Idea Generation, Creativity, and Prototypicality

Olivier Toubia and Oded Netzer

Columbia Business School

The authors would like to thank Alain Lemaire and Erin Michet for their outstanding research assistance on this project.

Idea Generation, Creativity, and Prototypicality

We explore the use of Big Data tools to shed new light on the idea generation process, automatically "read" ideas in order to identify promising ones, and help people be more creative. The literature suggests that creativity results from the optimal balance between novelty and familiarity, which should be measured based on the combinations of words in an idea. We build semantic networks where nodes represent word stems relevant to a particular idea generation topic, and edge weights capture the novelty vs. familiarity of word stem combinations (i.e., the weight of an edge that connects two word stems measures their scaled co-occurrence). Each idea contains a set of word stems, which form a semantic subnetwork. The edge weight distribution in that subnetwork reflects how the idea balances novelty with familiarity. Based on the "beauty in averageness" effect, we hypothesize that ideas with semantic subnetworks that have a more prototypical edge weight distribution are judged as more creative. We show this effect in eight studies involving over 4,000 ideas across multiple domains. Practically, we demonstrate how our research can be used to automatically identify promising ideas, and recommend words to users on the fly to help them improve their ideas.

1. Introduction

"Big Data" tools and methods have heavily focused on improving the effectiveness of advertising or other marketing vehicles. In this paper we explore whether and how Big Data tools may be leveraged in other marketing-related domains. In particular, we focus on idea generation, which is a critical aspect of product development, innovation, and advertising. We explore whether, and how Big Data tools may be leveraged to shed new light on the idea generation process, automatically "read" ideas in order to identify promising ones, and help people be more creative in practice.

We adopt a cognitive view of idea generation according to which generating ideas involves retrieving knowledge from long-term memory (Finke, Ward and Smith 1992). This memory retrieval stage of the idea generation process, in which people select the "ingredients" that will be combined to form a new idea, lends itself well to systematic, computer-based analysis. This raises the question of whether and how the judged creativity of an idea may be linked to its "ingredients," i.e., to the set of words present in the idea. To answer this question, we rely on the creativity literature that suggests that creativity lies in the optimal balance between novelty and familiarity. This raises three new questions: (*i*) How exactly should novelty and familiarity be *defined* in the context of idea generation? (*ii*) How may novelty and familiarity be *measured*? (*iii*) What constitutes an optimal balance between novelty and familiarity? To answer the first question, we rely on a literature that has established the associative nature of creativity, i.e., creativity relies on associations. Therefore, it is appropriate to relate novelty to uncommon associations of words and familiarity to common associations. For example, consider a recipe for a new dish. Novelty does not necessarily come from choosing novel ingredients for the recipe but rather from choosing ingredients that do not often appear together – both chicken and chocolate are very common and familiar ingredients in recipes but the combination of these two ingredients is novel. Because we focus on the association between words to represent novelty and familiarity, we turn to the rich literature in knowledge discovery and co-word analysis to answer the second question (e.g., Callon et al. 1986). Using standard text mining tools, we organize the word stems related to a given idea generation

topic into a semantic network. Nodes in this network represent word stems, and the weight of an edge that connects two word stems measures their scaled co-occurrence. A high edge weight means that the two corresponding word stems appear frequently with one another, i.e., their combination is familiar. Conversely, a low edge weight means that the two corresponding word stems appear infrequently with one another, i.e. their combination is novel. The subset of word stems involved in an idea form a semantic subnetwork. The edge weights in this subnetwork reflect a distribution between familiar (i.e., strongly connected) and novel (i.e., weakly connected) combinations of word stems. That is, the balance between novelty and familiarity is captured by the distribution of edge weights in the subnetwork. Finally, we answer the third question based on the "beauty in averageness" effect, which postulates that prototypes have inherent qualities and properties that robustly make them more appealing. This leads us to our hypothesis that *ideas with semantic subnetworks that have a more prototypical edge weight distribution tend to be judged as more creative*.

It is important to note that prototypicality of the edge weight distribution does not mean that word stems used in the idea are prototypical or common, but rather that the structure of the semantic relationships among these word stems is prototypical. Note that we define an "idea" as a document made of words that attempts to add value given a particular idea generation topic. Each word is associated with a unique word stem, and each stem may be associated with one or many words (e.g., the words "adventure," "adventures," "adventurous" all belong to the word stem "adventur").

We test and validate our hypothesis across eight studies, involving over 4,000 ideas generated by over 2,000 people. While we focus on judged creativity as our primary measure of quality, we show that the effect also holds with alternative measures of idea quality, coming from consumers or industry experts. Five of our studies were run in collaboration with companies that were interested in ideas for new products or services, or that host idea generation communities. Participants in our studies varied from Amazon Mechanical Turk to commercial online panels to members of an idea generation community. The idea generation topics varied from smartphone apps to oral care to insurance products. Our last study provides a proof of concept that our findings may be used to construct automatic tools to assist people in

the memory retrieval step of the idea generation process. In particular, we show that it is possible to build tools that text mine ideas in real time, and automatically recommend words or "ingredients" to help people improve their ideas.

The rest of the paper is organized as follows. In Section 2, we review some relevant literature and justify our main hypothesis. In Section 3 we introduce the various steps of our empirical approach, including constructing semantic networks, quantifying the prototypicality of edge weight distributions, as well as generating and evaluating ideas. In Section 4 we report the results of our studies. Section 5 concludes and offers suggestions for future research.

2. Theoretical Development

2.1. Idea Generation

Our research is based on a cognitive view of idea generation, which is based on the premise that one must rely on some type of stored information when developing new ideas (e.g., Goldenberg and Mazursky 2002; Simonton 2003). Indeed, it is well established that generating ideas involves retrieving knowledge from long-term memory (e.g., Nijstad and Stroebe 2006; Nijstad Stroebe and Lodewijkx 2003).

In particular, the Geneplore model of Finke, Ward and Smith (1992) suggests that the generation of creative ideas involves two phases that are performed iteratively: a *generative* phase in which mental representations called preinventive structures are constructed, and an *exploratory* phase in which these structures are interpreted, modified and combined in meaningful ways. Put simply, the Geneplore model realizes that new ideas are not constructed in a vacuum, but rather that some basic ingredients or starting points (preinventive structures) are necessary. Burroughs, Moreau and Mick (2008) define preinventive structures as "symbolic patterns, exemplars, mental models, or unique verbal combinations that are precursors to the creative process." Preinventive structures are typically constructed by retrieving relevant concepts from long-term memory (Finke, Ward and Smith 1992; Perkins 1981). Moreau and Dahl (2005) provide the vivid illustration of a consumer needing to cook dinner. In that case a set of ingredients (e.g.,

peanut butter, spaghetti noodles, carrots, etc.) will form a preinventive structure that will form the basis for a solution.

The type of preinventive structures retrieved during the generative phase of the idea generation process will obviously have an effect on the quality of the ideas developed. As Ward (1995) notes, "Any time a person develops a new idea, it will be based to some extent on recalled information; however, the exact manner or form in which information is recalled may affect the likelihood of a creative outcome." However, very little is known regarding the relation between the characteristics of the pre-inventive structures retrieved during the generative phase of the idea generation process and the quality of the ideas developed, i.e., between the set of words that form the "ingredients" of an idea and the quality of that idea. In this paper we explore this relationship, by drawing on research from various fields including psychology, text mining, and network analysis. Studying this relationship is not only interesting theoretically, it also has practical implications. Indeed, the generative phase of the idea generation process relies on retrieval from long-term memory, which can be at least partially automated or assisted by computers. Therefore, understanding the relationship between the set of words in an idea and its judged creativity opens the door for automated tools that not only identify promising ideas, but also help people find the right "ingredients" to include or add into their ideas.

2.2. Balancing Novelty with Familiarity

The study of creativity in various domains, from scientific discovery (e.g., Uzzi et al. 2013) to linguistics (e.g. Giora, 2003), has pointed to the robust conclusion that creativity results from the optimal balance between novelty and familiarity. For example, Uzzi et al. (2013) link the impact of scientific papers (as measured by the number of citations) to the network of journals cited in these papers (i.e., how frequently the journals cited in a paper tend to be cited together). They find that papers are more likely to have high impact if they combine novelty and conventionality, i.e., if they cite papers from journals that are commonly cited together on average, with some very unusual combinations. In a context even closer to

ours, Ward (1995) notes that "truly useful creativity may reflect a balance between novelty and a connection to previous ideas."

Therefore, based on the creativity literature we can argue that an optimal set of "ingredients" in an idea is one that balances novelty with familiarity. This raises three questions: (*i*) How exactly should novelty and familiarity be defined in the context of idea generation? (*ii*) How may novelty and familiarity be measured? (*iii*) What constitutes an optimal balance between novelty and familiarity? The following three subsections address each of these questions in turn.

2.3 The Associative Nature of Creativity

One might be tempted to define novelty and familiarity in our context based on whether the word stems present in the idea are inherently common or novel themselves. In that case, the novelty or familiarity of a particular word stem would be measured based how frequently it appears in language related to the idea generation topic under consideration. However, the literature suggests that it is preferable to define and measure novelty and familiarity based on the *combinations* of word stems in the idea, rather than the individual word stems themselves. As we discussed previously, an idea for a new recipe that combines chicken with chocolate would be uncommon because these two ingredients are rarely found together, even though both ingredients are common in recipes.

Indeed, the creativity literature has suggested that associations between concepts are the basis of creativity. Dahl and Moreau (2002) argue that "researchers in cognitive psychology generally agree that creativity consists of *reassembling* elements from existing knowledge bases in a novel fashion" (page 48, emphasis added). Finke, Ward and Smith (1992) argue that "the *merging* of concepts is an inherently creative process" (page 108, emphasis added), and that a moderate level of incongruity among the concepts in an idea is useful in creative discovery. Mednick (1962) defines the creative thinking process as "the forming of associative elements into new combinations which either meet specified requirements or are in some way useful" (page 221). As background to this definition, Mednick relays introspective statements by several well-known scientists and artists including Albert Einstein (who wrote that

"combinatory play seems to be the essential feature in productive thought"), André Breton (according to whom artistic creativity comes the "juxtaposition of distant realities") and Henri Poincaré (who wrote that "to create consists of making new combinations of associative elements which are useful"). More recently, Rothenberg (2015) interviewed 34 Nobel laureates in various domains and concluded that integration, where "multiple separate elements retain their discreteness and identity while connected and operating together in a whole" (page 9), is the characteristic result of the cognitive creative process. Although Rothenberg (2015)'s study focuses on creativity in the scientific domain, he notes that: "applications of all of the cognitive creative processes, in whole or in selective part, certainly must play a role in other types of everyday and work-day creativity, such as in business and advertising" (page 190).

