10, 2016), [available at `http://www.digitalmusicnews.com/2015/04/20/creating-and-maintaining-momentum-on-soundcloud/`].

Walker, Dylan and Lev Muchnik (2014), "Design of Randomized Experiments in Networks," *Proceedings of the IEEE*, 102 (12), 1940–1951.

Walster, Elaine, Vera Aronson, Darcy Abrahams, and Leon Rottman (1966), "Importance of Physical Attractiveness in Dating Behavior," *Journal of Personality and Social Psychology*, 4 (5), 508.

Wasserman, Stanley and Katherine Faust (1994), *Social Network Analysis: Methods and Applications*, Cambridge, MA: Cambridge University Press.

Watts, Duncan J. (2002), "A Simple Model of Global Cascades on Random Networks," *Proceedings of the National Academy of Sciences*, 99 (9), 5766–5771.

Watts, Duncan J. and Peter Sheridan Dodds (2007), "Influentials, Networks, and Public Opinion Formation," *Journal of Consumer Research*, 34 (4), 441–458.

Watts, Duncan J. and Steven H. Strogatz (1998), "Collective Dynamics of 'Small-World' Networks," *Nature*, 393 (6684), 440–442.

Weimann, Gabriel (1991), "The Influentials: Back to the Concept of Opinion Leaders?" *Public Opinion Quarterly*, 55 (2), 267–279.

Wellman, Barry and Stephen D. Berkowitz (1988), *Social Structures: A Network Approach*, vol. 2, New York, NY: Cambridge University Press.

Wickens, Thomas D. (2001), *Elementary Signal Detection Theory*, New York, NY: Oxford University Press.

Wixted, John T. (2007), "Dual-Process Theory and Signal-Detection Theory of Recognition Memory." *Psychological Review*, 114 (1), 152–176.

Yoganarasimhan, Hema (2012), "Impact of Social Network Structure on Content Propagation: A Study Using YouTube Data," *Quantitative Marketing and Economics*, 10 (1), 111–150.

Yonelinas, Andrew P. and Colleen M. Parks (2007), "Receiver Operating Characteristics (ROCs) in Recognition Memory: A Review." *Psychological Bulletin*, 133 (5), 800–832.

Tables and Figures

Table 1: Descriptive Statistics

		Sample 2009-2014[1]			Sample 2012-2015		
Descriptives		Mean	Median	Std	Mean	Median	Std
Indegree	Aug. 2011	134.77	19.00	532.88	-	-	-
	Mar. 2014	1254.81	59.00	31730.52	53.07	5.00	12.10
	Jun. 2015	-	-	-	20.01	8.00	102.10
Follows	sent	204.94	56.00	381.10	35.21	10.00	91.57
	received	1254.81	59.00	31730.52	20.01	8.00	102.10
Song Comments	sent	69.12	12.00	290.67	5.52	2.00	19.84
	received	132.31	14.00	711.46	12.33	3.00	97.49
Song Likes	sent	75.28	13.00	274.63	32.74	4.00	101.84
	received	552.34	20.00	7108.54	89.07	6.00	1014.00
Messages	sent	65.34	9.00	767.03	10.00	2.00	67.55
	received	54.24	9.00	149.68	5.25	1.00	68.72
Song Plays	sent	1703.26	390.00	3876.24	771.25	37.00	2885.87
	received	14378.87	380.00	222609.50	4785.35	176.00	59770.22
Song Reposts	sent	-	-	-	0.04	0.00	0.70
	received	-	-	-	0.02	0.00	1.17
Tracks	uploaded	31.12	9.00	108.10	10.13	3.00	37.85
Weekly Follows	sent	0.55	0.00	10.18	0.19	0.00	4.37
	received	4.12	0.00	362.48	0.10	0.00	1.13
Weekly Song Comments	sent	0.14	0.00	2.00	0.01	0.00	0.28
	received	0.32	0.00	4.84	0.01	0.00	0.32
Weekly Song Likes	sent	0.17	0.00	2.01	0.12	0.00	1.28
	received	1.50	0.00	49.93	0.07	0.00	3.19
Weekly Messages	sent	0.13	0.00	8.36	0.003	0.00	0.21
	received	0.15	0.00	1.01	0.004	0.00	0.25
Weekly Song Plays	sent	5.68	0.00	84.37	5.20	0.00	34.13
	received	50.29	0.00	1658.29	5.98	0.00	266.50
Weekly Song Reposts	sent	-	-	-	0.003	0.00	0.21
	received	-	-	-	0.004	0.00	0.25
Weekly Tracks	uploaded	0.12	0.00	1.66	0.01	0.00	0.35

[1] *We consider only the 24,020 creators of music and omit the 11,936 non-creators that signed up in the first quarter 2009.*

Table 2: Response Probabilities

	Seeding Target Types			
Creator Types	Type 1	Type 2	Type 3	Type 4
Type 1 Seeding Target	7.41%	3.31%	0.37%	0.03%
Type 2 Seeding Target	8.61%	4.97%	0.86%	0.05%
Type 3 Seeding Target	9.07%	7.30%	2.11%	0.22%
Type 4 Seeding Target	15.03%	16.46%	5.86%	0.75%

N = 4,964,174 follows, messages, song comments, and song likes of 18,005 creators of music over 1,959 days / Reaction period = 1 week

Table 3: Song Repost Probabilities

	Seeding Target Types			
Creator Types	Type 1	Type 2	Type 3	Type 4
Type 1 Seeding Target	0.109%	0.050%	0.007%	0.001%
Type 2 Seeding Target	0.174%	0.075%	0.025%	0.014%
Type 3 Seeding Target	0.457%	0.152%	0.078%	0.022%
Type 4 Seeding Target	0.372%	0.554%	0.123%	0.023%

N = 1,377,838 follows, messages, song comments, and song likes of 13,469 creators of music over 498 days / Reaction period = 1 week

Table 4: Expected Indirect Returns (and Standard Deviations) Given a Song Repost

	Follows	**Plays**
Type 1 Creator Reposted by		
Type 1 Seeding Target	0.01 (0.13)	0.48 (1.18)
Type 2 Seeding Target	0.41 (1.67)	6.51 (10.95)
Type 3 Seeding Target	2.22 (4.15)	43.19 (70.94)
Type 4 Seeding Target	7.60 (9.82)	281.60 (335.03)
Type 2 Creator Reposted by		
Type 1 Seeding Target	0.03 (0.19)	0.75 (1.55)
Type 2 Seeding Target	0.45 (1.14)	7.87 (12.68)
Type 3 Seeding Target	2.99 (5.95)	62.10 (112.72)
Type 4 Seeding Target	17.52 (26.94)	779.03 (1104.14)
Type 3 Creator Reposted by		
Type 1 Seeding Target	0.03 (0.20)	0.80 (1.55)
Type 2 Seeding Target	0.34 (0.91)	7.09 (12.96)
Type 3 Seeding Target	4.24 (10.32)	100.42 (194.90)
Type 4 Seeding Target	32.84 (42.81)	1144.18 (1719.02)
Type 4 Creator Reposted by		
Type 1 Seeding Target	0.03 (0.19)	1.13 (1.95)
Type 2 Seeding Target	0.26 (0.70)	7.66 (13.32)
Type 3 Seeding Target	3.15 (6.44)	113.13 (237.84)
Type 4 Seeding Target	55.86 (67.68)	2397.01 (3063.42)

N = 1,501,051 reposts of songs from 9,402 creators of music over 424 days / Reaction period = 1 week

Table 5: Expected Total Returns (and Standard Deviations)

	Seeding Target Types			
Creator Types	Type 1	Type 2	Type 3	Type 4
Type 1 Seeding Target	.0741 (.2619)	.0333 (.1828)	.0039 (.0721)	.0003 (.0446)
Type 2 Seeding Target	.0862 (.2807)	.0500 (.2199)	.0093 (.1401)	.0029 (.3757)
Type 3 Seeding Target	.0909 (.2876)	.0735 (.2629)	.0244 (.3421)	.0095 (.8038)
Type 4 Seeding Target	.1505 (.3576)	.1660 (.3749)	.0625 (.3443)	.0205 (.3390)

N = 4,964,174 follows, messages, song comments, and song likes of 18,005 creators of music over 1,959 days / Reaction period = 1 week

Table 6: Predictive Power of Alternative Model Specifications

			Social Capital Measures			
	t	**Naïve Model**	First-Degree Followers	Second-Degree Followers	First-Degree Clustering	Reciprocity
Average $PD(S)$	4	4.6%	3.7%	5.0%	6.4%	10.2%
	8	9.6%	10.1%	10.8%	13.0%	14.0%
	12	18.7%	18.2%	18.7%	19.7%	19.8%
Average $PF(S)$	4	87.3%	89.6%	86.3%	82.4%	72.1%
	8	73.6%	72.2%	70.5%	64.5%	61.7%
	12	49.1%	50.3%	49.2%	46.3%	45.9%

Table 7: Predictive Power of Benchmark Models

				Benchmark Models		
	t	**Naïve Model**	**Ordered Logit**	Logit	Firth	ReLogit
Average $PD(S)$	4	4.6%	10.2%	–	–	–
	8	9.6%	14.0%	16.1%	16.7%	7.5%
	12	18.7%	19.8%	25.4%	25.3%	7.1%
Average $PF(S)$	4	87.3%	72.1%	–	–	–
	8	73.6%	61.7%	77.4%	76.5%	89.8%
	12	49.1%	45.9%	61.5%	61.3%	90.2%

