Special Issue: Enhancing Organizations Research Involving Meaning and Culture Using Computational Text and Image Analyses

Organization Studies
1–32
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/01708406241286287
www.egosnet.org/os
Sage

Clearing Up Confusion: The effect of outlier similarity on IPO underpricing

Paul Gouvard
USI Lugano, Switzerland

Rodolphe Durand
HEC Paris, France

Abstract

Prior research demonstrates that audiences tend to converge in their valuations of firms similar to preexisting category prototypes or exemplars. Much less is known of the influence of salient outliers, specific firms that receive market-wide attention due to their extreme, ambiguous performance, on audiences' valuations. We argue that outlier similarity, by contrast with prototype similarity, leads to divergent valuations among individual investors. We explore this insight in the context of initial public offerings (IPOs). In this context, converging valuations among investors lead to limited information asymmetry concerns and hence reduced underpricing on the first day of trading of an issuing firm. Hence, we expect that prototype similarity leads to lower underpricing while outlier similarity leads to higher underpricing. We test our hypotheses using a sample of 2,488 United States IPOs from 1996 to 2015, measuring prototype and outlier similarity through a natural language processing technique applied to nearly 160,000 financial documents. We find that in low-tech industries, where prototypes are informative about category members, prototype similarity reduces underpricing, but not in high-tech industries. Additionally, we find that outlier similarity increases underpricing, especially for more recent outliers. This paper contributes to the literature on market valuation and market categories, and advances research on meaning and culture using new text-based computational methods.

Keywords

archival data, content analysis, economic sociology, IPOs, natural language processing, organizational ecology, population ecology

Introduction

Extant research on audiences' valuation in markets (Durand & Paolella, 2013; Hannan et al., 2019; Lamont, 2012) suggests that 'typical' firms, i.e. firms with high similarity to market categories' prototypes, as well as firms similar to well-known category exemplars (Barlow,

Paul Gouvard is now affiliated to ESSEC Business School, France

Corresponding author:
Paul Gouvard, ESSEC Business School, 3 Av. Bernard Hirsch, Cergy-Pontoise, 95000, France.
Email: paul.gouvard@essec.edu

Verhaal, & Angus, 2019; Pontikes & Barnett, 2017; Soublière & Gehman, 2020; Zhao, Ishihara, Jennings, & Lounsbury, 2018), enjoy a perceptual advantage (Leung & Sharkey, 2014; Negro & Leung, 2013): audience members more easily identify them, infer their unobservable attributes (Hsu & Grodal, 2015) and convergently estimate their value. By comparison, little attention has been dedicated to salient outliers, which stand out across existing and forming categories within a market owing to their recent extreme and ambiguous performance. Indeed, as extreme, the performance of salient outliers does not have clear attributable causes in firms' characteristics and clear implications for firms' underlying value. For instance, dramatic stock price increase, topping corporate social responsibility (CSR) charts, or 100% of IPOs' first-day returns are all extreme performance outcomes with ambiguous interpretations: different audiences may disagree on their causes and their consequences for firm value (e.g. DesJardine, Marti, & Durand, 2021). The inherent difference between salient outliers and category prototypes suggests that the concomitant study of similarity to both category prototypes and salient outliers may shed new light on firm valuation.

Thus, we investigate how prototype similarity and outlier similarity simultaneously affect how investors value an issuing firm in the IPO market. When a firm has high prototype similarity, investors easily identify its observable attributes, correctly infer its unobservable attributes, and use common interpretive schemas to relate this information to expected levels of performance. Hence, given high prototype similarity, investors tend to convergently infer firm value, reducing concerns that there is information asymmetry among investors (Akerlof, 1970). Conversely, when a firm has high outlier similarity, investors tend to dissent about how its observable attributes relate to its underlying value, to discrepantly infer its unobservable attributes and to use different interpretive schemas. Hence, investors tend to diverge in their individual valuations, leading to information asymmetry concerns. Since, in the IPO market, firms' value is discounted when firms are subject to information asymmetry concerns (Biais & Faugeron-Crouzet, 2002; Rock, 1986), the two similarities drive firm value in opposite directions.