Based on this perspective, it seems reasonable to define novelty in our context as the association of word stems that do not appear frequently together in text related to the topic under consideration; and familiarity as the association of word stems that appear frequently together. In other words, our initial statement may be refined as follows: an optimal set of "ingredients" in an idea is one that balances novel combinations of word stems with familiar combinations of word stems. Therefore, throughout the remainder of the paper, unless otherwise specified, "familiarity" and "novelty" refer to *combinations* of word stems.

2.4. Semantic Networks

We have argued that novelty and familiarity may be measured by the strength of association between word stems. The next step is to measure these associations. For this, we turn to the literature on semantic networks and co-word analysis (Anderson 1983; Collins and Loftus 1975). A semantic network is a network that represents associations among a set of words or word stems (we focus on word stems).

Today, semantic networks may be constructed relatively easily from primary or secondary data using text mining analysis. (See Feldman et al., 1998 for a general introduction to text mining.) In a semantic network the nodes are word stems and the edges are based on some measure of co-occurrence among word stems. Word stems that appear together more frequently in textual data are connected by edges that

have higher weights, and are therefore closer to each other in the semantic network (Netzer et al. 2012). Thus, the measure of edge weights in a semantic network is directly related to our proposed definition of familiarity and novelty as the scaled co-occurrence of combinations of word stems. Because words can have different meanings and associations in different contexts (Anderson 1983), we build context-specific semantic networks for each idea generation topic.[1] More details are provided in Section 3.1.

Figure 1 provides an illustration of a semantic network from one of our studies in which consumers generated ideas for new insurance products designed to improve financial stability. Note that such a figure was created only to illustrate the concept of a semantic network in the present paper, but that it was not shown to any participant in any of our studies. Each idea on the topic involves a subset of the nodes (word stems) in the general network, which form a semantic subnetwork. If the semantic subnetwork corresponding to a given idea has N nodes, there are $N(N\text{-}1)/2$ edges in the subnetwork, where the weight of each edge captures the strength of association between two nodes in the network. Familiar combinations of word stems have higher edge weights, i.e., they are commonly found together in natural text related to the topic. In contrast, novel combinations of word stems have lower edge weights, i.e., their combinations are more unusual.

We could describe a given semantic network based for example on the average weight of its edges, or based on other statistics such as the variance, median, minimum, maximum, etc. However, in order to capture the *balance* between novel and familiar combinations of words, we need to consider the entire *distribution* of edge weights in an idea's semantic subnetwork.

[INSERT FIGURE 1 ABOUT HERE]

2.5. "Beauty in Averageness" Effect

We have argued that the creativity of an idea should be linked to the edge weight distribution of the semantic subnetwork associated with that idea, and that the optimal distribution is one that balances

[1] Text mining has been proposed previously as a method for generating new ideas by automatically linking streams of literature. For example, Swanson (1988) found relationships between magnesium and migraine and between biological viruses and weapons by mining disjoint literatures. Similarly, Kostoff (2006) proposed literature-based discovery of ideas via text mining of the academic literature about a topic. In this paper we use text mining to better understand which type of semantic structures make for a good idea, focusing on the context of innovation.

novelty and familiarity. This leaves us with our last question of what constitutes an optimal balance, i.e., an optimal distribution of edge weights in a semantic subnetwork. For this, we turn to a large literature spanning Psychology, Biology, Art and Business, that has shown that prototypes have inherent qualities and properties that robustly make them more appealing. This effect is sometimes labeled the "beauty in averageness effect."

The most well-known demonstration of the beauty-in-averageness effect is probably in the domain of human faces. A large number of studies have shown that humans find faces with average features more beautiful and attractive (e.g., Langlois and Roggman 1990; Strzalko and Kaszycka 1991). This effect has also been demonstrated for music performances (Repp 1997), polygons, drawings and paintings (Martindale, Moore and Borkum 1990), and words / exemplars (Martindale, Moore and West 1988). Demonstrations of this effect in business applications include Landwehr, Labroo and Herrmann (2011) and Veryzer and Hutchinson (1998).

Several explanations have been proposed for this effect, often relying on biology and evolution (Grammer and Thornhill 1994; Langlois and Roggman 1990; Thornhill and Gangestad 1993), or fluency (Landwehr, Labroo and Herrmann 2011; Reber, Schwarz and Winkielman 2004; Winkielman et al. 2006). A more straightforward explanation, which is also more relevant in our context, relies on the "wisdom of the crowds" phenomenon (Surowiecki 2005). Domains in which the beauty-in-averageness effect holds tend to be ones in which quality relies on the optimal balance between various features or the optimal distribution of resources across various dimensions. For example, a beautiful face is one in which the nose is neither too narrow nor too wide, a beautiful piano performance is one in which the key strokes are neither too heavy nor too light, etc. Each stimulus may be viewed as one attempt to find an optimal distribution or allocation. Taking the average of a set of stimuli cancels out the small errors made by each stimulus and gives rise to a distribution that is closer to optimal (Halberstadt and Rhodes 2003; Repp 1997). Using the same reasoning, we should expect that taking the average distribution of edge weights across documents gives rise to a prototypical distribution that optimally balances novelty and familiarity.

Therefore, our main hypothesis is that *ideas with semantic subnetworks that have a more prototypical edge weight distribution tend to be judged as more creative*.

3. Empirical Approach

We test our hypothesis and study its managerial implications across eight studies, which we describe in the next section. In this section we describe our overall empirical approach, which requires the following steps. We start by building a baseline semantic network related to each idea generation topic. We construct a prototypical distribution of edge weights. We collect ideas and idea evaluations and measure the prototypicality of each idea's edge weight distribution. Finally, we explore the link between prototypicality and judged creativity statistically.

3.1. Construction of the Baseline Semantic Network

Extracting Textual Data for the Baseline Semantic Network

We need to identify a text corpus that will allow us to construct a baseline semantic network capturing the set of word stems commonly related to the idea generation topic at hand. This baseline semantic network should be exogenous to the ideas being tested, i.e., the semantic network should not be constructed based on the ideas themselves. Indeed, if the baseline semantic network is derived from the ideas themselves, our measure of the prototypicality of each idea's edge weight distribution would become a function of which other ideas are included in the analysis.

Across our eight studies, we use two different approaches for constructing this baseline semantic network. In Study 1 the baseline semantic network comes from a set of pretest ideas in which we ask consumers (different from those involved in the main study) to generate an initial set of ideas on the topic. Unfortunately, this approach is costly (both in time and money) and it cannot be fully automated.

Therefore, in Studies 2 to 6, we test an alternative approach that leverages Google, and that can be fully automated. We simply perform a search query on Google using the exact wording of the idea generation topic as the text of the query. For example, if a study asks consumers to generate ideas on the

following topic: “How could smartphones help their users be healthier?” we copy and paste this exact sentence into Google as a search query. We then download the html page source code of the top 50 search results provided by Google. Throughout the paper we refer to these documents as a “Google results” or “pages retrieved from Google.” The advantage of using top search results from Google is that this information is readily available and can be scraped automatically with no human effort. However, this approach is not without its limitations. For example, the pages retrieved from Google might be biased towards certain types of content. In addition, while some portions of the pages may be relevant to the idea generation topic, others may not. Therefore, it is an empirical question whether Google may be used as a reliable source of text to create the baseline semantic network and prototypical edge weight distribution.

Text Mining

Once the text corpus has been collected, we need to mine the text to extract relevant word stems. We use the text-mining infrastructure in R (Feinerer, Hornik and Meyer 2008). Our text mining process includes the following steps. First we clean the text from irrelevant information such as pictures and HTML signs. Next, we tokenize the text into words. In the next step we use the Porter stemming algorithm implemented in R (Porter 1980) to automatically stem words into their stems or roots (e.g., “adventur” is a stem for the words “adventure,” “adventures,” and “adventurous”). Human experts checked the list of stems and associated words manually, to remove stems that are too generic (e.g, “five”) or manually split/combine stems that were not appropriately allocated by the stemmer. This step requires approximately one hour of human labor per ideation topic. In Study 5, we omit the manual cleaning of the stemmed words to explore how our approach may be applied to field data in a fully automated way. Once a final list of word stems and associated words was obtained, we retained only those word stems that appeared frequently enough (in at least 5 of the ideas generated in the pretest in Study 1, and at least 10 of the 50 pages retrieved from Google in Studies 2-6).

We used similar text mining extraction and stemming processes to extract words from the ideas generated in our studies.

Edge Weights

Several measures are available to quantify the edge weights in our semantic network, i.e., the scaled co-occurrence of pairs of word stems. We use a common measure, the Jaccard index (see e.g., Netzer et al., 2012). Consider two word stems *A* and *B*. Let S_A (respectively, S_B) be the set of training documents (pretest ideas or web pages) that contain stem *A* (respectively, *B*). The Jaccard index between word stems *A* and *B* is defined as:

$$J_{A,B} = \frac{|S_A \cap S_B|}{|S_A \cup S_B|}$$

Where $|S|$ denotes the size of set *S*. The Jaccard index is the ratio between the number of documents that contain both *A* and *B* and the number of documents that contain *A* or *B*. It is the probability that *A* and *B* appear in a randomly selected document, given that *A* or *B* appears in that document. (The intuition behind the Jaccard index may be visualized easily with a Venn diagram: it is the area of the intersection of S_A and S_B divided by their union.) A high value means that the two word stems appear frequently with one another, over and beyond chance based on their separate occurrences. Thus, seeing these two word stems in an idea is not surprising. On the other hand, a low Jaccard index means that these two word stems do not appear commonly in the textual corpus, thus seeing them together in an idea could be considered novel or surprising. Each node in our baseline semantic network corresponds to one word stem, and the weights of the edges among all possible pairs of nodes are captured by an incidence matrix of Jaccard indexes.

3.2. Network Features

Several features have been proposed in the literature to describe and characterize the structure of networks. As reviewed in the previous section, our key descriptor of a network is the distribution of *edge weights* in the network, where the weight of an edge that connects nodes *i* and *j* measures the scaled co-occurrence of these two nodes using the Jaccard index.