Figure 1: A Priori Response Probabilities

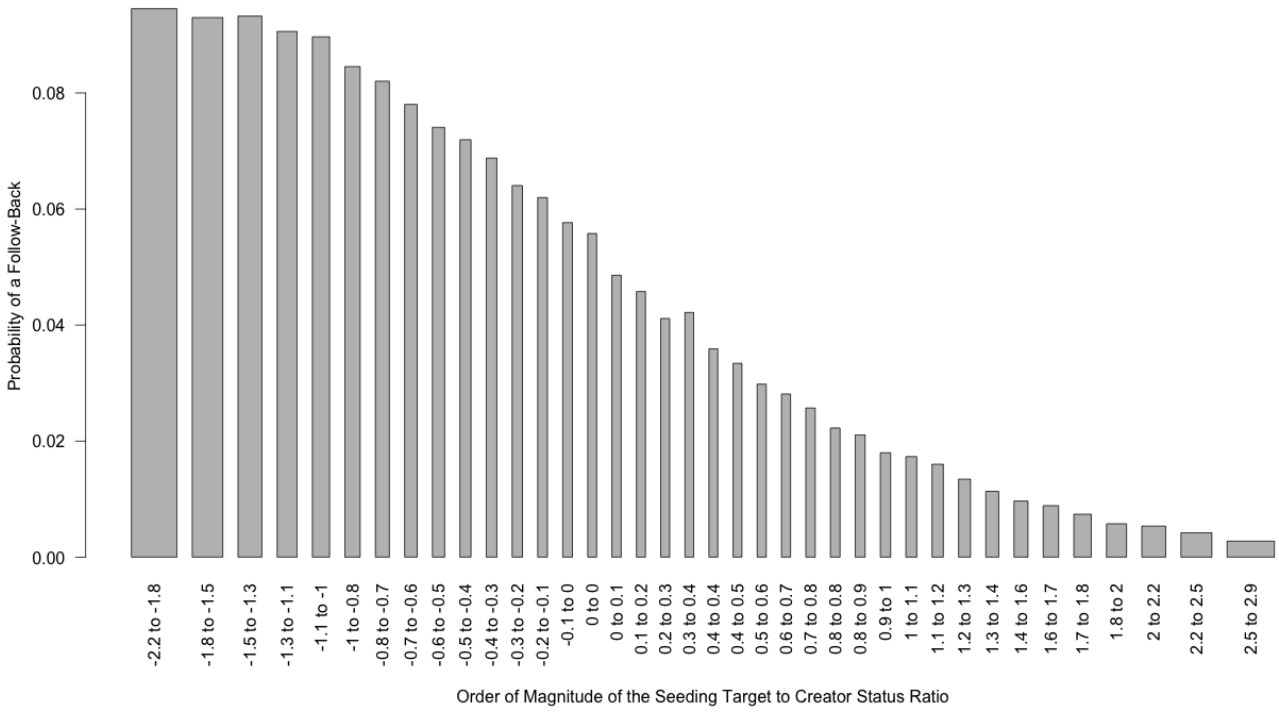


Figure 2: A Priori Response Probabilities: Low- and High-Status Creators

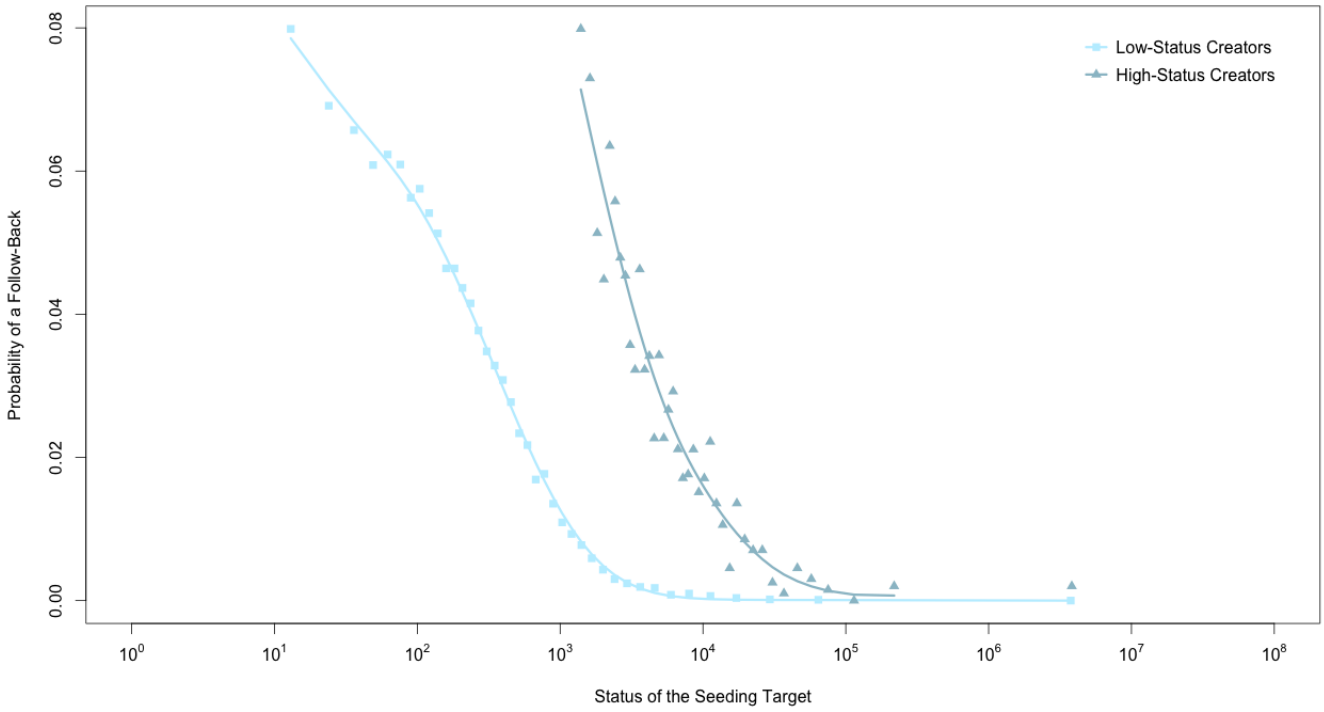


Figure 3: Portfolios of Creators

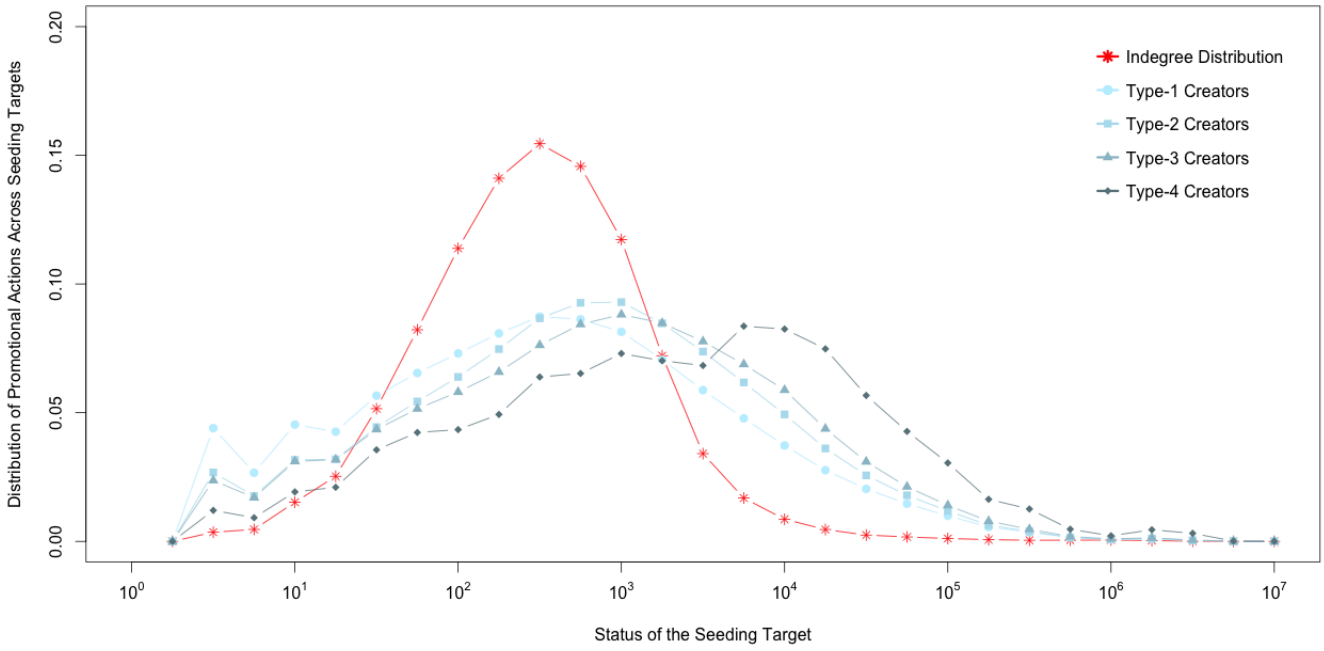


Figure 4: Portfolio of Type-1 Creators as a Function of the Budget

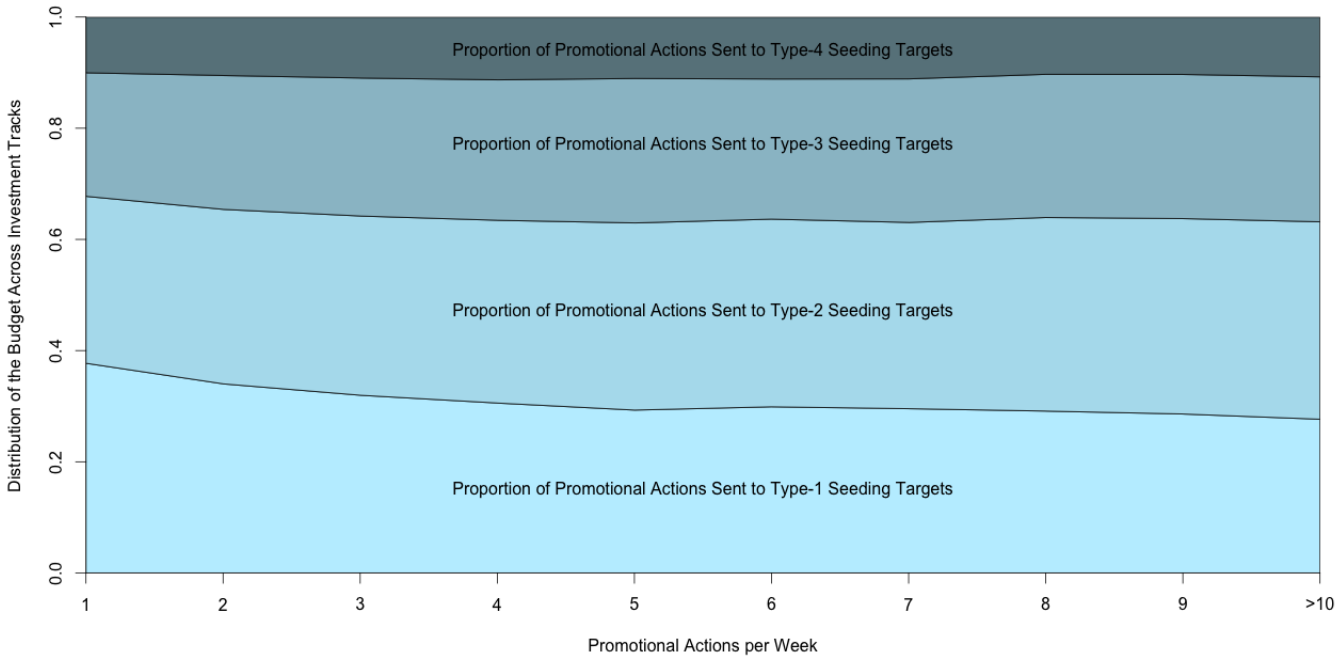


Figure 5: The Median Growth of the Follower Base: A Comparison of Three Seeding Policies

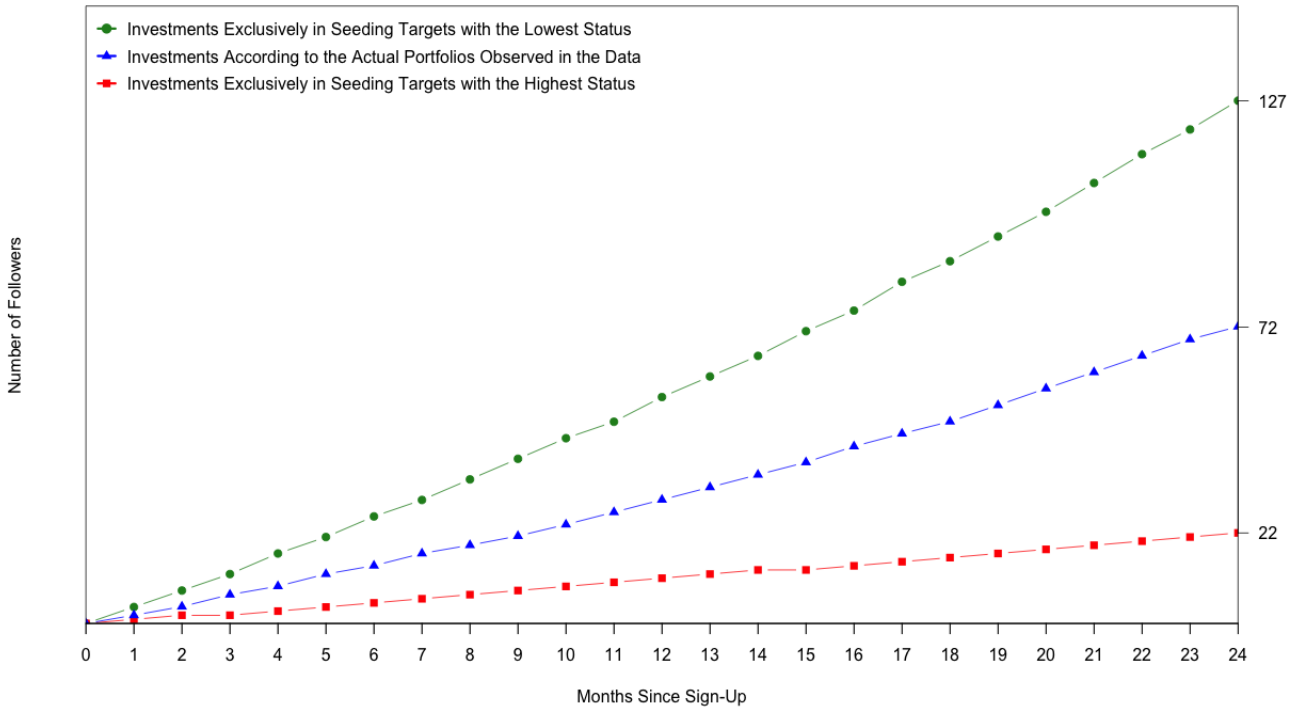


Figure 6: Probability of Detection and False Discovery as a Function of the Selection Size Four (Panels A and B), Eight (Panels C and D), and Twelve Weeks (Panels E and F) After Sign-Up