We test our hypotheses on a sample of 2,488 United States IPOs from 1996 to 2015. We use a document embedding model (Le & Mikolov, 2014) on 2,488 IPO prospectuses and 159,216 annual reports produced by a benchmark of 33,308 established firms to measure issuing firms' prototype similarity and outlier similarity. Our dependent variable is IPO underpricing, a measure of the discount applied to an issuing firm's value in response to information asymmetry concerns (Pollock & Rindova, 2003; Pollock, Rindova, & Maggitti, 2008). To partially account for regressor imbalance and endogeneity, we use coarsened exact matching (Blackwell, Iacus, King, & Porro, 2009) on top of OLS regressions. Surprisingly, while the relationship between prototype similarity and underpricing does not appear as significant, in additional analyses, we find such a negative and significant relationship in low-tech categories – i.e. when prototypes are informative about category members – but not in high-tech categories. The relationship between outlier similarity and underpricing is positive and significant, and is attenuated when the ambiguity that surrounds a salient outlier is reduced (as time passes). Overall, our findings support our theory and are robust to alternative measures of the two similarities, alternative specifications of the embedding model, and the use of exogeneous instruments.

This paper first contributes to the literature on market valuation (Gouvard & Durand, 2023; Hannan et al., 2019; Lamont, 2012) by contrasting the effect of prototype similarity on investors' valuation with that of outlier similarity. Second, this paper advances research on meaning and culture using computational methods (Aceves & Evans, 2023; Kozlowski, Taddy, & Evans, 2019; Poschmann, Goldenstein, Büchel, & Hahn, 2023) as it introduces new methods to measure both prototype similarity and outlier similarity in a high-dimensional semantic space. It further includes a detailed review of different computational methods to measure semantic similarities between texts in Appendix.

Theory Development

Valuation and the two similarities described

In this section, we focus on how prototype and outlier similarity relate to valuation, especially in the IPO context. We will further explore in additional analyses the mechanisms supporting our main hypotheses and develop theoretical implications in the Discussion section.

Market categories summarize information on ‘the symbolic and material attributes of products, firms, and industries that are both shared among actors and that distinguish these entities from others’ (Durand & Thornton, 2018, p. 632). Market categories are defined by their prototypes, abstract representations of their ‘average’ members (Mervis & Rosch, 1981; Reed, 1972). For example, we all share an abstract representation of the prototypical fast-fashion retailer that summarizes general information about its activities, e.g. fast-fashion products imitate haute couture’s trends, are affordable and are sold in large retail stores. Fast-fashion retailers that fit this representation appear typical (e.g. Zara), while retailers deviating from this prototype appear atypical (e.g. Boohoo.com). It is generally easier for market participants to interpret information and make inferences about firms or products with high prototype similarity, which facilitates their individual valuations and ensure their convergence (e.g. Hsu, 2006; Hsu, Koçak, & Hannan, 2009; Kennedy, Lo, & Lounsbury, 2010).[1]

Recent research suggests that beyond prototype similarity, similarity to salient reference points, i.e. specific firms or products that stand out within a particular market, explain audiences’ valuations. In the early video game market, successful games served as a benchmark to evaluate new games in the absence of pre-established categories (Zhao et al., 2018). In the phone app market, similarity to successful apps is conducive to more downloads (Barlow et al., 2019) while, on crowdfunding platforms, support for new projects is influenced by their similarity to past successes and failures in the same category (Soublière & Gehman, 2020). Finally, imitators crowd in market categories blessed with extreme successes but shun those with extreme failures (Pontikes & Barnett, 2017). A first limitation of this research is that it focuses mostly on exemplars of a specific market category whose influence on audiences’ valuations is limited within the confines of this category (e.g. on Kickstarter, tabletop games or video games). A second limitation is that this research focuses on reference points which are unambiguously related to success (or failure) within their category so that audiences value them convergently. Hence the mechanisms associating similarity to these reference points to audiences’ valuations are not essentially different from those associating prototype similarity to audiences’ valuations. Notably, they both result in convergence among audiences’ valuations.