We consider control variables derived from two additional standard network features. The first is the set of *frequencies* of the nodes in the network, where the frequency of a node is the frequency of

occurrence of the corresponding word stem in the training text (i.e., proportion of pretest ideas or results from Google in which the word stem appears). Note that node frequency describes the properties of the nodes present in the network, rather than their relationships to one another. The second feature is the set of *clustering coefficients* of the nodes in the network, where the clustering coefficient of node *i* measures how interconnected the nodes that connect to *i* are to each other. Readers are referred to Barrat et al. (2004) for more details on these standard network features.[2]

3.3. Constructing the Prototypical Distribution of Edge Weights

We construct a different prototypical distribution of edge weights for each domain-specific baseline semantic network. We first compute the distribution of edge weights in the subnetwork corresponding to each of the pretest ideas/Google results used to construct the baseline network. For example, a subnetwork with 5 nodes may be described by a set of $\binom{5}{2} = 10$ weights (one per edge), which are distributed between 0 and 1 according to some cumulative distribution function (cdf). For instance, if 2 of the 10 edge weights are smaller than or equal to 0.3, the cdf would have value 0.2 at x=0.3. We then construct a prototypical distribution by taking the average of the distributions across pretest ideas/Google results. That is, the value of the prototypical cdf at any value x is the average of the values of the cdf at x across all pretest ideas/Google results. For example, if 5% of the edge weights are smaller than or equal to 0.1 in one pretest idea and 10% are smaller than or equal to 0.1 in another, the average cdf across these two ideas would have value 0.075 for x=0.1. Future research may explore alternative ways to construct the prototypical distribution, e.g., by computing the median instead of the average distribution, although the literature reviewed in Section 2.5. suggests that the average is more appropriate. Building the prototypical distribution using the pretest ideas or the Google results ensures that our prototypical distribution is not a function of the particular set of ideas being tested. This prototypical distribution

[2] Unlike many networks found in marketing, our semantic networks are weighted networks, i.e., the relationship between two nodes (word stems) is captured by a continuous variable (the Jaccard index, which varies between 0 and 1) rather than a binary one. We use a set of features that generalize standard features developed for binary networks, to weighted networks (Barrat et al. 2004).

serves as our benchmark for the optimal balance between novelty and familiarity. Web Appendix A shows the prototypical cdf from each study.

Although using pages retrieved from Google rather than pretest ideas to build the prototypical distribution allows for faster, more convenient and automatic processes, it does not come without limitations. In particular, pages are selected by Google to be maximally relevant to the query, i.e., they are likely to be of "high quality." This introduces a risk that ideas with prototypical edge weight distributions are judged as more creative not because of how they balance novelty with familiarity, but because they are "similar" to "high quality" pages retrieved from Google. We address this concern in several ways. First, our first three studies do not rely on Google at all but rather on pretest ideas. Second, in Web Appendix C (subsection "Using the Ideas Themselves to Create the Prototypical Distribution"), we show that our results still hold when the prototypical edge weight distribution is based on the ideas themselves, rather than the Google results used to construct the baseline semantic network. Third, in Section 4.9. (subsection "Vector Space Representation vs. Edge Weight Distributions"), we explore directly whether ideas that are more "similar" to an average Google result in a traditional sense (i.e., they use similar word stems or topics) are indeed judged as more creative. We find that this is not the case.

3.4. Measuring the "Prototypicality" of an Idea's Edge Weight Distribution

The previous subsection described the construction of the prototypical distribution of edge weights. Each idea has its own semantic subnetwork (comprised of a subset of the nodes in the baseline network). This semantic subnetwork results in a distribution of edge weights, where the weight of an edge between two nodes (word stems) in the subnetwork is the same as the weight of the edge between these two nodes in the baseline network. We measure the "prototypicality" of that idea's edge weight distribution by comparing it to the prototypical distribution of edge weights described in the previous section. We use a simple and common measure of the distance between two distributions, the Kolmogorov-Smirnov statistic. The Kolmogorov-Smirnov statistic between two cumulative distributions is defined as the maximum absolute difference between the two distributions. One advantage of this measure, compared to

alternative measures such as the Kullback-Leibler divergence, is that it may be computed for any pair of distributions regardless of their support (we test the robustness of our results to the use of the Kullback-Leibler divergence measure in Web Appendix C). Ideas with semantic subnetworks that have a smaller Kolmogorov-Smirnov statistic have a "more prototypical" edge weight distribution. Conversely, ideas with semantic subnetworks that have a larger Kolmogorov-Smirnov statistic have a "less prototypical" edge weight distribution. It is important to keep in mind that a "prototypical" idea according to this measure does not have prototypical or "average" edge weights, but that the *distribution* of edge weights in the semantic subnetwork corresponding to that idea is similar to the prototypical distribution of edge weights. As can be seen in Web Appendix A, the prototypical distribution contains a whole range of edge weights and "prototypical" ideas have a balance between novelty (coming from the presence of smaller edge weights) and familiarity (coming from the presence of larger edge weights).

3.5.Idea Generation

We collected ideas in various ways across the eight studies, but here we provide an overview of our main approach. In all studies except Study 5, we collect ideas from a panel of consumers using a simple online interface developed by the authors using php (see Figure 2 for an example). The basic interface asks consumers to generate ideas on a specific topic by entering ideas one after the other until they do not wish to contribute more ideas. Ideas were screened manually by the authors in order to remove "junk" ideas that were clearly off-topic or nonsensical. In all studies we remove participants who only submitted "junk" ideas from the analysis.

In Studies 1, 2, 4 and 6, we allow respondents to enter as many ideas as they wish, as long as they enter at least one. In Study 3 we ask consumers to submit exactly three ideas, to reduce variations in the number of ideas across consumers. Study 5 uses secondary data from an online idea generation community and Study 6 uses an interactive aimed at improving the idea generation process on the fly. We describe these approaches in the corresponding sections.

[INSERT FIGURE 2 ABOUT HERE]

3.6. Idea Evaluation

The source of idea evaluations also varied slightly across our eight studies, and we describe here our main approach. In all studies except Study 5, we collected idea evaluations from a set of individuals who were different from those who generated the ideas, but who came from the same panel. This idea evaluation step was performed after all ideas had been collected, using an online interface developed by the authors using php. We follow standard practice (e.g., Kornish and Ulrich 2011; Luo and Toubia 2015; Toubia and Flores 2007) and ask each individual in the idea evaluation sample to evaluate a set of ideas one after the other on several dimensions. The set of ideas rated by each individual were randomly selected among the ideas that had received the fewest number of evaluations up to that point, in order to reduce the variance in the number of evaluations per idea. The average number of raters per idea varied between 18.05 and 26.22 across studies. Each idea was rated by each rater on four dimensions: *creativity* (e.g., "How creative is this app idea?"), *purchase interest* (e.g., "How likely would you be to download this app if it were available for $0.99?"), *predicted popularity* (e.g., "How popular do you think this app would be if it were available for $0.99?"), and *writing quality* (e.g., "Is the description of this app well written?"). Each item had a 5-point Likert scale.

In Study 4, we also collected idea evaluations from experts in our partner company. In Study 5 the evaluations of the ideas came from an online idea generation community.

3.7. Statistical Analysis

In all our studies, we test our hypothesis by regressing the average creativity rating of each idea (or its proportion of positive votes in Study 5) on the prototypicality of its edge weight distribution, controlling for a host of other factors. In all regressions, each observation corresponds to one idea. In Studies 1-4 and 6, we use a linear regression where the dependent variable is the average creativity rating.[3] In Study 5, we run a binomial regression where the dependent variable is the proportion of positive votes. Because ideas

[3] The average creativity rating for each idea is the average of approximately 20 independent evaluations, each of which is on a 5-point Likert scale. We approximate this average as a continuous variable and do not explicitly model the fact that it is truncated.

contributed by the same participant may be more likely to be of similar quality, we control for contributor heterogeneity by including random effects intercepts in all our regressions.

In our regressions, in addition to our primary independent variable (the Kolmogorov-Smirnov statistic between the edge weight distribution of the idea and the prototypical edge weight distribution), we control for the following characteristics of the idea's semantic subnetwork: average edge weight, coefficient of variation of edge weights, minimum edge weight, maximum edge weight, average node frequency, coefficient of variation of node frequencies, minimum node frequency, maximum node frequency, and the number of nodes in the subnetwork. In addition, we control for the length of the idea using its number of characters. It is important to control for the number of nodes and number of characters in the idea as larger semantic subnetworks tend to have smoother distributions of edge weights, which tend to be more prototypical.

In Studies 1a-1c, we also control for variables related to the clustering coefficient: average node clustering coefficient, coefficient of variation of node clustering coefficients, minimum node clustering coefficient, maximum node clustering coefficient. We were not able to control for these variables in the other studies in which the prototypical network was extracted from Google, due to a lack of variation in the clustering coefficients. Indeed, in these studies the network was very dense and almost all clustering coefficients were equal to 1, leading to poorly conditioned regressions.

In our robustness checks, we run additional specifications, accounting for other controls including word stem fixed effects. Finally, ideas with fewer than two nodes (i.e., no edge) in their semantic subnetwork were removed from the analysis.[4]

[4] In Studies 1a-1c where we also control for the clustering coefficient, ideas with fewer than *three* nodes in their semantic subnetwork were removed from the analysis, because at least three nodes are needed to compute the clustering coefficient.

4. Studies

We test our hypothesis and study its managerial implications across eight studies, five of which were run in collaboration with three different companies. Across studies we had over 4,000 ideas generated on 6 different topics by over 2,000 idea contributors. In Studies 1a-1c we test our hypothesis using a baseline semantic network and prototypical distribution obtained from a pretest. Study 2 replicates our finding using Google instead of a pretest. We adopt Google in all subsequent studies for its convenience. In Study 3 we ask each respondent to generate exactly three ideas, in order to reduce the variance in the number of ideas across contributors. In Study 4 we complement our consumer evaluations with company evaluations. In Study 5 we test our hypothesis in a typical managerial context, by using a secondary data set coming from an online idea generation community. In Study 6, we show how our findings may be used to help people generate better ideas. We develop and test a tool that leverages our findings to recommend words to consumers on the fly in order to help them improve their ideas. See Table 1 for an overview of our studies.

[INSERT TABLE 1 ABOUT HERE]

4.1.Studies 1a-1c

Method

Studies 1a-1c were conducted in collaboration with a large US-based insurance company that was looking for innovative ideas for new insurance products. The three studies were similar to each other in design and only differed in their idea generation topics. Participants in these three studies were recruited from Amazon Mechanical Turk. Participants were asked to generate ideas for new insurance products related to aging and being a senior (Study 1a), financial security (Study 1b), or unemployment (Study 1c).[5]

[5] In order to help participants structure their ideas and increase their relevance to the company, participants were asked to list three components in each insurance product idea: what may be lost by the customer, what the customer would get if the loss occurred, and what the customer had to give in exchange for this protection.