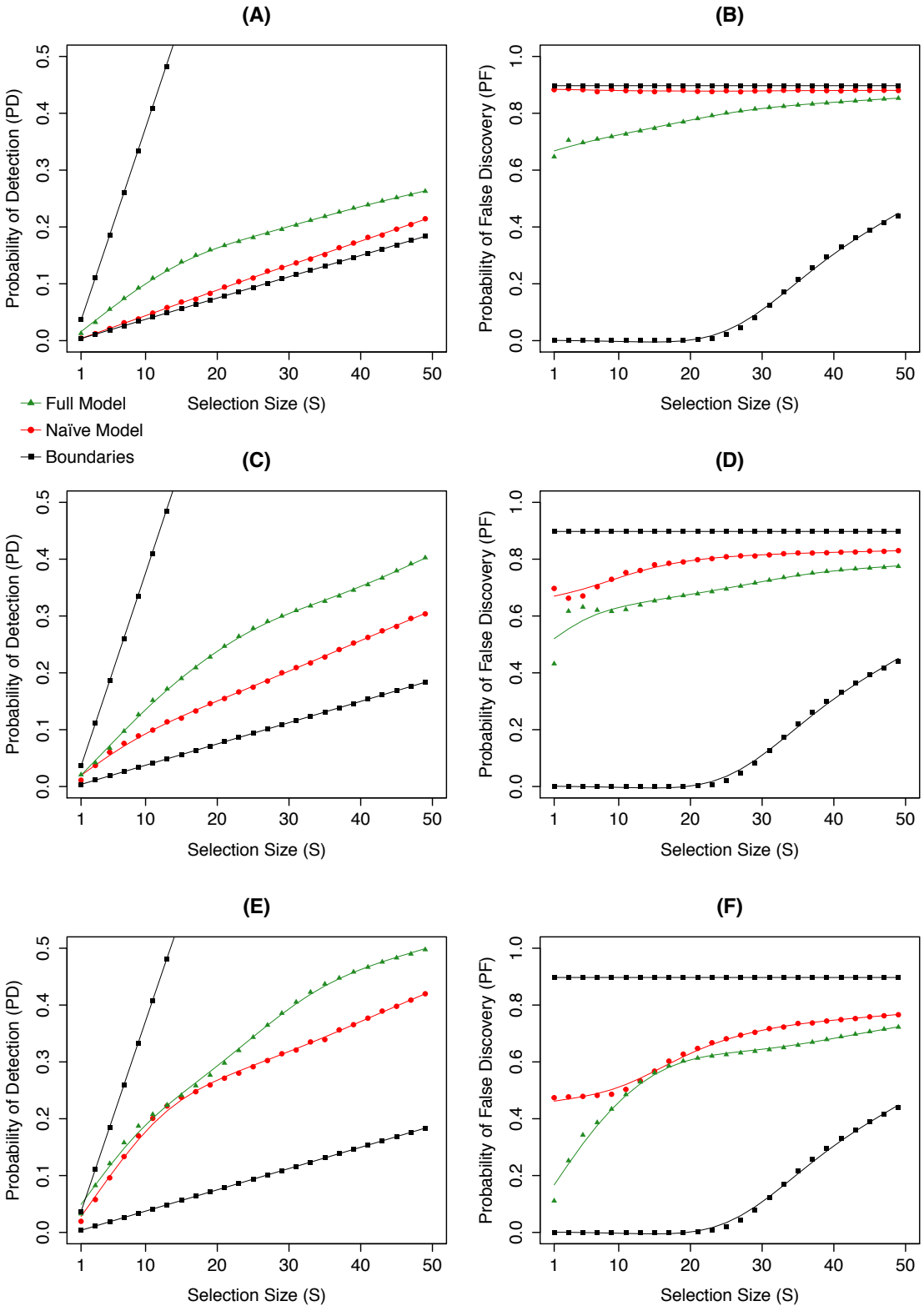


Figure 7: Profit Realization as a Function of the Return on Investment Four (Panel A), Eight (Panel B), and Twelve Weeks (Panel C) After Sign-Up

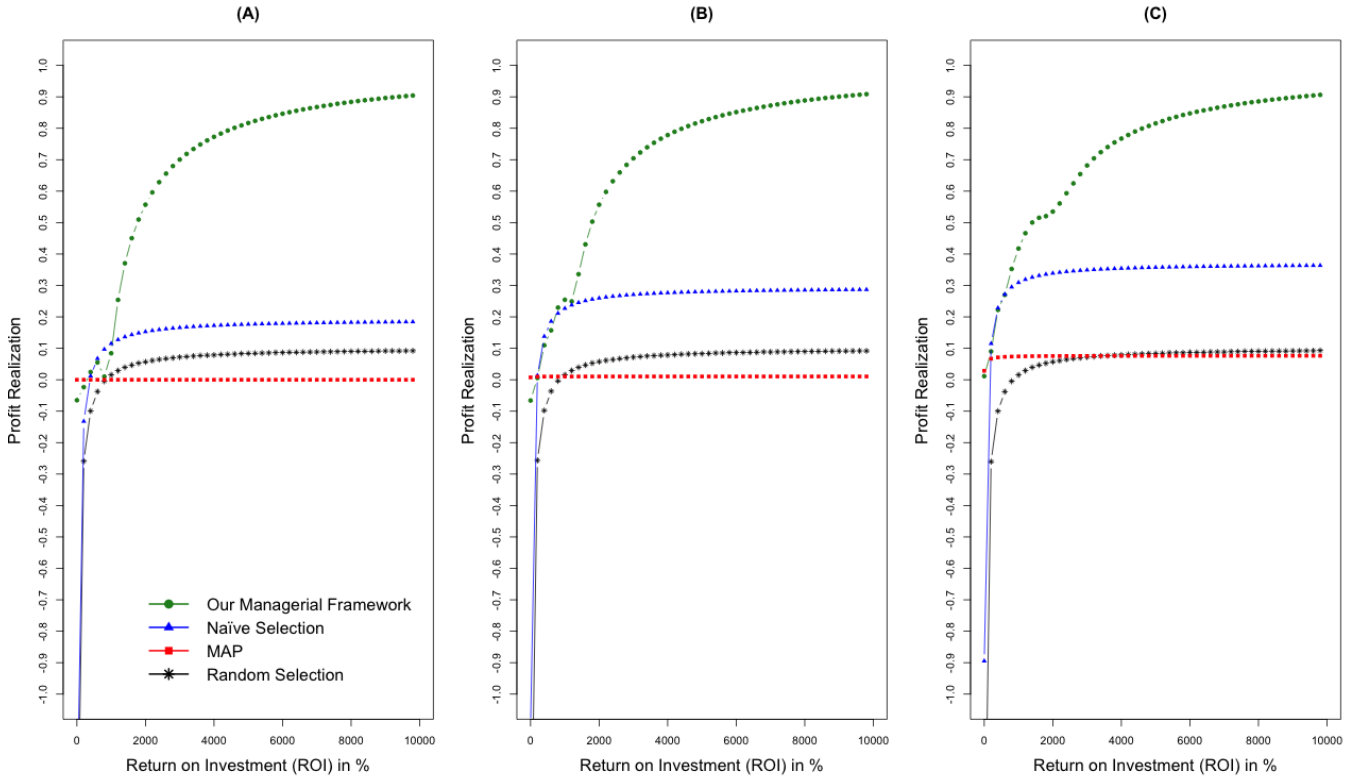
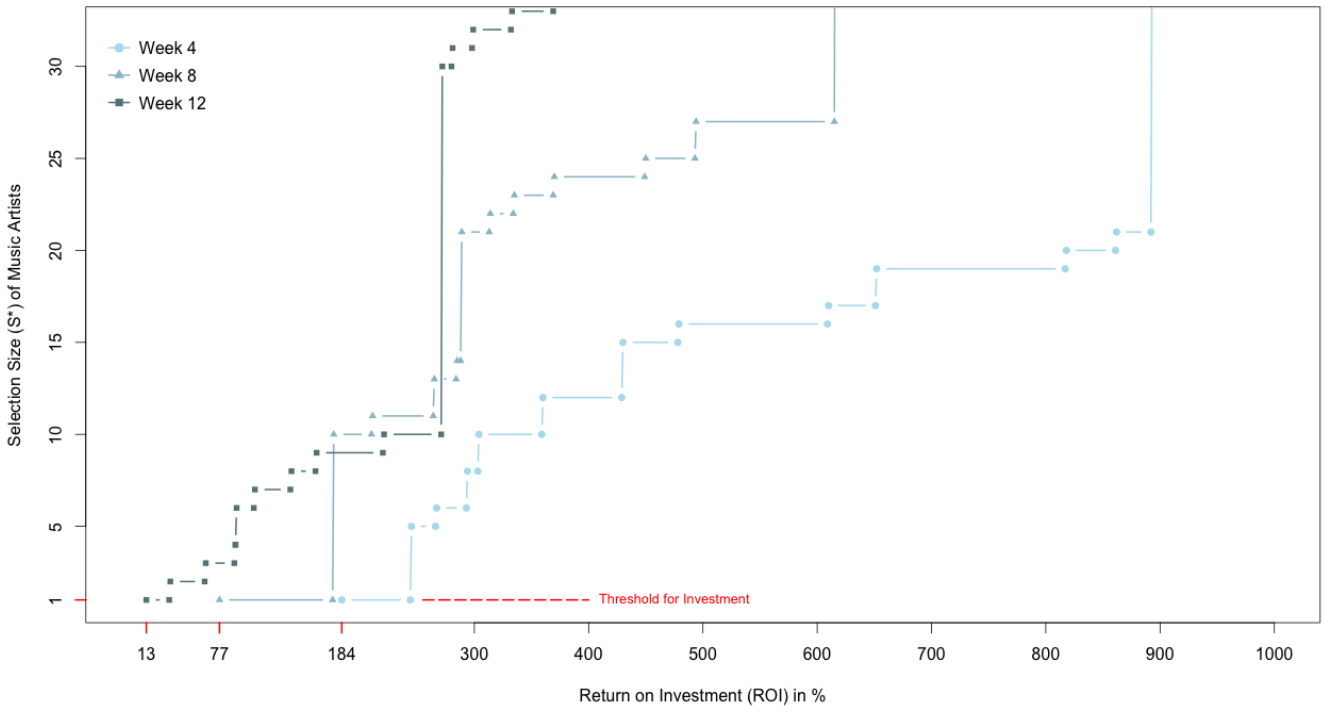


Figure 8: Optimal Selection Size of Music Artists as a Function of the Return on Investment



Appendix

A.1 Proofs of Propositions 1 – 6

Proof of Proposition 1. We follow the expected utility given by (3), i.e.,

$$EU = \sum_{x=0}^{X} \sum_{y=0}^{Y} \binom{X}{x} \binom{Y}{y} p_L^x \, p_H^y \, (1-p_L)^{X-x} \, (1-p_H)^{Y-y} \; U(xL+yH) \, ,$$

and the optimization problem given by (10), i.e.,

$$\underset{X,Y}{\text{Max}} \; EU(X,Y) \quad \text{s.t.} \quad X+Y=B \, .$$

Since a risk neutral creator does not differentiate between the expected utility and the expected return, i.e., $U(xL+yH) = xL+yH$, the optimization problem is as a result given by

$$\begin{aligned} \underset{X,Y}{\text{Max}} \; EU = & \left(\sum_{x=0}^{X} \binom{X}{x} p_L^x \, (1-p_L)^{X-x} \, x \right) \left(\sum_{y=0}^{Y} \binom{Y}{y} p_H^y \, (1-p_H)^{Y-y} \right) L \\ & + \left(\sum_{x=0}^{X} \binom{X}{x} p_L^x \, (1-p_L)^{X-x} \right) \left(\sum_{y=0}^{Y} \binom{Y}{y} p_H^y \, (1-p_H)^{Y-y} \, y \right) H \qquad \text{(A.1)} \\ = & \, Xp_L + Yp_H \, . \qquad \text{(A.2)} \end{aligned}$$