To expand this research in new directions, we contrast the influence of prototype similarity on audiences’ valuations with that of similarity to reference points which receive market-wide attention due to their extreme, ambiguous performance – which we label as salient outliers. Markets abound with examples of salient outliers: a firm may face widely inconsistent earnings predictions, may achieve market valuation well above expectations based on financial accounting ratios and/or, as in this paper, may experience exceptionally high first-day returns when going public. In all these cases, the firm exhibits an extreme outcome on a particular dimension of performance, but the interpretation of this outcome is unclear such that audiences might interpret it discrepantly. The market valuation of Tesla is a good example of an extreme and ambiguous performance outcome due to the large discrepancy between Tesla’s market valuation and the one that could be expected based on financial accounting ratios (Rothaermel, 2020). Some investors believe this valuation reflects Tesla’s future earnings while others find it absurd based on fundamentals.[2] In this sense, Tesla is a salient outlier in financial markets.[3]

As such, salient outliers are an important counterpoint to category prototypes. Unlike category prototypes, which generally go unquestioned, salient outliers are extensively discussed by market participants, specifically because their extreme performance is difficult to interpret. For instance, Tesla's market valuation is typically the object of regular debate among observers of financial markets. A related consequence of the ambiguity of the relationship between salient outliers' characteristics and their performance is that, while audience members generally share the same representation of category prototypes, they differ in their interpretation of salient outliers. For instance, some investors interpret firms with extremely high CSR performance as investing in long-term sustainability, but others see them as wasting resources (DesJardine et al., 2021). Finally, while category prototypes are relatively stable and enduring, salient outliers are relatively unstable and transient. They either fall from view once audiences' attention is attracted to new salient outliers or stay long enough in the limelight for audiences to resolve the ambiguity surrounding their extreme performance – thus ceasing to be salient outliers.

Overall, contrasting how prototype similarity and outlier similarity relate to valuation would usefully complement existing research. Before detailing our hypotheses, we present our empirical context, the IPO market, in more detail.

Context: IPO and underpricing

During an IPO, an issuing firm becomes a publicly traded company. A set of underwriters (investment banks) and the issuing firm's managers present the offering to investors and write an IPO prospectus, or Form S-1, a document required by the US Securities and Exchange Commission (SEC). The prospectus is the primary source of information about the issuing firm and influences investors' perceptions (Loughran & McDonald, 2013, 2017). Underwriters set the final offer price and allocate shares to investors who bid for them. On the first day of trading, investors who have been allocated shares can sell them to other investors.

In the IPO market, issuing firms are frequently underpriced; i.e. underwriters set the final offer price well below the expected market price (Cohen & Dean, 2005; Park & Patel, 2015). This underpricing is a discount applied to issuing firms when there appears to be significant heterogeneity among investors in the information available to them, so that individual investors tend to reach divergent valuations of the issuing firm, i.e. when information asymmetry concerns are high (Pollock & Rindova, 2003; Pollock et al., 2008). In the absence of underpricing, poorly informed investors would have difficulty taking part in profitable IPOs. Well-informed investors, who value IPOs accurately, would only bet for shares in profitable IPOs. Poorly informed investors would thus be crowded out of these IPOs and be allocated shares mainly in unprofitable IPOs (Biais & Faugeron-Crouzet, 2002; Rock, 1986). Hence, in the absence of underpricing, uninformed investors would have little interest in participating in IPOs. However, underwriters and issuing firms need all kinds of investors to participate in the IPO market to raise sufficient funds and guarantee liquid and efficient market exchanges. Thus, if information asymmetry concerns are high, underwriters underprice IPOs, enabling poorly informed investors to benefit from participating in them.