Before running these studies, we conducted a pretest for each study in which participants were asked to generate initial ideas on the topic. The number of participants in the pretest was 149, 101 and 98 for Studies 1a, 1b and 1c, respectively, and the number of initial ideas obtained was 447, 303, and 294, respectively. The baseline semantic network for each study was constructed as described in Section 3.1. The ideas from the pretest were not used in any other part of the analysis. The baseline semantic networks contained 314, 175, and 184 nodes in Studies 1a, 1b, and 1c, respectively.

After removing "junk" ideas and ideas with semantic subnetworks that had fewer than three nodes (in order to calculate clustering coefficient metrics), we were left with 276, 271 and 251 ideas from 178, 177 and 167 participants, respectively. The idea evaluation stage resulted in an average number of evaluators per idea equal to 18.05, 21.62 and 20.91 across studies (standard deviations of 0.59, 0.64 and 0.46, respectively).

Results and discussion

Descriptive statistics may be found in Web Appendix A, which shows the distribution of the size of the ideas' semantic subnetworks (i.e., number of nodes) as well as the distribution of prototypicality across ideas (i.e., the Kolmogorov-Smirnov statistic between the idea's edge weight distribution and the prototypical distribution). The statistical analysis of the link between prototypicality and judged creativity is reported in the 1st to 3rd columns of Table 2. The coefficient for prototypicality is negative and statistically significant in all three studies ($p<0.05$). That is, ideas with semantic subnetworks that have an edge weight distribution closer to the prototypical distribution, are judged as significantly more creative. Therefore, the results of these first studies are consistent with our hypothesis.

[INSERT TABLE 2 ABOUT HERE]

4.2. Study 2

Method

Study 2 replicates Study 1, using Google instead of a pretest to construct the baseline semantic network and the prototypical edge weight distribution. The baseline semantic network contained 485 nodes.

Participants in the idea generation and idea evaluation tasks were again recruited from the Amazon Mechanical Turk panel. Participants in our idea generation task were asked to generate ideas for new smartphone apps that will help their users keep a healthier lifestyle. Each participant received $1 as compensation. After removing "junk" ideas and ideas with semantic network that had fewer than two nodes (i.e, no edge), we were left with 555 ideas generated by 300 participants. A different group of 1,209 Amazon Mechanical Turk participants evaluated these ideas as described above, and were paid $0.50 each for their participation. Each participant evaluated 10 ideas, giving rise to an average of 20.31 evaluators per idea (standard deviation = 1.34).

Results and Discussion

As can be seen in Table 2, the coefficient for the prototypicality of the edge weight distribution is negative and statistically significant. Hence the results of Study 2 replicated those of Study 1a-1c in a different ideation domain. Moreover, this study suggests that the results are robust to the way the baseline semantic network is constructed, such that this network may be constructed based on an initial set of ideas coming from a pretest, or publicly available text such as webpages identified by Google. We adopt the latter approach throughout the rest of the paper, for its convenience.

Although we control for heterogeneity across participants in their ability to generate creative ideas using random effects, there is also heterogeneity in the number of ideas generated by participants and therefore some participants contribute more than others to the results. We address this concern in the next study.

4.3.Study 3

Method

The design of Study 3 was identical to that of Study 2, except that participants were forced to generate three ideas each. The idea generation topic and the baseline semantic network were identical to Study 2. Amazon Mechanical Turk panel members completed the idea generation task for $1 each. After removing "junk" ideas and ideas with semantic subnetworks that had fewer than two nodes, we were left with 173

ideas from 61 participants. A different group of Amazon Mechanical Turk participants evaluated these ideas, giving rise to an average of 20.53 evaluators per idea (standard deviation = 0.78).

Results and discussion

The results of our main regression are reported in the fifth column of Table 2. We see that the coefficient corresponding to prototypicality remains negative and statistically significant. Therefore Study 3 provides further replication of our main finding, keeping constant the number of ideas per participant.

4.4.Study 4

Method

Study 4 was conducted in collaboration with an international health and beauty company that was looking for ideas for new oral care solutions targeted to women over 40 years old. The idea generation topic was: "What new product could help women maintain healthy and beautiful oral features?" The baseline semantic network was constructed again by copying and pasting this idea generation topic into Google and mining the page source code of the top 50 search results. The resulting baseline semantic network contained 280 nodes.

This study differed from the previous ones in two major ways. First, ideas were evaluated by company experts, in addition to consumers. Second, participants were recruited from a commercial consumer panel maintained by Research Now, instead of Amazon Mechanical Turk.[6] Interestingly, compared to the Amazon Mechanical Turk participants in Studies 1, 2, and 3, the commercial panel participants generated fewer and shorter ideas (1.350 ideas vs. 1.646 in Amazon Mechanical Turk with an average of 85.7 characters vs. 300.0 characters on average in Amazon Mechanical Turk). After removing "junk" ideas, we were left with 220 ideas from 163 participants. The idea evaluation stage resulted in an average of 26.22 evaluators per idea (standard deviation = 1.13).

[6] Both for idea generation and for idea evaluation, respondents were screened to only include women over 40 years old who brushed their teeth at least once a day, had visited a dental professional at least once in the last two years, and suffered from at least one aging-related oral symptom from a list specified by the company.

In addition, the ideas were carefully evaluated by a group of experts from the company. These judges applied a screening process developed internally and reached a consensus on each idea through deliberation. The experts selected ideas that were on topic, addressed unsatisfied needs, and that were consistent with the company's strategy. The expert selection of the ideas was independent of our text mining analysis of the ideas.

Results and discussion

We first analyze the ideas based on the consumer evaluations. The results are reported in the 6th column of Table 2. Consistent with our hypothesis, the coefficient corresponding to prototypicality is marginally significant ($p<0.08$).

We now turn to the analysis of the company's evaluation of the ideas. Eighty nine out of all 220 ideas for which a prototypicality measure was available passed the company screening as being on topic, addressing an unsatisfied need and being consistent with the company's strategy. We find that the prototypicality of these 89 ideas was significantly higher compared to the ideas that were not selected by the firm's experts. Specifically, the distance to the prototypical distribution of edge weights (measured by the Kolmogorov-Smirnov statistic) was significantly lower for the ideas selected relative to the ideas not selected (means of 0.459 vs. 0.545, $p<0.01$).[7]

Therefore, Study 4 suggests that our results extend to evaluations that are not specifically limited to consumers evaluating the creativity of ideas, and that come from practitioners who are experts in product innovation. In addition, it shows that our results still hold when both idea generators and evaluators are selected from a commercial panel rather than Amazon Mechanical Turk.

[7] In Web Appendix B we distinguish between precision and recall using a Receiver Operating Characteristic (ROC) curve analysis. This analysis further confirms that our classification of ideas based on prototypicality (Kolmogorov-Smirnov statistic) is adequate.

4.5. Study 5

Method

Study 5 complements the previous studies by testing whether our findings apply in a typical online idea generation context. In practice, idea generation is often performed through online idea generation communities, such as the well-known My Starbucks Idea or Dell's Idea Storm. Instead of collecting new ideas experimentally like in the other studies, in this study we received secondary field data from an actual online idea generation community focused around Pro Tools, a digital audio workstation. Members of the online idea generation community submit new ideas that would improve the product, and evaluate ideas submitted to the community. Idea evaluation takes the form of binary votes ("thumbs up – I agree" vs. "thumbs down – I disagree"). Users may generate as many ideas as they wish and vote on as many ideas as they wish (although each user cannot evaluate the same idea multiple times and cannot vote on their own ideas). The company that manages and hosts this community made the data related to the ideas and their evaluations available to us. Our analysis focuses on the 1,735 ideas submitted by users in 2010, 2011 and 2012 that received at least one vote and that have semantic subnetworks with at least two nodes. The average number of votes per idea is 28.34. Because ideas that have received the most votes tend to be featured more prominently in the community (a common practice in online ideation communities), the standard deviation of the number of votes per idea is large, and equal to 49.11. Overall, 84.25% of the votes are positive.

Our baseline semantic network for this study was constructed based on Google (the text of the query was "Pro Tools"), and had 455 nodes. In order to assess whether the company hosting the community would be able to leverage our findings systematically and automatically, we did not go through the list of word stems manually when constructing the baseline semantic network (see Section 3.1). Similar results were obtained when this manual cleaning stage was applied (details available from the authors).

Results and discussion

Our statistical analysis in this study differs slightly from the other studies, given the nature of the evaluations. Instead of running a linear regression based on the average ratings across evaluators, we run a binomial regression based on the number of votes for each idea and the proportion of positive votes. We assume a logistic link between the proportion of positive votes and the independent variables and allow the residuals to be correlated between ideas submitted by the same user. The results are presented in the seventh column of Table 2. We see that the coefficient corresponding to prototypicality is statistically significant at $p<0.05$.

Therefore, this study further confirms our results using secondary field data coming from a popular form of idea generation, online idea generation communities. It also replicates our results with a much larger set of ideas than the ones used in the previous studies. Moreover, it suggests that our hypothesis still holds when the text mining process and measurement of prototypicality are completely automated and do not rely on any human input. Thus, our research provides firms hosting idea generation communities with a "free" measure of idea quality, which may be combined with other measures based on human judgment. With the advent of online ideation communities such as the one we studied here, the challenge of effectively screening a large number of ideas is more relevant today than ever (Simon 2014). We would not recommend making a final selection of ideas based on prototypicality only. Rather we envision our research being used in a first round of screening that flags a set of ideas worth considering carefully.

Our results so far have confirmed our main hypothesis that ideas with semantic subnetworks that have a more prototypical edge weight distribution tend to be judged as more creative. The results hold whether the baseline semantic network is constructed based on ideas from a pretest or based on web pages related to the topic. The results do not seem to be driven by differences in the quantity of ideas across consumers (Study 3), by whether the evaluations are performed by consumers or company experts (Study 4), and the results hold in field data coming from an online idea generation community (Study 5). Moreover, the results seem to extend to alternative measures of idea quality (company selection in Study

4 and votes from the community in Study 5). Our next and final study will further explore the practical implications of our main hypothesis. Before describing it, we first describe a set of robustness checks, explore alternative measures of fit, explore the extent to which our hypothesis applies to alternative dimensions of idea quality, and show various boundary conditions.