Resolving the above optimization problem results in

$$\frac{\partial EU}{\partial X} = p_L L \, , \qquad \text{(A.3)}$$

$$\frac{\partial EU}{\partial Y} = p_H H \, . \qquad \text{(A.4)}$$

Therefore, the optimal solution is a corner solution, namely if $p_H H > p_L L$, then $\frac{\partial EU}{\partial Y} > \frac{\partial EU}{\partial X}$, hence $Y^* = B$ and $X^* = 0$; otherwise, if $p_H H < p_L L$, then $\frac{\partial EU}{\partial Y} < \frac{\partial EU}{\partial X}$, hence $Y^* = 0$ and $X^* = B$. We conclude that if a creator is risk neutral, then the whole budget is invested in high-status individuals in case $p_H H > p_L L$; otherwise (if $p_H H < p_L L$) the whole budget is invested in low-status individuals. □

Proof of Proposition 2. Since $p_L = p_H = p$, $B = X + Y$, and $z = x + y$ for each pair (x, y), we can rewrite the expected utility given by (3) as follows:

$$EU(X,Y) = \sum_{x=0}^{X} \sum_{y=0}^{Y} \binom{X}{x} \binom{Y}{y} p^z \left(1-p\right)^{B-z} U(xL + yH) \,, \tag{A.5}$$

where the optimization problem is given by (10), i.e.,

$$\underset{X,Y}{\text{Max}}\, EU(X,Y) \quad \text{s.t.} \quad X + Y = B \,. \tag{A.6}$$

To contrast the above case where the creator invests a fraction of her budget in low-status individuals and a fraction of her budget in high-status individuals, i.e., $X \neq 0$ an $Y \neq 0$, we now look at the case where the whole budget is invested in high-status individuals, i.e., $X = 0$ and $Y = B$. Then the expected utility is given by

$$EU_H = \sum_{z=0}^{B} \binom{B}{z} p^z \left(1-p\right)^{B-z} \; U(zH) \,, \tag{A.7}$$

where z is the number of cases, which yield a return H when investing B promotional actions in high-status individuals. Since $z = x + y$, $0 \leq z \leq B$, and $0 \leq x \leq z$, the expected utility can be rewritten as

$$EU_H = \sum_{z=0}^{B} \sum_{x=0}^{z} \binom{X}{x} \binom{Y}{y} p^z (1-p)^{B-z} U(zH) \,. \tag{A.8}$$

As $y = z - x$ and $B = X + Y$, according to Vandermonde's Convolution

$$\sum_{x=0}^{z} \binom{X}{x} \binom{Y}{y} = \sum_{x=0}^{z} \binom{X}{x} \binom{Y}{z-x} = \binom{X+Y}{z} = \binom{B}{z} \,. \tag{A.9}$$

Moreover, because the return on low-status individuals is lower than on high-status individuals, i.e., $L < H$, it follows that

$$U(xL + yH) \leq U(zH) \,, \tag{A.10}$$

Then, based on (A.5), (A.8), and (A.10), we conclude that the expected utility is lower when investing in both high- and low-status individuals, i.e.,

$$EU \leq EU_H. \tag{A.11}$$

To generalize (A.11), we look again at a creator who gets in x cases a return L when investing X promotional actions in low-status individuals and in y cases a return H when investing Y promotional actions in high-status individuals. In this context, the creator gains return $R(x, y)$ from her investment, i.e., $X \neq 0$ an $Y \neq 0$. However, if the creator directs promotional actions only to high-status individuals, i.e., $X = 0$ and $Y = B$, then she gains return $R(0, z)$ from her investment, where $z = x + y$. For any return r,

$$Prob(R(x, y) \leq r) \geq Prob(R(0, x + y) \leq r)\,, \tag{A.12}$$

where $Prob(R(x, y) \leq r)$ is the probability to get a return R less than r, given that there are x and y cases that yield a return from low- and high-status individuals, respectively. $Prob(R(0, z) \leq r)$ is the probability to get a return R less than r, given that there are no cases that yield a return from low-status individuals and z cases that yield a return from high-status individuals. Following Hadar and Russell (1969), $Prob(R(0, x + y) \leq r)$ has second-order stochastic dominance over $Prob(R(x, y) \leq r)$. Hence, for all nondecreasing and concave utility functions, the expected utility is lower when investing in both high- and low-status individuals, i.e.,

$$EU \leq EU_H\,. \tag{A.13}$$

Thus, we conclude that if the response probability of high-status individuals equals the response probability of low-status individuals, i.e., $p = p_L = p_H$, then the whole budget is invested in high-status individuals, i.e., $Y^* = B$. □

Proof of Proposition 3. We follow the expected utility given by (3), i.e.,

$$EU = \sum_{x=0}^{X} \sum_{y=0}^{Y} \binom{X}{x} \binom{Y}{y} p_L^x\, p_H^y\, (1 - p_L)^{X-x}\, (1 - p_H)^{Y-y}\; U(xL + yH)\,,$$

and the optimization problem given by (10), i.e.,

$$\underset{X,Y}{\text{Max}}\, EU(X,Y) \quad \text{s.t.} \quad X+Y=B\,.$$

Let us define K such that

$$K \gg max(B, \frac{1}{p_L \underset{0\leq x\leq B}{min}(\frac{U(xL+L)}{U(xL+H)})})\,, \tag{A.14}$$

where $U(0)=0$. Thus, for any $p_H < \frac{1}{K}$ it holds that

$$p_H \ll \frac{1}{B}\,, \tag{A.15}$$

$$p_H \ll p_L \underset{0\leq x\leq B}{min}(\frac{U(xL+L)}{U(xL+H)})\,. \tag{A.16}$$

Since $p_H \ll \frac{1}{B}$, an investment in high-status individuals behaves like a Bernoulli coin with probability Yp_H. Thus, the expected utility given by (3) has the following form:

$$\begin{aligned} EU =& \sum_{x=0}^{X} \binom{X}{x} p_L^x \left(1-p_L\right)^{X-x} \\ & \left\{ \binom{Y}{0} p_H^0 (1-Yp_H) U(xL) + \binom{Y}{1} p_H U(xL+H) + 0(p_H^2) \right\} \qquad \text{(A.17)} \\ =& \sum_{x=0}^{X} \binom{X}{x} p_L^x \left(1-p_L\right)^{X-x} \left((1-Yp_H)U(xL) + Yp_H U(xL+H)\right)\,. \qquad \text{(A.18)} \end{aligned}$$

To contrast the above case where the creator invests a fraction of her budget in low-status individuals and a fraction of her budget in high-status individuals, i.e., $X \neq 0$ and $Y \neq 0$, we now look at the case where the whole budget is invested in high-status individuals, i.e., $X = B$ and $Y = 0$. By applying Vandermonde's Convolution, the expected utility can be

rewritten as

$$EU_L = \sum_{x=0}^{B} \binom{B}{x} p_L^x \left(1-p_L\right)^{B-x} U(xL) \tag{A.19}$$

$$= \sum_{x=0}^{X} \binom{X}{x} p_L^x \left(1-p_L\right)^{X-x} \sum_{y=0}^{Y} \binom{Y}{y} p_L^y \left(1-p_L\right)^{Y-y} U((x+y)L) \tag{A.20}$$

$$> \sum_{x=0}^{X} \binom{X}{x} p_L^x \left(1-p_L\right)^{X-x} \left(U(xL) + Y p_L U(xL+L)\right) . \tag{A.21}$$

As $p_H \ll p_L \min\limits_{0\leq x\leq B}(\frac{U(x+1)L}{U(xL+H)})$, it follows that

$$p_H U(xL+H) < p_L U(xL+L) , \tag{A.22}$$

and therefore

$$\sum_{x=0}^{X} \binom{X}{x} p_L^x \left(1-p_L\right)^{X-x} \left(U(xL) + Y p_L U(xL+L)\right)$$

$$\geq \sum_{x=0}^{X} \binom{X}{x} p_L^x \left(1-p_L\right)^{X-x} \left(U(xL) + Y p_H U(xL+L)\right) \tag{A.23}$$

$$\geq \sum_{x=0}^{X} \binom{X}{x} p_L^x \left(1-p_L\right)^{X-x} \left((1-Yp_H) U(xL) + Y p_H U(xL+L)\right) . \tag{A.24}$$

Thus, we conclude that if the response probability of high-status individuals is extremely low, i.e., $p_H \ll \frac{1}{B}$, then the whole budget is invested in low-status individuals, i.e., $X^* = B$, independent of the individual aversion to risk. □

Proof of Proposition 4. We follow the expected utility given by (3), i.e.,

$$EU = \sum_{x=0}^{X} \sum_{y=0}^{Y} \binom{X}{x} \binom{Y}{y} p_L^x \, p_H^y \left(1-p_L\right)^{X-x} \left(1-p_H\right)^{Y-y} \, U(xL+yH) ,$$

and the optimization problem given by (10), i.e.,

$$\underset{X,Y}{\text{Max}} \; EU(X,Y) \quad \text{s.t.} \quad X+Y = B .$$

Since we assume that the response probability of high-status individuals is extremely low, i.e., $p_H \ll \frac{1}{B}$, an investment in high-status individuals behaves like a Bernoulli coin with probability Yp_H. Thus, the expected utility given by (3) has the following form:

$$EU = \sum_{x=0}^{X} \binom{X}{x} p_L^x \left(1-p_L\right)^{X-x} \left((1-Yp_H)U(xL) + Yp_H U(xL+H)\right) . \tag{A.25}$$

Let us assume that the return on low-status individuals is much lower than on high-status individuals such that $BL \ll H$. Since $\sum_{x=0}^{X} \binom{X}{x} p_L^x \left(1-p_L\right)^{X-x} = 1$ and $U(xL+H) \approx U(H)$, the expected utility is thus given by

$$EU = V(X, p_L) + Yp_H(U(H) - V(X|p_L)) , \tag{A.26}$$

where $V(X, p_L) = \sum_{x=0}^{X} \binom{X}{x} p_L^x \left(1-p_L\right)^{X-x} U(xL)$. Naturally, an increase in the status of the creator leads to an increase in the expected utility. Along these lines, we aim to show that the order of magnitude of the change in the expected utility due to an increase in the response probability of low-status individuals, i.e., $O(\Delta EU_{p_L})$, is lower than the order of magnitude of the change in the expected utility due to an increase in the response probability of high-status individuals, i.e., $O(\Delta EU_{p_H})$. We assume that the return on low-status individuals, i.e., L, is not changing much as it is typically close to 1 since the indirect return is marginal. The changes in the expected utility, i.e., ΔEU_{p_L} and ΔEU_{p_H}, have the following forms:

$$\Delta EU_{p_L} = \Delta V(X, p_L)(1 - Yp_H) , \tag{A.27}$$

$$\Delta EU_{p_H} = Yp_H(U(H) - V(X, p_L))\frac{\Delta p_H}{p_H} . \tag{A.28}$$

Let us look at the effect of an increase in the response probability of low-status individuals, i.e.,

$$p_L \to p_L + \Delta p_L , \tag{A.29}$$

$$V(X, p_L) \to V(X, p_L + \Delta p_L) . \tag{A.30}$$

We define $q_L = 1 - p_L$ and $\Delta q_L = -\Delta p_L$. Since $\frac{\Delta p_L}{p_L} \ll 1$ and further $(1+\delta)^X \approx 1 + X\delta$ as

well as $(1+\delta_1)(1+\delta_2) \approx 1+\delta_1+\delta_2$ if the δ's are small, then

$$V(X, p_L + \Delta p_L) = \sum_{x=0}^{X} \binom{X}{x} \left[p_L \left(1 + \frac{\Delta p_L}{p_L}\right)\right]^X \left[q_L \left(1 + \frac{\Delta q_L}{q_L}\right)\right]^{X-x} U(xL) \tag{A.31}$$

$$\approx \sum_{x=0}^{X} \binom{X}{x} p_L^x \left(1 + x\frac{\Delta p_L}{p_L}\right) q_L^{X-x} \left(1 + (X-x)\frac{\Delta q_L}{q_L}\right) U(xL) \tag{A.32}$$

$$\approx \sum_{x=0}^{X} \binom{X}{x} p_L^x q_L^{X-x} U(xL) \left\{1 + \frac{x\Delta p_L}{p_L(1-p_L)} - X\frac{\Delta p_L}{1-p_L}\right\} \tag{A.33}$$

$$= V(X, p_L) + \frac{\Delta p_L}{p_L(1-p_L)} \sum_{x=0}^{X} \binom{X}{x} p_L^x q_L^{X-x} U(xL)(x - Xp_L)\,. \tag{A.34}$$