This feature of the IPO market makes it an ideal setting to explore the effect of prototype and outlier similarity on investors' valuations. Higher (lower) levels of underpricing suggest a higher (lower) discount of a firm's value in response to the two similarities as a function of the effects that they might have on the divergence (convergence) of investors' valuations.

The two similarities and underpricing

When considering in which firms to invest, investors, like other audiences, use market categories to identify potential targets and facilitates information processing about them (Smith, 2011; Wry, Lounsbury, & Jennings, 2014). Investors notably rely on industry categories to help them define potential targets' activities (Zuckerman, 1999, 2017). Hence, we define the relevant prototype for a given issuing firm as the prototypical member of its main industry category.

Prototype similarity covaries with information asymmetry concerns for three reasons. First, when an issuing firm has high prototype similarity, investors recognize many observable attributes of the firm and associate them with expected performance levels – which in turn influences their valuations (Zuckerman, 2017). Second, with higher levels of prototype similarity, investors make better inferences about the firm's unobserved features and their implication for the firm's value (Leung & Sharkey, 2014; Murphy & Ross, 2005; Naumovska & Zajac, 2022; Negro & Leung, 2013). Third, the more prototypical an issuing firm is, the more investors will rely on common interpretive schemas associated with the prototype's category to connect the issuing firm's observable and unobservable attributes to expected performance levels (Hsu, Roberts, & Swaminathan, 2012; Zuckerman, 2004).

Hence, when an issuing firm has high prototype similarity, investors tend to (1) analyse more convergently observable attributes, (2) more convergently infer how the firm's unobservable attributes relate to its value, and (3) apply more common interpretive schemas than when prototype similarity is low. As a result, with higher prototype similarity, investors tend to converge in their individual valuations. Thus, information asymmetry concerns are limited; hence, the level of underpricing of the issuing firm is lower:

Hypothesis 1: *The relationship between issuing firms' prototype similarity and underpricing is negative.*

In the IPO market, we define salient outliers as any issuing firm from any industry category having achieved especially high first-day returns in the recent past. Highly underpriced firms constitute clear instances of extreme performance in this market. As Pollock et al. (2008) mention: 'Some IPOs exhibit dramatic differences between their offering and closing prices (Ritter & Welch, 2002). Such large deviations from the offering price are unusual and surprising, so high levels of underpricing are likely to be noticed and discussed, and likely to become widely available information' (p. 340). High levels of underpricing generate investors' interest and correlate with increased web traffic to the issuing firm's website following its IPO (Demers & Lewellen, 2003). The financial press routinely comments on first-day 'pops', whose causes and consequences generally appear ambiguous.[4] For instance, when the restaurant chain Cava went public in June 2023 it received substantial attention due to its extreme first-day returns of 99%. Commentators speculated about which attributes of Cava might have explained such first-day returns and what they meant for the IPO market.[5] Such debates occur around many salient outliers in the IPO market, such as Beyond Meat (IPO in 2019, 163% first-day returns),[6] Airbnb (IPO in 2020, 113% first-day returns),[7] or Poshmark (IPO in 2021, 142% first-day returns).[8]

Salient outliers influence IPO investors' valuations of similar firms. First, since salient outliers tend to receive media coverage, which weights into the decision of IPO investors (Pollock & Rindova, 2003), we would expect IPO investors to be well aware of recent salient outliers and sensible to an issuing firm's similarity to them. Second, valuing issuing firms is difficult due to limited access to public information on those firms (Pollock et al., 2008). We would thus expect IPO investors to support their valuations of issuing firms on multiple reference points beyond category prototypes. Due to the attention received by salient outliers, the similarity between a focal issuing firm and a salient outlier would likely be perceived as worthy of consideration when valuing the issuing firm. For instance, before Cava went public in 2023, it was compared to Sweetgreen, a similar company, which also experienced a significant first-day 'pop' in 2021 (76%).[9] Third, due to the lack of information on issuing firms, IPO investors pay particular attention to IPO prospectuses (Loughran & McDonald, 2013; Martens, Jennings, & Jennings, 2007). IPO investors are thus likely sensitive to similarities between issuing firms' prospectuses and especially between those of salient outliers having captured market attention and those of issuing firms in which they consider investing – these similarities may in turn influence their valuations.[10]