4.6. Robustness Checks

Web Appendix C reports a series of robustness checks, which we briefly summarize here.

-To test whether creativity is driven by a set of word stems that are considered creative, we include fixed effects in the regression for the most commonly used word stems. We find that our results are robust to the introduction of these fixed effects, despite the reduction in statistical power.

-We find that our results in Studies 2-6 are robust to using the ideas submitted by participants to create the prototypical edge weight distribution, instead of using pages retrieved from Google. This helps address the concern that ideas with prototypical edge weight distributions might be judged as more creative only because they are "similar" to pages selected by Google for their attractive properties, not because of their edge weight distribution per se. Using the ideas submitted by participants to create the prototypical edge weight distribution also makes it possible to measure prototypicality using the Kullback-Leibler divergence instead of the Kolmogorov-Smirnov statistic. We find that our results still hold with this alternative measure.

-We test for possible asymmetry in the effect of prototypicality, by using a signed measure of the Kolmogorov-Smirnov statistic that captures whether the edge weight distribution corresponding to each idea is above (Kolmogorov-Smirnov>0) or below (Kolmogorov-Smirnov<0) the prototypical distribution at the point at which the two distributions are maximally distant. We find the same effect of prototypicality on judged creativity for ideas whose distributions are above the prototypical distribution (i.e., more small weights and therefore more novel combinations) as well as for ideas whose distributions are below the prototypical distribution (i.e., more large weights and therefore more familiar combinations), although the effect is stronger for the former type of ideas.

-We find that our results still hold when edge weights are measured using the Salton Cosine (Salton and McGill 1983) instead of the Jaccard index.

-We test whether our results in Studies 2-6 are robust to reducing the number of top search results from Google used to construct the baseline semantic network and prototypical edge weight distribution. While in some studies the effect of prototypicality is significant with as few as 20 pages, we recommend mining at least 50 pages to obtain robust and significant effects.

-We find that our results are robust to an alternative regression specification that removes the average of the edge weight, node frequency, and node clustering coefficient distributions, which is likely to be highly correlated with the sum of the minimum and the maximum.

-We test whether our results are impacted by the use of synonyms, by identifying word stems that are synonyms and combining them in our analysis. Our conclusions are unchanged.

-We find that our results in Studies 1-4 and 6 still hold when the dependent variable is the proportion of creativity ratings of 4 or 5 out of 5 received by each idea, rather than the average creativity rating. This addresses the concern that ideas with an average edge weight distribution might present a "compromise" that is judged as more creative on average, but that is not necessarily seen as creative by many judges.

4.7.Alternative Measures of the Relationship between Prototypicality and Judged Creativity

One of the practical implications of our research is helping companies identify promising ideas from a large set of ideas without the need for any human involvement. To shed more light on the ability of our approach to identify promising ideas, we look at the rank-order correlation between the fitted and the observed creativity ratings of ideas based on the regressions from Table 2. The average correlation, across studies, is r=0.44 ($p<0.001$). See Web Appendix D for details. This analysis provides additional support for the use of our research as a tool for flagging ideas that are worth considering carefully.

4.8. Alternative Measures of Idea Quality

Our analysis so far has focused primarily on the judged creativity of ideas, with the exception of the company expert evaluations in Study 4 and the binary votes from online community members in Study 5.

In all studies (except Study 5), all ideas were rated on four dimensions: purchase interest, predicted popularity, writing quality, and creativity. We explore the use of alternative measures of idea quality as dependent variables, and tests whether the effect found on judged creativity is mediated by any of these alternative measures. See Web Appendix E for details.

These analyses suggest that while other measures of idea quality are also related to the prototypicality of the edge weight distribution, the relationship is strongest for judged creativity. Furthermore, our alternative measures of idea quality do not mediate the relationship between prototypicality and judged creativity, providing empirical support for the use of creativity as the dependent variable. This is consistent with our theoretical development from Section 2, which relied specifically on the link between creativity and the balance of novelty vs. familiarity.

Of particular interest is the use of writing quality as the dependent variable. Our results suggest that prototypicality has a positive effect on the judged writing quality of an idea (which does *not* mediate the effect on judged creativity). This finding may be relevant to the literature on automated essay scoring (e.g., Attali and Burstein 2006; Landauer, Laham and Foltz 2003), which is very relevant to online academic testing (e.g., GRE, GMAT). While the algorithms used by companies such as ETS are proprietary and not fully public, to the best of our knowledge this literature has not considered using the prototypicality of the structure of an essay's semantic network as a measure of writing quality.

4.9.Boundary Conditions

Alternative Measures of Prototypicality

We have argued, based on the creativity literature, that an appropriate measure for prototypicality in the context of idea generation is one that captures the distribution of edge weights, thereby quantifying the balance between novel and familiar combinations of word stems. Here we test some boundary conditions of our results by measuring prototypicality based on the distribution of two other popular network features: node frequency and clustering coefficient. We construct these two alternative prototypicality measures using the approach described in sections 3.3 and 3.4.

Results are provided in Web Appendix F. When prototypicality is measured based on the distribution of node frequency, the coefficient corresponding to prototypicality is directionally consistent with the hypothesis in 6 out of 8 studies, but significant at $p<0.05$ in only one of them. When prototypicality is measured based on the distribution of the clustering coefficient (which we are able to do in Studies 1a-1c only), the coefficient is actually in the opposite sign, significantly so in one study. Also, the fit in these regressions is worse compared to the regressions in Table 2. These analyses confirm our theoretical argument that prototypicality should be measured in a way that captures the relationships among the word stems present in the ideas, as well as the tradeoff between familiarity and novelty.

Vector Space Representation vs. Edge Weight Distributions

The previous subsection explored alternative ways to measure prototypicality given a baseline semantic network and a set of ideas. In this subsection we explore the relevance of using a semantic network in the first place. The concept of a semantic network is central to our theoretical argument because it captures the balance between novelty and familiarity. We compare it to a more direct approach inspired by analogies with the Information Retrieval literature.

Indeed, our approach may be compared and contrasted with a traditional Information Retrieval model, where our idea generation topic would be equivalent to a query, and our goal would be to assess which documents (i.e., ideas) are "relevant" to that query. Our approach compares documents to a prototypical distribution derived from a set of training documents related to the query (pretest ideas or Google results). A standard approach for making this comparison would be to represent documents as vectors of word stems and compute the distance between vectors corresponding to various documents, similar to the standard Rocchio classifier (Feldman and Sanger 2007, page 74).

To test such an alternative approach, we represent each document as a vector with dimensionality equal to the number of word stems in our dictionary (i.e., number of nodes in our semantic network). We use a standard term frequency – inverse document frequency (*tf-idf*) approach (see for example Manning, Raghavan and Schutze 2008). We measure prototypicality for a given idea using the Euclidean distance

between the vector representing that idea and the average vector among training documents. See details of this analysis and the results in Table F3 in Web Appendix F. We find that measuring the prototypicality of an idea using the distance between this idea and an average document does not give rise to a robust significant link between prototypicality and judged creativity. In fact, in all studies the coefficient associated with the distance to the prototypical document is positive (it is statistically significant at $p<0.05$ in three studies and at $p<0.10$ in two). That is, ideas that are *further away* from a prototypical document in a vector space representation tend to be judged as more creative.

We also explore representing documents by *topics* rather than actual words. We perform Latent Dirichlet Allocation (LDA) on each set of training documents (pretest ideas or Google results) to identify a set of topics and associated words (Blei et al., 2003; Tirunillai and Tellis, 2014). Details of the LDA estimation are provided in Web Appendix F. Each idea is represented as a vector with dimensionality equal to the number of topics. We compute the Euclidean distance between the vector representing each idea and the average vector from the training documents. Results of the regressions are presented in Table F4 in Web Appendix F. Again, we find no significant robust relationship between distance and judged creativity.

This analysis underscores the importance of defining prototypicality with respect to the balance between novel and familiar combinations of word stems, which calls for a semantic network. This analysis also informs the potential concern that ideas with prototypical edge weight distributions are judged as more creative only because they are "similar" to pages selected by Google for their attractive properties, not because of their edge weight distribution per se. In particular, the results suggest that ideas that are more "similar" to an average Google result in a traditional sense (i.e., they use similar word stems or similar topics) in fact tend to be judged as less creative.

Misspecification of the Baseline Semantic Network

We have argued that the baseline semantic network and the prototypical edge weight distribution should be specific to each idea generation topic. Here we explore the consequences of using a baseline semantic

network and prototypical edge weight distribution from a different idea generation topic. Studies 1a-1c were all related to insurance, but each study focused on a different insurance domain: aging and being a senior (Study 1a), financial security (Study 1b), and unemployment (Study 1c). This provides us with an opportunity to explore situations where the baseline semantic network and its corresponding prototypical edge weight distribution come from a domain that is related but different from the idea generation topic being considered. For each of these studies, we replicate our analysis using the baseline semantic network and prototypical distribution from the two other studies. See details of this analysis and results in Web Appendix F. We find that the relationship between prototypicality and judged creativity is neither consistent nor significant when the baseline semantic network (and its corresponding prototypical edge weight distribution) is taken from a different, albeit related, ideation topic. This underlines the need to construct baseline semantic networks and prototypical edge weight distributions that are specific to each idea generation topic. Luckily, this may be done efficiently and with no need for incremental human labor, using Google.

4.10. Study 6

The previous studies have demonstrated the link between the prototypicality of the edge weight distribution of an idea's semantic subnetwork and the judged creativity of the idea. One first practical implication of this finding is that it provides firms with an automatic measure that may be used to identify promising ideas, thereby reducing the costs involved in idea screening. In our final study, we explore a second practical implication. In particular, we explore leveraging our finding to help people improve the creativity of their ideas. We develop an online idea generation tool in which participants enter their ideas, and sets of words are suggested to them on the fly to help them improve their ideas. We compare the judged creativity of ideas when we recommend words to users that would improve the prototypicality of their ideas' edge weight distribution vs. words based on other criteria vs. when no recommendations are made. This study presents a proof of concept of using "Big Data" tools to foster creativity.