Recall that $\frac{\Delta p_L}{p_L} - \frac{\Delta q_L}{q_L} = \frac{\Delta p_L}{p_L} + \frac{\Delta p_L}{1-p_L} = \frac{\Delta p_L}{p_L(1-p_L)}$ and $\frac{\Delta q_L}{q_L} = -\frac{\Delta p_L}{1-p_L}$. To evaluate the order of magnitude of the change in the above component of the expected utility, i.e., $O(\Delta V(X, p_L))$, we approximate the binomial distribution $\binom{X}{x} p_L^x q_L^{X-x}$ by a rectangular, which is concentrated around the average, i.e., Xp_L, has a width of two standard deviations, i.e., 2σ, and is normalized:

$$O\left(\binom{X}{x} p_L^x q_L^{X-x}\right) \approx \begin{cases} \frac{1}{4\sigma} & -2\sigma < x < 2\sigma\,, \\ 0 & \text{otherwise}\,, \end{cases} \tag{A.35}$$

namely we rely on the fact that the distribution is centered around the average. It follows that

$$\sum_{x=0}^{X} \binom{X}{x} p_L^x q_L^{X-x} U(xL)(x - Xp_L) \tag{A.36}$$

$$= \sum_{x=Xp_L-2\sigma}^{Xp_L} \binom{X}{x} p_L^x q_L^{X-x} U(xL)(x - Xp_L) \tag{A.37}$$

$$+ \sum_{x=Xp_L}^{Xp_L+2\sigma} \binom{X}{x} p_L^x q_L^{X-x} U(xL)(x - Xp_L) \tag{A.38}$$

$$\approx 2\sigma \left\{ \frac{1}{4\sigma} U((Xp_L - \sigma)L)(-\sigma) \right\} + 2\sigma \left\{ \frac{1}{4\sigma} U((Xp_L + \sigma)L)\sigma \right\} \tag{A.39}$$

$$= \frac{\sigma}{2} \{ U((Xp_L + \sigma)L) - U((Xp_L - \sigma)L) \} \tag{A.40}$$

$$= \sigma^2 L \left\{ \frac{U((Xp_L + \sigma)L) - U((Xp_L - \sigma)L)}{2\sigma L} \right\} \approx \sigma^2 L \frac{\partial U}{\partial Z}\Big|_{Z=Xp_L L} . \tag{A.41}$$

Note that $x - Xp_L = \pm O(\sigma)$. Hence, note that if $x \leq Xp_L$, then $U(xL) = U((Xp_L - \sigma)L)$ and $(x - Xp_L) \approx -\sigma$. If $x \geq Xp_L$, then $U(xL) = U((Xp_L + \sigma)L)$ and $(x - Xp_L) \approx \sigma$. On the other hand, $V(X, p_L) = \sum_{x=0}^{X} \binom{X}{x} p_L^x q_L^{X-x} U(xL) \approx \sum_{x=Xp_L-2\sigma}^{Xp_L+2\sigma} \frac{1}{4\sigma} U(Xp_L L) = U(Xp_L L)$ and therefore $\frac{\partial V}{\partial X} \approx p_L L \frac{\partial U}{\partial Z}\Big|_{Z=Xp_L L}$. Thus,

$$O(\Delta V(X, p_L)) = \frac{\Delta p_L}{p_L(1 - p_L)} \sum_{x=0}^{X} \binom{X}{x} p_L^x q_L^{X-x} U(xL)(x - Xp_L) \tag{A.42}$$

$$= \frac{\Delta p_L}{p_L(1 - p_L)} \sigma^2 \frac{1}{p_L} \frac{\partial V}{\partial X} \tag{A.43}$$

$$= X \frac{\Delta p_L}{p_L} \frac{\partial V}{\partial X} , \tag{A.44}$$

where we assume that $O(\sigma^2) = Xp_L(1 - p_L)$ as the distribution is approximately binomial. Hence, the order of magnitude of the change in the expected utility due to an increase in the response probability of low-status individuals, i.e., $O(\Delta EU_{p_L})$, and the order of magnitude of the change in the expected utility due to an increase in the response probability of high-status individuals, i.e., $O(\Delta EU_{p_H})$, are given by

$$O(\Delta EU_{p_L}) = \Delta p_L X \frac{\partial V}{\partial X}(1 - Yp_H) , \tag{A.45}$$

$$O(\Delta EU_{p_H}) = Yp_H(U(H) - V(X, p_L)) \frac{\Delta p_H}{p_H} . \tag{A.46}$$

Following the expected utility given by (A.26), the marginal-rate-of-substitution-equation

has the following form:

$$\frac{\partial EU}{\partial X} = \frac{\partial EU}{\partial Y} \tag{A.47}$$

$$\Rightarrow \frac{\partial V}{\partial X}(1 - Yp_H) = p_H(U(H) - V(X, p_L)) \tag{A.48}$$

$$\Rightarrow O\left(\frac{\partial V}{\partial X}(1 - Yp_H)\right) = O(p_H(U(H) - V(X, p_L))) \equiv M \tag{A.49}$$

Therefore, the order of magnitude of the change in the expected utility due to an increase in the response probability of low-status individuals, i.e., $O(\Delta EU_{p_L})$, and the order of magnitude of the change in the expected utility due to an increase in the response probability of high-status individuals, i.e., $O(\Delta EU_{p_H})$, given by (A.45) and (A.46), respectively, can be rewritten as

$$O(\Delta EU_{p_L}) = X\frac{\Delta p_L}{p_L}\frac{\partial V}{\partial X}(1 - Yp_H) = XM\frac{\Delta p_L}{p_L}\,, \tag{A.50}$$

$$O(\Delta EU_{p_H}) = Yp_H(U(H) - V(X, p_L))\frac{\Delta p_H}{p_H} = YM\frac{\Delta p_H}{p_H}\,, \tag{A.51}$$

and since $\frac{\Delta p_L}{p_L} \ll \frac{\Delta p_H}{p_H}$, it follows that

$$O(\Delta EU_{P_L}) \ll O(\Delta EU_{P_H})\,. \tag{A.52}$$

As a result, we assume that an increase in the status of the creator leads to an increase in the response probability of high-status individuals, i.e., p_H, whereas the response probability of low-status individuals, i.e., p_L, stays approximately constant. In the data we observe that an increase in the status of the creator increases both p_L and p_H but $\frac{\Delta p_L}{p_L} \ll \frac{\Delta p_H}{p_H}$, e.g., if the status of the creator increases from type 1 to type 2, then the increase in the response probability of a type 1 seeding target, i.e., $\frac{\Delta p_L}{p_L}$, is 16% and the increase in the response probability of a type 3 seeding target, i.e., $\frac{\Delta p_L}{p_L}$, is 132%.

Lemma 1. *If the response probability of high-status individuals and low-status individuals, i.e., p_H and p_L, increases such that $\frac{\Delta p_L}{p_L} \ll \frac{\Delta p_H}{p_H}$, then the expected utility increases in the manner that the order of magnitude of the increase in p_L, i.e., $O(\Delta EU_{P_L})$, is lower than the order of magnitude of the increase in p_H, i.e., $O(\Delta EU_{p_H})$.*

Corollary 2. *Based on the finding that* $\frac{\Delta p_L}{p_L} \ll \frac{\Delta p_H}{p_H}$*, if the status of a creator increases, then we can assure that the response probability of high-status individuals, i.e.,* p_H*, increases whereas the response probability of low-status individuals, i.e.,* p_L*, stays approximately constant.*

Thus, we conclude that if the status of a creator, i.e., S, increases, then the more is invested in high-status individuals, i.e., $\frac{dX^*}{dS} < 0$ and $\frac{dY^*}{dS} > 0$. □

Proof of Proposition 5. We follow the expected utility given by (3), i.e.,

$$EU = \sum_{x=0}^{X} \sum_{y=0}^{Y} \binom{X}{x} \binom{Y}{y} p_L^x \, p_H^y \, (1-p_L)^{X-x} \, (1-p_H)^{Y-y} \; U(xL+yH) \, ,$$

and the optimization problem given by (10), i.e.,

$$\underset{X,Y}{\text{Max}} \; EU(X,Y) \quad \text{s.t.} \quad X+Y=B \, .$$

Since we assume that the response probability of high-status individuals is extremely low, i.e., $p_H \ll \frac{1}{B}$, an investment in high-status individuals behaves like a Bernoulli coin with probability Yp_H. Thus, the expected utility given by (3) has the following form:

$$\begin{aligned} EU =& \sum_{x=0}^{X} \binom{X}{x} p_L^x \, (1-p_L)^{X-x} \\ & \left\{ \binom{Y}{0} p_H^0 (1-Yp_H) U(xL) + \binom{Y}{1} p_H U(xL+H) + 0(p_H^2) \right\} \qquad \text{(A.53)} \\ =& \sum_{x=0}^{X} \binom{X}{x} p_L^x \, (1-p_L)^{X-x} \left((1-Yp_H) U(xL) + Yp_H U(xL+H) \right) \, . \qquad \text{(A.54)} \end{aligned}$$

Let us assume that the return on low-status individuals is much lower than on high-status individuals such that $BL \ll H$. Since $\sum_{x=0}^{X} \binom{X}{x} p_L^x \, (1-p_L)^{X-x} = 1$ and $U(xL+H) \approx U(H)$, then

$$EU = V(X|p_L) + Yp_H (U(H) - V(X|p_L)) \, , \qquad \text{(A.55)}$$

where $V(X|p_L) \equiv \sum_{x=0}^{X} \binom{X}{x} p_L^x \, (1-p_L)^{X-x} \, U(xL)$. The first-order condition is given by the marginal rate of substitution, i.e.,

$$\frac{\partial EU}{\partial X}\Big|_{X^*,Y^*} = \frac{\partial EU}{\partial Y}\Big|_{X^*,Y^*} \tag{A.56}$$

$$\Rightarrow \; \frac{\partial V}{\partial X}\Big|_{X^*}(1 - Y^* p_H) = p_H(U(H) - V(X^*|p_H)) \, . \tag{A.57}$$

Let us look at the effect of an increase in the budget, i.e., $B \to B + \Delta B$, and apply comparative statics. In case of an internal solution, (A.57) has the following form:

$$\begin{aligned} &\frac{\partial EU}{\partial X} + \frac{\partial^2 EU}{\partial X^2}\frac{dX^*}{dB}\Delta B + \frac{\partial^2 EU}{\partial X \partial Y}\frac{dY^*}{dB}\Delta B \\ =&\frac{\partial EU}{\partial Y} + \frac{\partial^2 EU}{\partial X \partial Y}\frac{dX^*}{dB}\Delta B + \frac{\partial^2 EU}{\partial Y^2}\frac{dY^*}{dB}\Delta B \end{aligned} \tag{A.58}$$

$$\Rightarrow \; \frac{\partial^2 EU}{\partial X^2}\frac{dX^*}{dB} + \frac{\partial^2 EU}{\partial X \partial Y}\frac{dY^2}{dB} = \frac{\partial^2 EU}{\partial X \partial Y}\frac{dX^*}{dB} + \frac{\partial^2 EU}{\partial Y^2}\frac{dY^*}{dB} \, , \tag{A.59}$$

and the budget constraint, i.e., $X + Y = B$, is given by

$$X^* + \frac{dX^*}{dB}\Delta B + Y^* + \frac{dY^*}{dB}\Delta B = B + \Delta B \tag{A.60}$$

$$\Rightarrow \; \frac{dX^*}{dB} + \frac{dY^*}{dB} = 1 \tag{A.61}$$

$$\Rightarrow \; p_H \frac{\partial V}{\partial X}\left(\frac{dX^*}{dB} + \frac{dY^*}{dB}\right) = p_H \frac{\partial V}{\partial X} \, . \tag{A.62}$$