An important specificity of salient outliers is that their extreme performance is ambiguous – some audience members consider it as an aberration, while others strive to make causal associations between the outlier's characteristics and its extreme performance. Due to this ambiguity, some investors are likely misguided in their interpretation of how the salient outlier's attributes relate to its performance and in their interpretation of this performance for the salient outlier's underlying value. Hence, there likely are divergences among investors' individual valuations of salient outliers, and in turn outlier similarity may thus foster divergence among investors' valuations of issuing firms.

First, when a focal issuing firm is similar to a salient outlier, this indicates that they share some observable attributes across their respective industry categories. Through analogical reasoning (Durand & Thornton, 2018; Etzion & Ferraro, 2010; Ketokivi, Mantere, & Cornelissen, 2017), some investors may thus base their interpretation of how those observable attributes relate to the issuing firm's performance and underlying value on their interpretation of the salient outlier's own observable attributes. Second, if outlier similarity is high, some investors may further infer that the issuing firm possesses unobserved attributes similar to those of the salient outlier. They might again base their interpretation of how those unobservable attributes relate to the issuing firm's performance and underlying value on their interpretation of the salient outlier's own unobservable attributes. However, in both cases, since some investors' individual interpretation of the salient outlier's attributes and their relationship to performance and valuation are likely misguided and divergent from that of better-informed investors, their individual valuations of the focal issuing firm may be misguided and divergent as well. Finally, due to the ambiguity surrounding salient outliers' extreme performance, investors cannot rely on shared interpretive schemas to make sense of available information both about the outlier's performance and the performance of similar others. As a result, outlier similarity leaves wide open the choice of interpretive schemas mobilized by various investors. Therefore, outlier similarity may increase the likelihood of diverging value estimates.[11]

Hence, when an issuing firm has high outlier similarity, investors tend to diverge in their individual valuations because they (1) do not convergently analyse how observable attributes relate to an issuing firm's value, (2) infer in discrepant ways the firm's unobservable attributes and how they relate to its value, and (3) apply multiple and diverse interpretive schemas. As a result, with higher outlier similarity, information asymmetry concerns among investors are likely high; thus, the level of underpricing of the issuing firm is higher:[12,13]

> **Hypothesis 2:** *The relationship between issuing firms' outlier similarity and underpricing is positive.*

Data, Methods and Results

Data

We collected data on US IPOs from 1996 to 2015 from the SDC Platinum new issues database and Professor Jay Ritter's database (see Loughran & Ritter, 2004: Appendix B). We collected stock-level data for firms in our sample from CRSP and fundamentals data from Compustat. We excluded

IPOs initially priced below $5 (i.e. penny stocks), and IPOs for financial institutions, closed-end funds, American depository receipts and real estate investment trusts. We collected S-1 forms submitted by the 2,488 IPOs in our sample for which we had a valid Central Index Key (CIK), which was used to identify financial documents in the SEC database. As we measured prototype similarity using a benchmark of publicly traded established firms, we downloaded 159,216 annual reports (Form 10-K) for all 33,308 US publicly listed firms included in Compustat for which we had a CIK between 1995 and 2015 from the SEC website.

Natural language processing method to measure prototype and outlier similarity

To measure firms' prototype and outlier similarities, we trained a document embedding model on our corpus of 159,216 annual reports and 2,488 IPO prospectuses. A document embedding model learns *document embeddings*, i.e. multidimensional vectors representing documents in a semantic space by sliding a window over documents and trying to predict a target word within this window based on neighbouring words and document identity (Le & Mikolov, 2014). Through this process, documents containing semantically similar words are mapped to the same region of a high-dimensional semantic space. Similarities between firms can then be evaluated based on similarities between their document embeddings. This reveals similarities that a diligent examination of firms' attributes – generally not possible with large samples – cannot precisely uncover. For an in-depth discussion of our measurement strategy, see Appendix A.