Method

We used the same idea generation topic (smartphone apps that would help their users be healthier), baseline semantic network, and prototypical edge weight distribution as in Studies 2 and 3. In the idea generation phase of the study, we assigned participants randomly to one of four conditions. In all conditions, participants navigated between two types of interfaces, coded in php: an idea collection interface and an idea modification interface. The idea collection interface looked similar to the interface used in Studies 1-4. It gave participants the opportunity to submit new ideas that were not related to any of their previous ideas. This interface was identical across conditions. The idea modification interface appeared after the submission of each idea, giving participants the opportunity to modify/improve the idea they had just submitted. On the idea modification interface, a participant could either submit a modified version of their last idea (based on a set of suggested words when applicable), or indicate that they had no more modification to make and go back to the idea collection interface. The idea modification interface always loaded with the response box pre-populated with the last idea submitted by the participant, to make it easier for participants to modify this idea. This process was repeated until the participant stated they had no more idea to contribute. Screenshots are provided in Figure 3. In both types of interfaces and in all conditions, a log of the ideas submitted by that participant up to that point was provided at the bottom of the screen.

In the *Control* condition, the idea modification interface simply invited participants to modify/improve their last idea ("Please modify/improve your idea. If you do not wish to improve your previous idea, please select 'I am done with this idea.'). See the middle panel of Figure 3.

In the other three conditions (*Random Words*, *Minimum Distance* and *Maximum Prototypicality*), the idea modification interface showed groups of words selected to help participants improve their last idea. Each group of words corresponded to one node (word stem) in the baseline semantic network, e.g., the words corresponding to the stem "electronic" were "electronically, electronic, electronics." A set of 10 word stems was selected for each new idea. Participants could cycle through the 10 word stems at will, and modify their ideas with or without using the suggested words. See bottom panel of Figure 3. The

only difference between the *Random Words*, *Minimum Distance* and *Maximum Prototypicality* conditions was the way the set of 10 nodes was selected. Each idea was text mined upon being submitted by a participant and the semantic subnetwork corresponding to that idea was constructed. All computations in all conditions were completed on the fly with no noticeable delay.

In the *Random Words* conditions, 10 nodes were randomly selected for each idea among those that were in the baseline semantic network but not used in the idea. For example, if an idea's semantic subnetwork contained 15 nodes and if the baseline semantic network contained 485 nodes (as was the case in our study), the 10 nodes were randomly selected without replacement from the 470 nodes that were not already part of the idea's subnetwork.

In the *Minimum Distance* condition, the distribution of edge weights in the idea's semantic subnetwork was computed, and a score for each potential new node was computed, equal to the average edge weight that would result from adding this node to the subnetwork. Consider again our example with 15 nodes in the idea's semantic subnetwork and 485 nodes in the baseline semantic network. For each of the 470 nodes that are not part of the subnetwork, we would compute the average of the $\binom{16}{2}$ edge weights in the new subnetwork that would result from adding this new node to the current subnetwork. The 10 nodes selected using this rule would maximally increase the average edge weight in the idea's semantic subnetwork, i.e., decrease the average distance between the nodes. The idea behind this rule is to suggest word stems that are most closely related to the words already used in the idea. We expected this selection rule to make it easy for participants to modify their ideas, but that these modifications would not necessarily improve the idea's creativity because the relationship may be too obvious or too familiar.

In the *Maximum Prototypicality* condition, the score for each potential new node was equal to the prototypicality (1-Kolmogorov-Smirnov statistic) of the edge weight distribution that would result from adding this node to the idea's current subnetwork. In our previous example, for each of the 470 nodes that are not currently in the network, we would compute the Kolmogorov-Smirnov statistic for the distribution of edge weights that would be obtained by adding this node to the current subnetwork. The 10 nodes

selected using this rule would maximally increase the prototypicality of the idea's edge weight distribution. We expected this selection rule to give rise to sets of words that would best allow participants to improve their ideas.

Amazon Mechanical Turk Participants completed the idea generation task in exchange for $1. After removing "junk" ideas as well as participants who only entered "junk" ideas, we were left with respectively 100, 100, 98 and 95 participants in the *Control*, *Random Words*, *Minimum Distance*, and *Maximum Prototypicality* conditions.[8] Idea evaluation was performed similarly to the other studies. A different group of 2,000 participants from Amazon Mechanical Turk evaluated (both the original and the modified) ideas in exchange for $0.50. Each idea received an average of 20.43 evaluations (standard deviation = 0.53).

[INSERT FIGURE 3 ABOUT HERE]

Results

We classify ideas into two types based on how they were submitted: "original ideas" are those submitted in the idea collection interface, and "modified ideas" are those submitted in the idea modification interface, i.e., they are modified versions of a previous idea.

First, for the "original" ideas, pooled across conditions, we replicate our findings from the previous studies using the same of set of regressions as earlier (see the 8th column in Table 2 and the tables in the various web appendices). We limit the analysis to original ideas in order to ensure statistical independence between ideas from the same author. The same conclusions are reached if we include all ideas in the regressions.

Next, we turn to the comparison between conditions. For each participant, we compute the number of "original" ideas and the average number of "modified" ideas per "original" idea. For each "original" idea that was modified at least once, we compute the difference between the judged creativity of its last

[8] Note that the four conditions had identical interfaces until after the submission of the participant's first idea. Therefore it is unlikely that some conditions made participants more likely to submit only "junk" ideas.

modification vs. the original idea (i.e., if an idea was modified three times we compare the judged creativity of the last idea in that stream to that of the original idea). The results are reported in Table 3 (more detailed analyses can be found in Web Appendix G). We find that the *Maximum Prototypicality* condition is the only one that gives rise both to a significantly greater propensity to modify ideas compared to the *Control* condition, and to modifications that are significant improvements over the original ideas. The *Random Words* condition did not significantly increase participants' propensity to modify their ideas. The *Minimum Distance* condition significantly helped participants *modify* their ideas (compared to the *Control* condition), but the modified ideas were not significantly better than the original ones.

We also explore asymmetries in the results using a signed Kolmogorov-Smirnov statistic. We separate the original ideas that were modified at least once between those with a positive Kolmogorov-Smirnov statistic (i.e., more small weights and therefore more novel combinations) vs. negative Kolmogorov-Smirnov statistic (i.e., more large weights and therefore more familiar combinations). We find that the *Maximum Prototypicality* condition worked primarily by helping participants with ideas that were too familiar increase the novelty of their ideas, but that participants with ideas that were too novel were not able to increase familiarity in a meaningful way using the suggested words. See detailed results in Web Appendix G.

[INSERT TABLE 3 ABOUT HERE]

Discussion

Study 6 not only replicated the findings from the other studies, it also demonstrated that the link between prototypicality and creativity may be leveraged in practice to create tools that help people improve their ideas. Generating new ideas involves retrieving knowledge from memory. We have shown that it is possible to use computers to assist people in this memory retrieval process, by developing an online interface that provides participants on the fly with possible words that may help them improve their ideas.

The tool we developed here is a proof of concept. We have developed a publicly available version of this tool, available at newtopic.protoideation.org.[9] We hope that future research will develop more sophisticated and powerful tools. For example, with access to individual-level data, it would be possible to build individual-specific baseline semantic networks based on the documents to which a particular individual was exposed in the past (such data are available to companies that track user behavior online). We could envision an online tool similar to Google in which a user would enter a problem they wish to solve or a topic on which they wish to ideate, and the tool would provide them with a customized set of possible words that could be basic ingredients to a solution, or a set of documents that are likely to contain useful information.

The tool we developed in this study may be viewed as an extension of the popular "random stimulation" technique developed by De Bono (De Bono, 1992). De Bono's method consists in drawing random words one at time and attempting to generate new ideas based on these words. Interestingly, De Bono writes (p. 182): "How do we find the 'best' random words? The simple answer is you cannot… There is no way of finding the 'best' random word because it would then no longer be random." Our research suggests that the words used as inspiration may in fact be "optimized," and that selecting words that will help users improve the prototypicality of their ideas' semantic subnetworks is more efficient than showing them random words.

5. Conclusions

In this paper we have uncovered and documented what appears to be a robust, fundamental property of creative ideas. We have shown ideas that balance well familiarity and novelty, as measured by the combination of "ingredients" in the idea, are judged as more creative. More specifically, ideas that are more prototypical in terms of the edge weight distribution of their semantic subnetwork tend to be judged as more creative. We have demonstrated the link between prototypicality and judged creativity across

[9] The development of this publicly available version was made possible by a generous grant from the Marketing Science Institute. Readers should contact the authors directly with questions or requests about this tool.

eight studies in which over 2,000 people generated over 4,000 ideas in total. Five of our studies were run in collaboration with companies. Across studies, we have varied the source of participants, the format of the idea generation task, the idea generation topic, the type of evaluations and the source of these evaluations. We have also used both primary and secondary sources of data. Managerially, we have shown that our findings can be leveraged not only to identify promising ideas automatically, but also to develop tools that can help people improve their idea generation output by proposing words that may serve as "ingredients" for their ideas.

We believe that many exciting opportunities for future research may be identified, in addition to those already mentioned throughout the paper. First, driven by our theoretical development and our need to capture the co-occurrence of word stems, we mapped ideas onto semantic networks. However, this approach does not capture *how* words are combined, and it does not allow interpreting ideas. Future research might extend the analysis in such directions. Second, future research may explore the extent to which our findings apply both to incremental and radical innovations. Although ideas in our studies were evaluated both by consumers and experts, they were all generated by consumers, and therefore may have been skewed towards incremental innovations. The literatures on which our theoretical argument is built have heavily focused on creativity in the domains of science and art, which one may argue are the bedrock of radical innovations. For example, the issue of balancing novelty with familiarity has been studied in the history of science literature, the literature on the associative nature of creativity was inspired by prominent scientists and artists, and the beauty-in-averageness effect has been found in various artistic domains. Therefore we expect our findings to generalize to ideas generated by professionals searching for radical innovation opportunities. Third, prototypicality may be considered as a new metric in the automated evaluation of other types of textual data, such as essays (e.g., Attali and Burstein 2006; Landauer, Laham and Foltz 2003), movie scripts (Eliashberg, Hui and Zhang 2007), or academic articles (Uzzi et al. 2013). Fourth, the insights and tools from our research can be applied to the domain of recommendation systems. For example, it might be possible to identify products that best complement the set of products a consumers already owns, based on the properties of the subnetwork

formed by these products (e.g., identifying which new book would best complement the user's personal library based on the properties of the network of books in her library, Oestreicher-Singer et al. 2013). Similar recommendations may be made in the domain of scientific citations (e.g., identify a set of papers that would best complement the set of papers already cited in one's manuscript). Finally, this paper provides one example of exploring the use of Big Data tools in new ways that may have a positive impact on people's lives and on society. A large proportion of the information to which we are exposed today is recorded electronically. This information is often used by marketers to target advertising and other marketing vehicles. We hope that our research will help open the door for new applications of these data that may offer new benefits to users.