On one hand, $\frac{\partial^2 EU}{\partial Y^2} = 0$ because

$$\frac{\partial EU}{\partial Y} = p_H(U(H) - V(X|p_L)) = \text{const. in } y \, , \tag{A.63}$$

and, on the other hand, $\frac{\partial^2 V}{\partial X \partial Y} < 0$ because

$$\frac{\partial^2 EU}{\partial X \partial Y} = \frac{\partial}{\partial Y}\left\{\frac{\partial EU}{\partial X}\right\} = \frac{\partial}{\partial Y}\left\{(1 - Y p_H)\frac{\partial V}{\partial X}\right\} = -p_H \frac{\partial V}{\partial X} < 0 \, . \tag{A.64}$$

From (A.55) we further find that

$$\frac{\partial^2 EU}{\partial X^2} = \frac{\partial^2 V}{\partial X^2}(1 - Y p_H) \, , \tag{A.65}$$

and thus plugging (A.61) and (A.62) into (A.59) (recall that $\frac{\partial^2 EU}{\partial Y^2} = 0$) gives

$$(1 - Yp_H)\frac{\partial^2 V}{\partial X^2}\frac{dX^*}{dB} + p_H\frac{\partial V}{\partial X}\left(\frac{dX^*}{dB} - \frac{dY^*}{dB}\right) = 0\,. \tag{A.66}$$

From (A.62) it follows that

$$\begin{aligned}(1 - Yp_H)\frac{\partial^2 V}{\partial X^2}\frac{dX^*}{dB} + p_H\frac{\partial V}{\partial X}\left(\frac{dX^*}{dB} - \frac{dY^*}{dB}\right)\\ + p_H\frac{\partial V}{\partial X}\left(\frac{dX^*}{dB} + \frac{dY^*}{dB}\right) = p_H\frac{\partial V}{\partial X}\end{aligned} \tag{A.67}$$

$$\Rightarrow\ (1 - Yp_H)\frac{\partial^2 V}{\partial X^2}\frac{dX^*}{dB} + 2p_H\frac{\partial V}{\partial X}\frac{dX^*}{dB} = p_H\frac{\partial V}{\partial X}\,, \tag{A.68}$$

and hence the change of investments in low- and high-status individuals, respectively, have the following forms:

$$\frac{dX^*}{dB} = \frac{p_H\frac{\partial V}{\partial X}}{(1 - Yp_H)\frac{\partial^2 V}{\partial X^2} + 2p_H\frac{\partial V}{\partial X}} \quad \xrightarrow[p_H \to 0]{<}\ 0\,, \tag{A.69}$$

$$\frac{dY^*}{dB} \to 1 > 0\,. \tag{A.70}$$

Note that if at the beginning $\frac{\partial EU}{\partial X} > \frac{\partial EU}{\partial Y}$ (since $\frac{dV}{dX}(1 - Yp_H) > p_H(U(H) - V(x))$ and as the response probability of high-status individuals, i.e., p_H, is very low), then there is a corner solution in which the whole budget is invested in low-status individuals, i.e., $X^* = B$ and $Y^* = 0$. An increase in the budget is followed by an increase of investments in low-status individuals, which in turn leads to a decrease in $\frac{\partial EU}{\partial X}$ (decreasing marginal returns) to the point where $\frac{\partial EU}{\partial X} = \frac{\partial EU}{\partial Y}$. From this point on, all additional budget is invested in high-status individuals, i.e., Y. Thus, we conclude that if a creator is endowed with a larger budget, then the more is invested in high-status individuals, i.e., $\frac{dY^*}{dB} > 0$. □

Proof of Proposition 6. We follow the expected utility given by (3), i.e.,

$$EU = \sum_{x=0}^{X}\sum_{y=0}^{Y}\binom{X}{x}\binom{Y}{y}p_L^x\, p_H^y\,(1 - p_L)^{X-x}\,(1 - p_H)^{Y-y}\ U(xL + yH)\,,$$

and the optimization problem given by (10), i.e.,

$$\underset{X,Y}{\text{Max}}\; EU(X,Y) \quad \text{s.t.} \quad X+Y=B\,.$$

To contrast the above case where the creator invests a fraction of her budget in low-status individuals and a fraction of her budget in high-status individuals, i.e., $X \neq 0$ an $Y \neq 0$, we now look at the case where the whole budget is invested in low-status individuals, i.e., $X = B$ and $Y = 0$. Then, according to Vandermonde's Convolution, the expected utility is given by

$$EU = \sum_{x=0}^{X} \binom{X}{x} p_L^x (1-p_L)^{X-x} \sum_{y=0}^{Y} \binom{Y}{y} p_L^y (1-p_L)^{Y-y} U(xL+yL) \tag{A.71}$$

$$= \sum_{z=0}^{B} \binom{B}{z} p_L^z (1-p_L^z)^{B-z} U(zL)\,, \tag{A.72}$$

where z is the number of cases, which yield a return L when investing B promotional actions in low-status individuals. Note that $z = x + y$. The entire budget, i.e., B, is invested in low-status individuals, if the expected return on high-status individuals is lower than on low-status individuals, i.e., $p_H H < p_L L$. Along these lines, we aim to show that for each portfolio, i.e., X and Y, it exists that

$$\sum_{z=0}^{B} \binom{B}{z} p_L^z (1-p_L)^{B-z} U(zL) \tag{A.73}$$

$$\geq \sum_{x=0}^{X} \sum_{y=0}^{Y} \binom{X}{x} \binom{Y}{y} p_L^x (1-p_L)^{X-x} p_H^y (1-p_H)^{Y-y} U(xL+yH) \tag{A.74}$$

$$= \sum_{x=0}^{X} \binom{X}{x} p_L^x (1-p_L)^{X-x} \sum_{y=0}^{Y} \binom{Y}{y} p_H^y (1-p_H)^{Y-y} U(xL+yH)\,. \tag{A.75}$$

More specifically, we aim to show that for any x (namely for each inequality separately) it exists that

$$\sum_{y=0}^{Y} \binom{Y}{y} p_L^y (1-p_L)^{Y-y} U(xL+yL) > \sum_{y=0}^{Y} \binom{Y}{y} p_H^y (1-p_H)^{Y-y} U(xL+yH)\,. \tag{A.76}$$

Let us assume that the creator gains with certainty at least the return of investments in low-status individuals, i.e., xL. As a result, the extra utility is given by

$$V(R) \equiv U(xL + R) - U(xL)\,, \tag{A.77}$$

where $U(xL)$ is constant and V is concave as U is invariant to translations. Hence, the inequality given by (A.76) can be rewritten as

$$\sum_{y=0}^{Y} \binom{Y}{y} p_L^y (1-p_L)^{Y-y} V(yL) > \sum_{y=0}^{Y} \binom{Y}{y} p_H^y (1-p_H)^{Y-y} V(yH)\,, \tag{A.78}$$

where $V(0) = 0$. By elimination, if we assume that V is concave, it holds that $\frac{V(R)}{R}$ is decreasing because $\frac{d}{dR}\frac{V(R)}{R} = \frac{RV' - V}{R^2} < 0$. Furthermore, since $R > 0$ it follows that $R' < \frac{V}{R}$. Moreover, let us assume that high-status individuals are more powerful than low-status individuals, i.e., $H > BL \geq YL$, and that $\frac{V(YL)}{YL} < \frac{V(yL)}{yL} < \frac{V(L)}{L}$. As $\frac{V(R)}{R}$ is decreasing and $yH > YL$, it follows that $\frac{V(yH)}{yH} < \frac{V(YL)}{YL}$ for any $1 \leq y \leq Y$. Since $V(0) = 0$ and $V\frac{V(YL)}{YL} < \frac{V(yL)}{yL}$ holds for any y, it follows that

$$\sum_{y=0}^{Y} \binom{Y}{y} p_L^y (1-p_L)^{Y-y} V(yL) = \sum_{y=1}^{Y} \binom{Y}{y} p_L^y (1-p_L)^{Y-y} \frac{V(yL)}{yL} yL \tag{A.79}$$

$$\geq \frac{V(YL)}{YL} \sum_{y=1}^{Y} \binom{Y}{y} p_L^y (1-p_L)^{Y-y} yL \tag{A.80}$$

$$= \frac{V(YL)}{YL} Y p_L L \tag{A.81}$$

$$> \frac{V(YL)}{YL} Y p_H H \tag{A.82}$$

$$= \frac{V(YL)}{YL} \sum_{y=1}^{Y} \binom{Y}{y} p_H^y (1-p_H)^{Y-y} yH\,. \tag{A.83}$$

As $p_H H < p_L L$ and since for any y it holds that $YL < yH$, and hence $\frac{V(YL)}{YL} > \frac{V(yH)}{yH}$, we conclude that

$$\frac{V(YL)}{YL} \sum_{y=1}^{Y} \binom{Y}{y} p_H^y (1-p_H)^{Y-y} yH > \sum_{y=1}^{Y} \binom{Y}{y} p_H^y (1-p_H)^{Y-y} V(yH)\,, \tag{A.84}$$

and further

$$\sum_{y=0}^{Y} \binom{Y}{y} p_L^y (1-p_L)^{Y-y} V(yL) > \sum_{y=0}^{Y} \binom{Y}{y} p_H^y (1-p_H)^{Y-y} V(yH) \,. \tag{A.85}$$

Thus, we conclude that if the return on high-status individuals is much higher than on low-status individuals such that $H > BL$ but the expected return on high-status individuals is lower than on low-status individuals, i.e., $p_H H < p_L L$, and there is an optimal solution such that a creator invests in high-status individuals, i.e., $Y^* \neq 0$, then the creator cannot be risk averse (and the individual utility function U is not concave). □

A.2 Expected Total Return on a Seeding Target

On one hand, let P_{ST} be the a priori probability that the seeding target follows the creator within a week, where S stands for the creator type and T for the seeding target type. The creator and the seeding target can take any of the four different user types, i.e., $S = \{1, 2, 3, 4\}$ and $T = \{1, 2, 3, 4\}$. On the other hand, let R_{ST} be the a priori probability that the seeding target reposts a song of the creator within a week and $f_{TS}(z)$ be the a priori probability distribution that the creator gets as a result z followers within a week. Finally, we combine both the a priori response and song repost probabilities in order to calculate the expected total return on each seeding target, given the status of the creator. So let P_{zST} be the a priori probability that the creator gets z followers when choosing a certain type of seeding target. The a priori probabilities for all possible returns z are given by

$$P_{0ST} = (1 - P_{ST}) \left[R_{ST} f_{TS}(0) + (1 - R_{ST}) \right] , \quad \text{if } z = 0 \tag{A.86}$$

$$P_{1ST} = P_{ST}(1 - R_{ST}) + (1 - P_{ST}) R_{ST} f_{TS}(1) + P_{ST} R_{ST} f_{TS}(0) , \quad \text{if } z = 1 \tag{A.87}$$

$$P_{zST} = P_{ST} R_{ST} f_{TS} \left(z - 1 \right) + (1 - P_{ST}) \, R_{ST} f_{TS}(z) , \quad \text{if } z > 0 \tag{A.88}$$

where $\sum_z f_{TS}(z) = 1$ and $P_{0ST} + P_{1ST} + \sum_z P_{zST} = 1$. Then, the expected total return μ_{ST} and the standard deviation s_{ST} are given by

$$\mu_{ST} = \sum_z Z P_{zST} , \tag{A.89}$$

$$s_{ST} = \sqrt{\sum_z (z - \mu_{ST})^2 P_{zST}} . \tag{A.90}$$

A.3 Simulation Study: Comparing the Effectiveness of Seeding Policies

The randomized dissemination processes comparing the three seeding policies, i.e., investments (1) according to the actual portfolios observed in the data, (2) only in seeding targets with the highest status, and (3) only in seeding targets with the lowest status, are based on the status-dependent probability of a non-zero return on a seeding target and further on the status-dependent probability of either a direct or indirect return. For this reason, let us define the following probabilities:

- p_D is the probability of a direct return in the form of a follow-back from the seeding target,
- p_R is the probability of a song repost from the seeding target,
- $p_{D\&R}$ is the joint probability of a direct and indirect return,
- $p_{\bar{D}\&R}$ is the joint probability of no direct but an indirect return,
- $p_{R|D}$ is the probability of a song repost given a direct return,
- $p_{R|D} = \frac{p_{D\&R}}{P_D}$ is the probability of a song repost given a direct return,
- $p_{R|\bar{D}} = \frac{p_{\bar{D}\&R}}{p_{\bar{D}}}$ is the probability of a song repost given no direct return,
- $p_{I|R}$ is the probability of a non-zero indirect return in the form of follows from the seeding target's egocentric network given a song repost.