Preprocessing of documents. We preprocessed documents in several steps. First, we extracted the main texts of the annual reports or the IPO prospectuses. We reduced these texts to lists of tokens (i.e. words), removed punctuation and digits, and lowered words. We removed stop words and words characteristic of SEC filings rather than of firms themselves by removing all words that appeared in more than 50% of documents of each type. We also removed very infrequent words and focused on the 10,000 most frequent words.

Document embedding model specifications. Building on standard specifications (Dai, Olah, & Le, 2015; Lau & Baldwin, 2016; Le & Mikolov, 2014), we fixed the number of dimensions of the document embeddings learned by the model at 300 and adopted a window size of 5. The model was trained by passing 5 times over the entire corpus using an initial learning rate of 0.025, which decreased linearly at each pass to a minimum of 0.005. To improve model quality, we downsampled words that appeared more than 10^{-5} times (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Our results are robust to the use of alternative specifications of the model (see Appendix B).

Variables

Dependent variable. Our dependent variable is the *underpricing* for a focal issuing firm, measured as the difference between the market price at the end of the first day of trading and the offer price divided by the offer price. The average *underpricing* in our sample is 29%, with a standard deviation of 55%. The variable skewed to the right due to some firms achieving very high underpricing (up to 606% for the IPO of theGlobe.com in 1998). These descriptives were comparable to those of other studies using underpricing as their main dependent variable (e.g. Loughran & McDonald, 2013; Park, Borah, & Kotha, 2016).

Independent variables. To measure *prototype similarity*, we first created prototypes for each 3-digit SIC code industry category.[14] For each year and industry, we considered the set of all established firms in the industry and took the centroid of the document embeddings associated with their annual reports:

$$Prototype_{I^y} = \frac{1}{\left|I^y\right|}\sum_{f \in I^y} Embedding_{f,y}$$

where f indexes established firms, I^y indexes the set of firms in industry I in year y based on Compustat data, $\left|I^y\right|$ is the cardinal of this set and $Embedding_{f,y}$ is the document embedding associated with the form 10-K of established firm f in year y.

We operationalized the *prototype similarity* of an IPO as the similarity of the document embedding of its IPO prospectus (Form S-1) to its industry centroid and mean-centred it[15]:

$$Prototype\,similarity_g = Cosine\,similarity\,(Embedding_g, Prototype_{G^{y--1}})$$

where g indexes the issuing firm, y indexes the year of issuing firm g's IPO, G^{y--1} indexes the industry of issuing firm g in year $y-1$, $Embedding_g$ refers to the document embedding associated with the form S-1 of issuing firm g and $Prototype_{G^{y-1}}$ is the prototype of industry G in year $y-1$.

We measured an issuing firm's *outlier similarity* as its similarity to the IPO with the highest level of underpricing in the preceding year. Formally:

$$Outlier\,similarity_g = Cosine\,similarity(Embedding_g, Embedding_{Outlier_g})$$

where g indexes the issuing firm, $Outlier_g$ indexes the outlier used for firm g, i.e. the most highly underpriced IPO in the year preceding g's IPO, $Embedding_g$ is the document embedding associated with Form S-1 of issuing firm g, and $Embedding_{Outlier_g}$ is the embedding associated with Form S-1 of $Outlier_g$.

Note that unlike prototype similarity, outlier similarity is measured as a function of one's similarity to the most salient outlier irrespective of industry boundaries. This is in line with our theorization of outliers as affecting the valuations of investors throughout the entire IPO market due to their sudden and extreme underpricing irrespective of industry boundaries. Finally, we mean-centred outlier similarity. Figure 1 plots the distribution of both similarities.