References

Anderson, JR. 1983. "A Spreading Activation Theory of Memory." *Journal of Verbal Learning and Verbal Behavior*. 22: 261-295.

Attali, Y and Burstein, J. 2006. "Automated Essay Scoring with e-Rater® V.2." *Journal of Technology, Learning, and Assessment.* 4(3).

Barrat, A, Barthélemy, M, Pastor-Satorras, R, and Vespignani, A. 2004. "The Architecture of Complex Weighted Networks." *Proceedings of the National Academy of Science*. 101(11): 3747-3752.

Blei, DM, Ng, AY, and Jordan, MI. 2003. "Latent Dirichlet Allocation." Journal of Machine Learning. 3. 993-1022.

Burroughs, JE, Moreau, CP, and Mick, DG. 2008. "Toward a Psychology of Consumer Creativity." in *Handbook of Consumer Psychology*.

Callon, M, Law, J, and Rip, A. 1986. *Mapping the dynamics of science and technology: sociology of science in the real world.* Houndmills, Basingstoke: The Macmillan Press Ltd.

Collins, AM and Loftus EF. 1975. "A Spreading Activation Theory of Semantic Processing." *Psychological Review*. 82(6): 407-428.

Dahl, DW, and Moreau, P. 2002. "The Influence and Value of Analogical Thinking During New Product Ideation." *Journal of Marketing Research*. 49 (February). 47-60.

De Bono, E. 1992. *Serious Creativity*. HarperCollins: New York, NY.

Eliashberg, J, Hui, SK, and Zhang, ZJ. 2007. "From Story Line to Box Office: A New Approach for Green-lighting Movie Scripts." *Management Science*. *53*(6): 881-893.

Feinerer, I, Hornik, K, and Meyer, D. 2008. "Text Mining Infrastructure in R." *Journal of Statistical Software.* 25(5): 1-54.

Feldman, R, Fresko, M, Kinar, Y, Lindell, Y, Liphstat, O, Rajman, M, Schler, Y, and Zamir O. 1998. "Text Mining at the Term Level." in *Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery:* Springer-Verlag. 65-73.

Feldman, R, and Sanger J. 2007. *The Text Mining Handbook*. Cambridge University Press, Cambridge, UK.

Finke, RA, Ward, TB, and Smith, SM. 1992. *Creative Cognition*, MIT Press: Cambridge, MA.

Giora, R. 2003. *On Our Mind: Salience, Context and Figurative Language*. New York: Oxford University Press.

Goldenberg, J, and Mazursky, D. 2002. *Creativity in Product Innovation*. Cambridge University Press.

Grammer, K and Thornhill, R. 1994. "Human (*Homo sapiens*) Facial Attractiveness and Sexual Selection: The Role of Symmetry and Averageness." *Journal of comparative psychology*. 108(3): 233-242.

Halberstadt, J and Rhodes, G. 2003. "It's Not Just Average Faces that are Attractive: Computer-Manipulated Averageness makes Birds, Fish, and Automobiles Attractive." *Psychonomic Bulletin & Review.* 10(1): 149-156.

Kornish, L and Ulrich, K. 2011. "Opportunity Spaces in Innovation: Empirical Analysis of Large Samples of Ideas." *Management Science.* 57(1): 107-128.

Kostoff, RN. 2006. "Systematic acceleration of Radical Discovery and Innovation in Science and Technology." *Technological Forecasting and Social Change*. *73*(8): 923-936.

Landauer, T, Laham, D, and Foltz, P. 2003. "Automated Essay Assessment." *Assessment in Education*. 10(3): 295-308.

Landwehr, JR, Labroo, AA, and Herrmann, A. 2011. "Gut Liking for the Ordinary: Incorporating Design Fluency Improves Automobile Sales Forecasts." *Marketing Science*. 30(3): 416-429.

Langlois, JH and Roggman, LA. 1990. "Attractive Faces are Only Average." *Psychological science*. 1(2): 115-121.

Luo, L and Toubia, O. 2015. "Improving Online Idea Generation Platforms and Customizing the Task Structure based on Consumers' Domain Specific Knowledge." *Journal of Marketing*.

Manning, CD, Raghavan, P, and Schűtze H. 2008. *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Martindale, C, Moore, K, and Borkum J. 1990. “Aesthetic Preference: Anomalous Findings for Berlyne's Psychobiological Theory.” *The American Journal of Psychology*. 103(1): 53-80.
Martindale, C, Moore, K, and West, A. 1988. “Relationship of Preference Judgments to Typicality, Novelty, and Mere Exposure.” *Empirical Studies of the Arts*. 6(1): 79-96.
Moreau, CP, and Dahl, DW. 2005. “Designing the Solution: The Impact of Constraints on Consumers’ Creativity.” *Journal of Consumer Research*. 32(1): 13-22.
Mednick, SA. 1962. “The Associative Basis of the Creative Process.” *Psychological Review*. 69(3): 220-232.Moreau, CP, and Dahl, DW. 2005. “Designing the Solution: The Impact of Constraints on Consumers’ Creativity.” *Journal of Consumer Research*. 32(1): 13-22.
Netzer, O, Feldman, R, Goldenberg, J, and Fresko, M. 2012. “Mine Your Own Business: Market-Structure Surveillance Through Text Mining.” *Marketing Science.* 31(3): 521-543.
Nijstad, BA, Stroebe, W, and Lodewijkx, HFM. 2003. “Production Blocking and Idea Generation: Does Blocking Interfere with Cognitive Processes?” *Journal of Experimental Social Psychology*. 39(6): 531-548.
Nijstad, BA, and Stroebe, W. 2006. “How the Group Affects the Mind: A Cognitive Model of Idea Generation in Groups.” *Personality & Social Psychology Review*. 10(3): 186-213.
Oestreicher-Singer, G, Libai, B, Sivan, L, Carmi, E, and Yassin, O. 2013. The Network Value of Products. *Journal of Marketing*. 77(3): 1-14.
Perkins, DN. 1981. *The Mind’s Best Work*. Harvard University Press: Cambridge, MA.
Porter, MF. 1980. “An Algorithm for Suffix Stripping.” *Program*. 14(3): 130–137.
Reber, R, Schwarz, N, and Winkielman, P. 2004. “Processing Fluency and Aesthetic Pleasure: is Beauty in the Perceiver's Processing Experience?” *Personality and social psychology review*. 8(4): 364-382.
Repp, BH. 1997. “The Aesthetic Quality of a Quantitatively Average Music Performance: Two Preliminary Experiments.” *Music Perception: An Interdisciplinary Journal*. 14(4): 419-444.
Rothenberg, A. 2015. *Flight from Wonder – An Investigation of Scientific Creativity*. Oxford University Press.
Salton, G and MJ McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill: New York, NY.
Simon, R. 2014. “One Week, 3,000 Product Ideas.” *The Wall Street Journal*, 07/03/2014.
Simonton, DK. 2003. “Scientific Creativity as Constrained Stochastic Behavior: The Integration of Product Person, and Process Perspectives.” *Psychological Bulletin*. 129(4): 475-494.
Strzalko, J and Kaszycka, KA. 1991. “Physical Attractiveness: Interpersonal and Intrapersonal Variability of Assessments.” *Social Biology*. 39:170-176.
Surowiecki, J. 2005. *The Wisdom of Crowds*. Random House LLC.Swanson, DR. 1988. "Migraine and Magnesium: Eleven Neglected Connections." Perspectives in Biology and Medicine. 31(4): 526-57.
Swanson, DR. 1988. “Migraine and Magnesium: Eleven Neglected Connections.” *Perspectives in Biology and Medicine*. 31(4): 526-57.
Thornhill, R and Gangestad, SW. 1993. “Human Facial Beauty.” *Human Nature.* 4(3): 237-269.
Tirunillai, S and Tellis, GJ. 2014. Mining Marketing Meaning from Online Chatter: Brand Analysis of Big Data Using Latent Dirichlet Allocation. *Journal of Marketing Research*, 51(4): 463-479.
Toubia, O and Florès, L. 2007. “Adaptive Idea Screening Using Consumers.” *Marketing Science.* 26(3): 342-360.
Uzzi, B, Mukherjee, S, Stringer, M, and Jones, B. 2013. “Atypical Combinations and Scientific Impact.” *Science*. 342(6157): 468-472.
Veryzer, RW and Hutchinson, JW. 1998. “The Influence of Unity and Prototypicality on Aesthetic Responses to New Product Designs.” *Journal of Consumer Research*. 24(4): 374-385.
Ward, TB. 1995. “What’s Old about New Ideas?” in *The Creative Cognition Approach*, Smith SM, Ward TB, and Fiske RA, eds., Cambridge, MA: MIT Press, 157-178.
Winkielman P, Halberstadt J, Fazendeiro T, and Catty S. 2006. “Prototypes are Attractive Because They are Easy on the Mind.” *Psychological Science*. 17(9): 799-806.

Figures and Tables

	Study 1a	Study 1b	Study 1c	Study 2	Study 3	Study 4	Study 5	Study 6
Idea Generation Topic	Insurance plans related to aging and becoming a senior	Insurance plans related to financial security	Insurance plans related to being or staying unemployed	Health-related smartphone apps	Health-related smartphone apps	Oral care solutions for women over 40	Pro Tools	Health-related smartphone apps
Source of Baseline Semantic Network / Prototypical Edge Weight distribution	Pretest with consumers	Pretest with consumers	Pretest with consumers	Google	Google	Google	Google	Google
Participants in Idea Generation	AMT	AMT	AMT	AMT	AMT	Commercial panel	Online ideation community	AMT
Participants in Idea Evaluation	AMT	AMT	AMT	AMT	AMT	Commercial panel + company executives	Online ideation community	AMT
Purpose of the Study	Initial test of hypothesis	Initial test of hypothesis	Initial test of hypothesis	Test the use of Google	Hold number of ideas per participant constant	Include company evaluations + use commercial panel	Test results in real-world ideation community	Test practical tool for improving ideas

Table 1. Overview of Studies.

Note: AMT stands for Amazon Mechanical Turk.