Hence, the probability of gaining only a *direct return* is $p_D(1 - p_{R|D}) + p_D\, p_{R|D}(1 - p_{I|R})$, whereas the probability of gaining only an *indirect return* is $(1 - p_D)p_{R|\bar{D}}\, p_{I|R}$. Finally, the probability to gain a *direct and indirect return* is $p_D\, p_{R|D}\, p_{I|R}$ and thus the probability of a *zero return* on a seeding target is

$$P_0 = (1 - p_D)\left[p_{R|\bar{D}}(1 - p_{I|R}) + (1 - p_{R|\bar{D}})\right] . \tag{A.91}$$

As a result, the probability of a *non-zero return* on a seeding target is

$$P = 1 - P_0 . \tag{A.92}$$

The status-dependent probability of a non-zero return, given by (A.92), is applied in every

time step of the randomized dissemination process and on each target that is seeded using the monthly budget of promotional actions, i.e., 40 times in each of the 24 months. If the return in a given month for a given seeding target is non-zero, then there are two different scenarios. On one hand, the investment in this seeding target can yield both a direct and indirect return, i.e., a follow-back from the seeding target and follows from subsequent song reposts. On the other hand, the investment in this seeding target can yield only an indirect return, i.e., follows from a song repost.

Given a non-zero return, let $q_{D|P}$ be the probability of only a direct return and $q_{D\&I|P}$ be the joint probability of a direct and (non-zero) indirect return. Then, the *conditional probability that given a non-zero return on a seeding target there is also an indirect return* is

$$Q = q_{D|P} + q_{D\&I|P} \tag{A.93}$$

$$= \frac{p_D(1 - p_{R|D}) + p_D\, p_{R|D}(1 - p_{I|R})}{P} + \frac{p_D\, p_{R|D}\, p_{I|R}}{P} = \frac{p_D}{P}\,. \tag{A.94}$$

It follows that given a non-zero return, with probability Q the investment in a seeding target yields a direct return H_D in the form of a follow-back from this seeding target. As there might also be an additional indirect return associated with the direct return, we account for the follows resulting from subsequent song reposts. Therefore, let H be the average number of follows from the seeding target's egocentric network resulting from a single song repost. By further considering the average number of song reposts a over a time period τ, we allow for a longterm indirect return on a follow-back. Thus, the *direct and indirect return within a time period* τ is given by

$$H_D = 1 + a\tau H\,. \tag{A.95}$$

With probability $1 - Q$ the investment in a seeding target yields only an indirect return H, which is the average number of follows from the seeding target's egocentric network resulting from the single song repost. Hence, if an investment in a seeding target is successful, i.e., $c_i = 1$, then let the conditional probability to gain return z_i on a promotional action i be

$$Prob(z_i|c_i = 1) \begin{cases} Q & H_D = 1 + a\tau H \\ 1 - Q & H \end{cases} \tag{A.96}$$

Note that if an investment in a seeding target is not successful, i.e., $c_i = 0$, then the return is zero, i.e., $z_i = 0$.

To sum up, the status-dependent probability of a non-zero return, given by (A.92), is applied in every time step of the randomized dissemination process and on each target that is seeded using the monthly budget of promotional actions, i.e., 40 times in each of the 24 months. If the return in a given month for a given seeding target is non-zero, then there are two different scenarios. On one hand, the investment in this seeding target can yield both a direct and indirect return, i.e., H_D with probability Q. On the other hand, the investment in this seeding target can yield only an indirect return, i.e., H with probability $1-Q$. Note that we consider the average number of song reposts from a seeding target over a year, i.e., $\tau = 1$ year. This scheme overestimates the return on high-status individuals, hence, overestimating the power of influencers.

Furthermore, we account for the status-dependent natural baseline follows, which is on average 0.93 follows per month for an unknown creator. This in turn increases the creator's status, which goes hand in hand with higher a priori probabilities and thus expected total returns on each seeding target. Both the natural baseline follows and the a priori probabilities including the expected returns on each seeding target are updated in multiples of 25 followers with regard to the growing follower base, i.e., after a reaching a community size of ≥ 25 followers, ≥ 50 followers, and so forth. To sum up, the randomized dissemination process first takes into account the status-dependent probability of a non-zero return on a seeding target and subsequently considers whether the non-zero return is realized directly or indirectly.

A.4 Elaborations on Equations 6 – 9

The decision problem of selecting S products by means of a prediction model involves considerations about profits Π. More specifically, the manager considers expectations about revenues, $(1 - PF(S)) \cdot S \cdot R$, as well as associated costs, $S \cdot C$. While deciding on the size

of the selection of products, the manager maximizes the profits given by

$$\Pi = (1 - PF(S)) \cdot S \cdot R - S \cdot C \quad \text{s.t.} \quad S \geq 0\,, \tag{A.97}$$

which is in fact equivalent to the profits defined in Equation 10, i.e.,

$$\Pi = \overline{\Pi} - LR(S) - WI(S) \quad \text{s.t.} \quad S \geq 0\,, \tag{A.98}$$

where $\overline{\Pi}$ is the maximum profit potential given by

$$\overline{\Pi} = n \cdot (R - C)\,, \tag{A.99}$$

$LR(S)$ is the lost return given by Equation 9, i.e.,

$$LR(S) = (1 - PD(S)) \cdot n \cdot (R - C)\,, \tag{A.100}$$

and $WI(S)$ is the wasted investment given by Equation 8, i.e.,

$$WI(S) = PF(S) \cdot S \cdot C\,. \tag{A.101}$$

Since the probability of detection and probability of false discovery given by Equation 5 and Equation 6, respectively, are two sides of the same coin and given by Equation 7, i.e.,

$$PD(S) = \frac{S}{n} \cdot (1 - PF(S))\,, \tag{A.102}$$

we can substitute Equation A.102 in Equation A.100, which gives

$$LR(S) = n \cdot (R - C) - (1 - PF(S)) \cdot S \cdot (R - C)\,. \tag{A.103}$$

Then, the substitution of Equation A.99, Equation A.103, and Equation A.101 in Equation A.98 results in the optimization problem defined in Equation A.97.

To guarantee that S is non-negative, we apply the Kuhn-Tucker conditions and maximize profits by applying the method of Lagrange multipliers, namely

$$\underset{s,t,\lambda}{\text{Max}}\ \mathcal{L} = \Pi - \lambda(S - t^2)\,, \tag{A.104}$$

where the respective FOCs are

$$(1) \quad \frac{\partial \mathcal{L}}{\partial S} = \frac{\partial \Pi}{\partial S} + \lambda = 0\,, \tag{A.105}$$

$$(2) \quad \frac{\partial \mathcal{L}}{\partial t} = -2\lambda t = 0\,, \tag{A.106}$$

$$(3) \quad \frac{\partial \mathcal{L}}{\partial x} = s - t^2 = 0\,. \tag{A.107}$$

From Equation A.106 we find that either $\lambda = 0$ or $t = 0$. If $\lambda = 0$, then from Equation A.97 and Equation A.105 we obtain that

$$\frac{\partial \Pi}{\partial S} = -\frac{\partial PF(S)}{\partial S} \cdot S \cdot R + (1 - PF(S)) \cdot R - C = 0\,, \tag{A.108}$$

and hence

$$\frac{\partial PF(S)}{\partial S} \cdot S + PF(S) = \frac{ROI}{1 + ROI}\,, \tag{A.109}$$

where ROI is the proportion of profits made with the investment in a successful product, $ROI = \frac{R-C}{C}$. If $t = 0$, then from Equation A.107 we obtain that $S^* = 0$ and in this case $\Pi = 0$ (see Equation A.97). Namely, we choose S^* that resolves Equation A.109 if and only if profits Π are non-negative, i.e.,

$$\Pi = (1 - PF(S)) \cdot S \cdot R - S \cdot C > 0\,, \tag{A.110}$$

which can be rewritten as

$$PF(S^*) < \frac{ROI}{1 + ROI}, \tag{A.111}$$

otherwise $S^* = 0$ and no selection is made at all.

A.5 Elaborations on the Empirical Test and Application

We use unique longitudinal data from SoundCloud to study music artists' success over several years since their sign-up. Below, we offer a comprehensive description of the data and the early-stage predictors for success, which is followed by a detailed description of the modeling approach.

Data

We use unique data from SoundCloud, the biggest user-generated content network in the domain of music with more than 175 million users (Pierce 2016). SoundCloud facilitates the exchange of information among users and is specifically geared towards music artists and their fan communities. Artists upload their latest songs to their user profiles, which can then be listened to, commented on, reposted, and liked by other users. As on any other directed user-generated content network such as Twitter, users can exchange private messages and also follow each other to receive notifications about their latest activities. Users on SoundCloud have different objectives: Music artists seek to build and increase their follower base to generate more song-plays, whereas fans follow their favorite artists, listen to their songs and connect with other fans. To trigger follow backs (and thus increase their follower base and generate more song-plays), music artists can reach out to other users by following them, sending them private messages, reposting their songs, commenting on their songs, and liking their songs.

Our longitudinal dataset comprises 35,000 users (4,978 music artists and 30,022 fans) who signed up in the first week of March 2013 and covers all incoming and outgoing activities of each user, including follows, messages, song-plays, song-reposts, song-comments, and song-likes, over a period of more than two years (March 2013 to July 2015). Table 8 presents relevant descriptive statistics about the 35,000 users in our dataset.

— Insert Table 8 about here —

To test our proposed managerial framework, our first goal is to identify predictors to effectively detect future successful music artists on SoundCloud several weeks after their sign-up. We thus focus on artists who are initially unknown – with negligible popularity before sign-up – and who invest time in self-promotion activities on SoundCloud. To achieve a sample of artists with a high activity level and negligible popularity before sign-up in March 2013, we define active but unknown music artists on SoundCloud as all users who (1) uploaded at least one song within three months after sign-up, (2) received fewer than 100 song-plays in the first month, and (3) made at least one follow, private message, song-repost, song-comment, or song-like in each quarter after sign-up. The resulting sample comprises 534 music artists. As predictors for success we consider (1) artists' activities and (2) their evolving social capital on SoundCloud.

Music Artists' Activity Measures

Since music artists on SoundCloud use the social networking platform for self-promotion purposes to generate more song-plays, they reach out to other users by following them, sending them private messages, reposting their songs, commenting on their songs, and liking their songs. In fact, they aim to trigger follow backs to increase their follower base, which in turn generates more song-plays and makes them more successful. Saboo, Kumar, and Ramani (2015), who study the influence of network activities on sales based on data from multiple social networks in the music domain, find evidence that network activities drive sales directly but also indirectly, by triggering cascades of further activities via newly acquired followers. Similarly, Ansari et al. (2016) find that music artists can capitalize on network activities to alter and shape their egocentric network structure, which subsequently influences the number of song-plays on their profiles. They find that specifically increasing the clustering within the egocentric network impacts long-term success. Hence, in our specification of the prediction model, we assume that promotional outgoing activities directly affect success (i.e., song-plays), and indirectly (through the artist's egocentric network). We therefore include activity measures as predictors of success and let $promo_{it}$ denote artist i's average monthly

outgoing promotional activities (i.e., follows, messages, etc.) in the period from sign-up to present time t. Additionally, let $songs_{it}$ denote the number of artist i's uploaded songs in the period from sign-up to present time t.