		Study 1a	Study 1b	Study 1c	Study 2	Study 3	Study 4	Study 5	Study 6
	Distance to prototypical edge weight distribution	-3.294**	-2.410**	-3.116**	-0.380**	-0.962**	-0.411*	-0.402**	-0.401**
Edge Weight	Average	-8.046	-12.924*	-7.644	0.633	-0.375	-1.080	1.532**	-0.599
	Coef. of var.	0.696	1.043*	0.560	-0.131	-0.221	0.625	-0.655**	0.341
	Min	-4.983	0.706	6.361	-0.499	-0.021	1.858	1.839**	0.374
	Max	-0.844	-1.313	-0.351	0.267	0.127	-0.364	0.734**	0.132
Node frequency	Average	-10.781**	6.373	11.729**	-0.806	0.717	-0.975	-4.177**	-0.609
	Coef. of var.	1.529**	1.028	1.945**	0.175	0.530	-0.092	-0.967**	-0.171
	Min	14.117	10.754**	12.400	0.420	0.298	0.678	-1.874**	0.255
	Max	-0.397	2.685	-5.132**	0.137	-0.373	0.299	1.215**	0.449
Node clustering coef.	Average	1.530	-1.320	0.539	N/A	N/A	N/A	N/A	N/A
	Coef. of var.	0.291	-4.399**	-3.440	N/A	N/A	N/A	N/A	N/A
	Min	-0.221	-2.461**	-2.716**	N/A	N/A	N/A	N/A	N/A
	Max	-4.271*	5.446	3.964	N/A	N/A	N/A	N/A	N/A
	Size of semantic subnetwork	-0.035**	-0.012	0.014	0.007	-0.002	0.034	-0.002	-0.022**
	Number of characters/1000	1.321**	1.158**	0.914**	-0.032	1.016**	0.940	-0.228**	1.299**
	# observations	276	271	251	555	173	220	1735	648
	# groups (authors)	178	177	167	300	61	163	703	391
	R^2 / χ^2	0.272	0.371	0.246	0.192	0.287	0.072	293.78	0.268

*: $p<0.1$, **: $p<0.05$.

Table 2. Judged creativity vs. prototypicality.

Note: Each column corresponds to one random effects regression with one observation per idea. The dependent variable is the average judged creativity rating of the idea across evaluators (except in Study 5 in which it is the proportion of positive votes on the idea). We are able to control for measures related to the clustering coefficient only in Studies 1a-1c. We capture heterogeneity across participants using random effects.

Condition	Average number of original ideas per participant	Average number of modifications per original idea	Average difference in judged creativity between last modification and original idea
Control	1.590	0.333	0.236
Random Words	1.600	0.489	0.166
Minimum Distance	1.874	0.838	0.032
Maximum Prototypicality	1.643	0.604	0.220

Table 3. Study 6 results.

Note: The Maximum Prototypicality condition is the only one that gives rise to both a significantly greater propensity to modify ideas compared to the Control condition ($p<0.05$), and to modifications that are significant improvements over the original ideas ($p<0.01$).

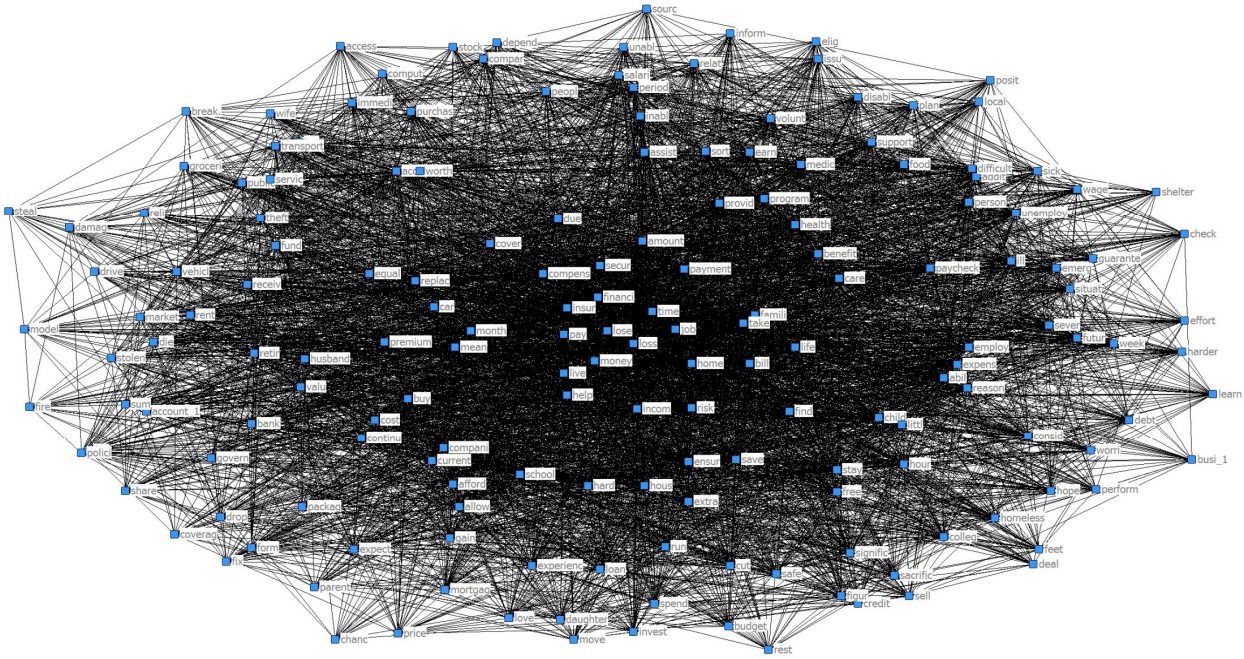

Figure 1. Example of a baseline semantic network.

Note: each node represents a word stem. Each edge captures the scaled co-occurrence between two word stems.

HOW COULD SMARTPHONES HELP THEIR USERS BE HEALTHIER?

We are interested in new ideas for smartphone (e.g., iPhone, Android) apps that will help theirs users keep a healthier lifestyle. In the space provided below, please enter a new idea for a smartphone app priced at $0.99 that would help its users keep a heathier lifestyle. Please be as specific as possible and describe the main features of the app.
You will be asked to enter 3 ideas in total, one after the other.
(Note: Your ideas will be shared with other participants, in an anonymous fashion.)

Please enter exactly one app idea in the box below.

submit app idea - continue

Figure 2. Typical idea generation interface (from Study 1).

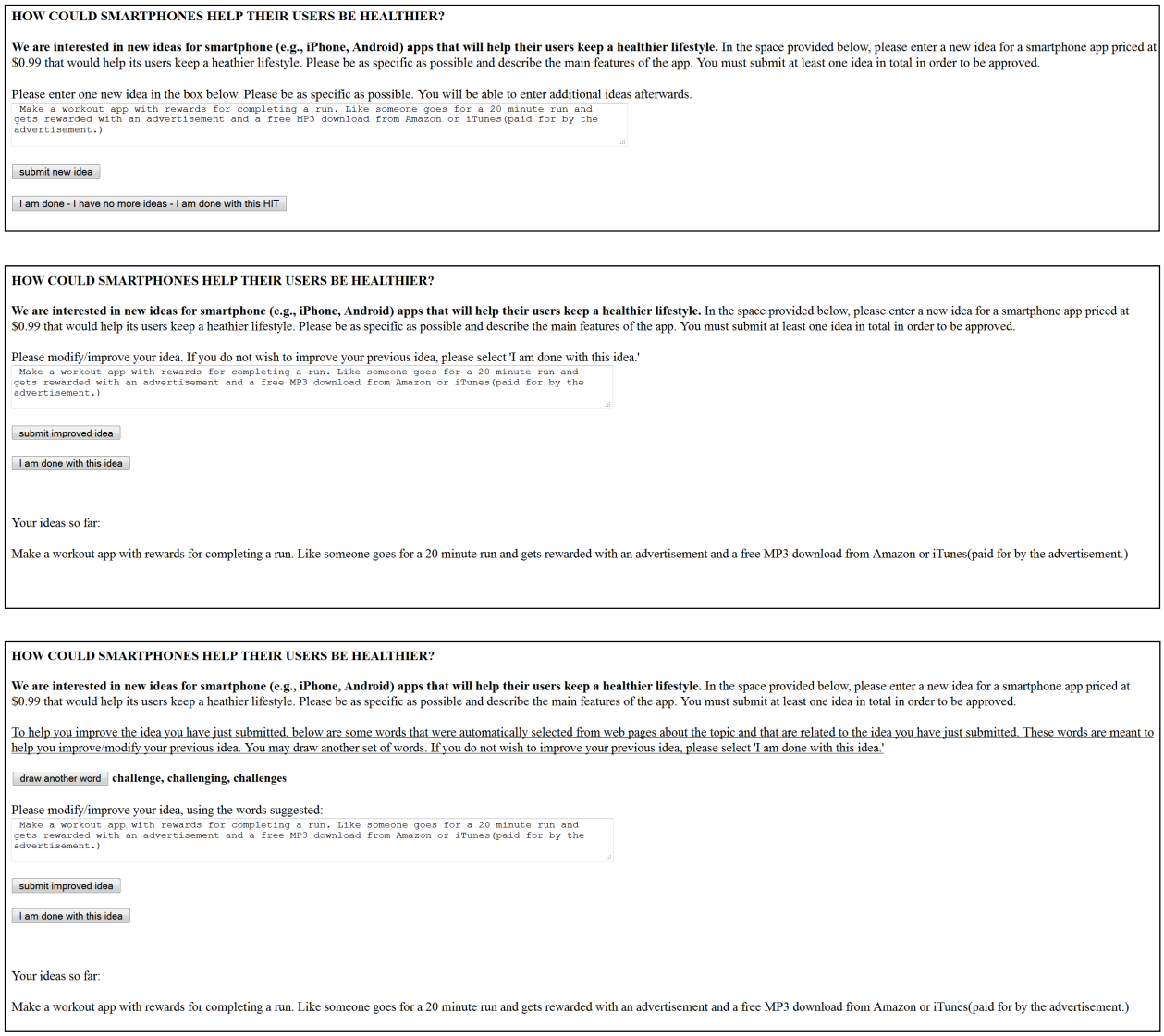

Figure 3. Screenshots from Study 6.

Top: idea collection interface (common across all conditions). Clicking “submit new idea” submits the idea and switches to the idea modification interface. Clicking “I am done” terminates the session. Middle and Bottom: idea modification interface in the control condition (Middle) and the other three conditions (Bottom). The submission box comes pre-loaded with the last idea submitted by the participant. Clicking “submit improved idea” submits the modified idea and re-loads the idea modification interface, allowing the participant to modify their idea further. Clicking “I am done with this idea” switches to the idea collection interface. In the non-Control conditions, words were presented to help participants improve their previous ideas. Clicking “draw another word” cycles through the 10 word stems associated with the idea.