Music Artists' Social Capital Measures

Besides the direct influence of outgoing promotional activities on success, there exists an indirect influence through the artist's egocentric network as outlined above. The value of an individual egocentric network is known as social capital (e.g., Coleman 1988, 1990; Putnam 2001) and encompasses all "[...] resources embedded in a social structure which are accessed and/or mobilized in purposive actions" (Lin 1999, p. 35). Hence, social capital determines the music artist's constraints and opportunities (Wellman and Berkowitz 1988). The general concept has emerged in a large field of research with applications in many disciplines of social sciences (for an overview see Kwon and Adler 2014). Following network theorists in sociology (e.g., Burt 1992), social capital takes an inside-out view, focusing on the music artist (ego) as well as on the different layers of the artist's follower base (egocentric network). In line with social capital measures (Borgatti, Jones, and Everett 1998; Burt 1983), these layers are most commonly characterized by the structure or density of the egocentric network and, furthermore, by centrality measures (degree, betweenness, and closeness centrality) that assess the music artist's position within the entire network (Freeman 1978; Wasserman and Faust 1994).

Although a comprehensive assessment of social capital, including computations of closeness and betweeness centralities, is rarely possible as it requires full information on the entire network, we use several measures of artists' social capital on SoundCloud. The focus of our study is on the dissemination of information about the music artist itself, which spreads exclusively through the artist's indegree, the users who follow the artist. For example, all followers of an artist receive a notification in their news feed if the artist uploads a new song. Hence, especially indegree (first-degree followers) plays a potential crucial role for an artist's

future success. Moreover, the number of first-degree followers is the most common yet basic operationalization of an artist's network status (Hu and Van den Bulte 2014; Sauder, Lynn, and Podolny 2012), which accounts for the adjacent network users (Gould 2002; Shaw 1954; Stuart 1998; Wasserman and Faust 1994). Along these lines, let $socap_{cit}$ denote a social capital measure c at a certain present time t, where the first of four measures – *the first-degree followers of artist i at time t* – is defined by

$$socap_{1it} = FD_{it} = \sum_k a(k, i, t) \qquad \textit{(First-Degree Followers)}, \tag{A.112}$$

where $a(k, i, t)$ is the adjacency matrix of the network at time t such that $a(k, i, t) = 1$ if and only if a SoundCloud user k follows artist i at time t. By definition, no one can follow him- or herself, $a(i, i, t) = 0$ for all i's at all t's.

According to Yoganarasimhan (2012), however, the number of first-degree followers given by Equation A.112 is also a deceptive connectivity measure, because it fails to capture the notion of potential: After uploading a song, the music artist's first-degree followers can trigger cascades (Watts 2002; Watts and Dodds 2007) by means of song-reposts, which then spread and appear in the news feeds of their own followers (the artist's second-degree follower base) and further. The second-degree follower base thus represents the artist's as-yet-unfulfilled potential of follower base growth (Yoganarasimhan 2012), where the speed and magnitude of growth depends on the number of hubs who follow the music artist (Goldenberg et al. 2009). Hence, our second social capital measure is the indirect adjacency – *the unique second-degree followers of artist i at time t* – given by

$$socap_{2it} = SD_{it} = \sum_j b(j, i, t) \qquad \textit{(Second-Degree Followers)}, \tag{A.113}$$

where $b(j, i, t) = 1$ (SoundCloud user j is a second-degree follower of music artist i at time t) if and only if $a(j, i, t) = 0$ (j does not follow i) but there exists at least one other SoundCloud user k ($k \neq i$ and $k \neq j$) such that $a(j, k, t) = 1$ (j follows k) and also $a(k, i, t) = 1$ (k follows i).

Future success is also affected by the degree of connectivity within the artist's follower base. The clustering coefficient, which determines the number of closed triangles (Watts and Strogatz 1998), provides insight into the density within the direct adjacency. According to Ansari et al. (2016), high connectedness among the first-degree followers of a music artist (first-degree clustering) impacts long-term success. Furthermore, higher first-degree clustering also speeds up the spread of information (Jackson 2010) and goes hand in hand with stronger network multiplier effects (Watts and Dodds 2007). Therefore, our third social capital measure – *the first-degree clustering of artist i at time t* – is defined by

$$socap_{3it} = C_{it} = \frac{\sum_{j,k} a(j,i,t) \cdot a(k,i,t) \cdot a(k,j,t)}{FD_{it} \cdot (FD_{it} - 1)} \qquad \textit{(First-Degree Clustering).} \tag{A.114}$$

where a triangle is closed if and only if $a(k,i,t) = 1$ and $a(j,i,t) = 1$ (SoundCloud users k and j follow artist i at time t) and at least $a(k,j,t) = 1$ or $a(j,k,t) = 1$ (at least one of the SoundCloud users k and j follows the other). From Equation A.114 it follows that the higher the first-degree clustering, the more connected the first-degree follower base.

Reciprocity – the artist's share of reciprocal ties – is another form of density, with greater focus on the music artist (Newman 2010). After receiving a follow, the artist can strengthen this tie by following back and establishing a bi-directional connection (Granovetter 1973). As a core component of exchange (Blau 1963), reciprocity potentially affects future success by fostering trust between the music artist and his or her followers (Coleman 1988). A follow back may also be considered a favor or benefit that an artist grants to a fan, which might then be reciprocated in the form of voluntary endorsement (Hargadon and Sutton 1997), which represents a form of generalized reciprocity, which Putnam (2000, p. 134) sees as the "touchstone of social capital". Hence, the fourth social capital measure – *the reciprocity of artist i at time t* – is defined by

$$socap_{4it} = R_{it} = \frac{\sum_{k} a(k,i,t) \cdot a(i,k,t)}{FD_{it}} \qquad \textit{(Reciprocity).} \tag{A.115}$$

where a follow is reciprocated if and only if $a(k,i,t) = 1$ (SoundCloud user k follows music artist i at time t) and also $a(i,k,t) = 1$ (i follows k). Put differently, this measure calculates

the proportion of follows that are doubled through establishing a bi-directional connection. Therefore, the more a music artist follows back if being followed, the stronger his or her bonding with the first-degree follower base.

We use these four egocentric network measures (first-degree followers, second-degree followers, first-degree clustering, and reciprocity) to assess different aspects of an artist's evolving social capital on SoundCloud.

Modeling Approach

In our sample of 534 unknown music artists with a high activity level but negligible popularity outside of SoundCloud when signing up, only 55 artists (10%) achieve to exceed more than three orders of magnitude in regard to his or her average received monthly song-plays (i.e., more than 1,000 song-plays) – the highest being of order five (i.e., 100,000 song-plays). Of the 534 artists, 479 (90%) were unsuccessful, based on this criteria. 135 manage to achieve between 100 and 1,000 average monthly song-plays and 344 artists received fewer than 100 average monthly song-plays.

Our longitudinal dataset spans the 123-week period from March 2013 to July 2015. We define the target period for prediction T as the final 12 weeks of our longitudinal dataset, namely the period from week 111 to 123, i.e., $T = [111, 123]$. We further define an artist's success in the target period for prediction T by y_{iT}, where $y_{iT} = \{1, 2, 3\}$, depending on the order of magnitude of the artist's average monthly song-plays. In our application, artist i is defined as successful if he or she achieves to exceed more than three orders of magnitude in regard to his or her average received monthly song-plays (i.e., $> 1,000$) in the target period for prediction T, i.e., $y_{iT} = 3$.

For each music artist in our sample, we define an early-stage observation period for prediction t, which starts with his or her sign-up on SoundCloud. To identify early-stage predictors for success at T, we focus on four, eight, and twelve weeks after an artist's sign-

up. Therefore, the early-stage observation period for prediction t may take three values, $t = \{4, 8, 12\}$. As stated, $songs_{it}$ denotes the number of uploaded songs and $promo_{it}$ the average monthly outgoing promotional activities to present time t. Also, $socap_{cit}$ denotes the social capital measure c at time t, where $c = \{1, 2, 3, 4\}$, namely (1) first-degree followers FD_{it}, (2) second-degree followers SD_{it}, (3) first-degree clustering C_{it}, and (4) reciprocity R_{it}. The resulting ordered-logit model can thus be described as follows:

$$Pr(y_{iT} = L) = Pr(\kappa_{L-1} < score_{it} \leq \kappa_L)\,, \tag{A.116}$$

where $score_{it} = \beta_1 plays_{it} + \beta_2 promo_{it} + \beta_3 songs_{it} + \beta_4 socap_{cit} + u_i$. Note that $plays_{it}$ denotes the artist's i average received monthly song-plays to present time t, κ_0 is defined as $-\infty$ and κ_k as $+\infty$, and u_i is assumed to be logistically distributed.[14]

Firth (1993) penalized logit model and the oversampling technique including bias correction suggested by King and Zeng (2001a,b), are both widely used in the case of rare events, where both aim to produce less biased a posteriori probabilities of success. Hence, to assess the predictive power of alternative model specifications, we first evaluate early-stage predictors for success using the above-described ordered-logit model. Given the most effective predictors for success, we then compare the ordered-logit model to the following models typically used for rare events: (1) *logit*, (2) $Firth$[15] (Firth 1993; Heinze 2006; Heinze and Schemper 2002), and (3) $ReLogit$[16] (Imai, King, and Lau 2008; King and Zeng 2001a,b).

[14]Adding more than one social capital measure does not improve the predictive power of the model in terms of probability of detection or false discovery, because variables are calibrated in the training set.

[15]Embedded in the logistf package of statistical software R, the Firth procedure is the bias reduction method that penalizes the log-likelihood using the Jeffreys invariant prior.

[16]Embedded in the Zelig package of the statistical software R, the ReLogit procedure is the rare events logistic regression for dichotomous dependent variables embedded in the Zelig package.

Table 8: Descriptive Statistics

		Sample 2012-2015		
Descriptives		Mean	Median	Std
Indegree	Mar. 2014	53.07	5.00	12.10
	Jun. 2015	20.01	8.00	102.10
Follows	sent	35.21	10.00	91.57
	received	20.01	8.00	102.10
Song-Comments	sent	5.52	2.00	19.84
	received	12.33	3.00	97.49
Song-Likes	sent	32.74	4.00	101.84
	received	89.07	6.00	1014.00
Messages	sent	10.00	2.00	67.55
	received	5.25	1.00	68.72
Song-Plays	sent	771.25	37.00	2885.87
	received	4785.35	176.00	59770.22
Song-Reposts	sent	0.04	0.00	0.70
	received	0.02	0.00	1.17
Tracks	uploaded	10.13	3.00	37.85
Weekly Follows	sent	0.19	0.00	4.37
	received	0.10	0.00	1.13
Weekly Song-Comments	sent	0.01	0.00	0.28
	received	0.01	0.00	0.32
Weekly Song-Likes	sent	0.12	0.00	1.28
	received	0.07	0.00	3.19
Weekly Messages	sent	0.003	0.00	0.21
	received	0.004	0.00	0.25
Weekly Song-Plays	sent	5.20	0.00	34.13
	received	5.98	0.00	266.50
Weekly Song-Reposts	sent	0.003	0.00	0.21
	received	0.004	0.00	0.25
Weekly Tracks	uploaded	0.01	0.00	0.35

Curriculum Vitae

2013–2017	**Ph.D. in Business Administration** (June 2, 2017) *University of Mannheim* | *Mannheim Business School*, Mannheim DE
2011–2013	**Master of Arts in Business Administration** (October 23, 2013) *University of Zurich*, Zurich CH
2007–2011	**Bachelor of Arts in Management** (August 31, 2011) *University of Fribourg*, Fribourg CH