Controls. Since the level of 'hype' for a particular industry may be correlated with both the outlier similarity of issuing firms within it and the level of underpricing, we control for it using two proxies: the *average stock return* and the *average stock price volatility* over the past quarter within a focal issuing firm's industry.[16]

As the level of investor enthusiasm in the financial market, which has known important variations during our period, impacts underpricing, we control for it using a rolling average of the VIX, the so-called 'fear index', over the three months preceding a focal issuing firm's IPO.[17]

Since high-tech IPOs tend to be subject to more information asymmetry (Carpenter, Pollock, & Leary, 2003; Ozmel, Reuer, & Wu, 2017), we created a tech dummy variable that took the value of 1 for issuing firms with three-digit SIC codes associated with *high-tech* industries (Kile & Phillips, 2009)[18] and 0 otherwise.

As the time elapsed since the rise of the outlier and the hype surrounding it may be correlated with underpricing, we measure the *temporal distance from the outlier* using the log of the number of days since the most highly underpriced IPO in the preceding year.[19] We further control for the *log of the number of days since the beginning of the year*.

Prior to an IPO, underwriters set a price range for the offering. *Offer price revision* is the percentage change between the final offer price set by underwriters and the middle of this initial price

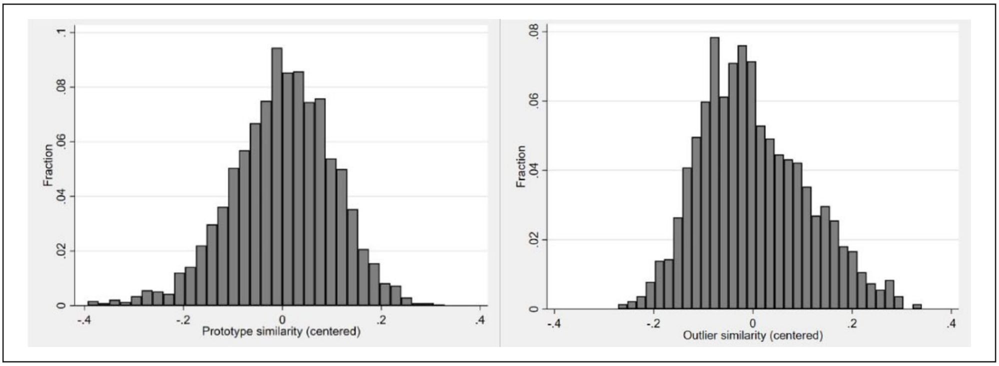


Figure 1. Distribution of prototype similarity (centred) and outlier similarity (centred).

range (Hanley, 1993; Loughran & Ritter, 2002). To control for *IPO market hotness*, we used the percentage of IPOs with offer prices above the midpoint of the initial price range in the preceding month (Ibbotson, Sindelar, & Ritter, 1994). We also controlled for whether the IPO received venture capital (*VC*) support prior to the IPO, as its presence (value = 1) or absence (value = 0) influences underpricing (Arthurs, Hoskisson, Busenitz, & Johnson, 2008; Lee & Wahal, 2004). We further controlled for firms' *size* (log of assets) and *age* (log of the number of years since founding plus 1).

Main results from OLS regression with multiple fixed effects

We estimated ordinary least squares (OLS) regressions of underpricing on prototype similarity and outlier similarity. In all models, we included *industry fixed effects* using 2-digit SIC codes to control for time-invariant unobserved heterogeneity between industries. We included lead *underwriter fixed effects* to control for stable unobserved factors, such as a lead underwriter's prestige (Carter & Manaster, 1990), connections with institutional investors (Goldstein, Irvine, & Puckett, 2011; Gondat-Larralde & James, 2008), or propensity to underprice shares. We included *stock exchange fixed effects* to ensure that our results are not driven by systematic differences in IPO returns as a function of the market in which it occurs. We also included *year effects* to control for year-specific trends. Although including these fixed effects helps account for omitted variable bias, we acknowledge that our modeling strategy does not yield causal estimates, a limitation to which we return in the Discussion section. We clustered errors by industry. Table 1 presents descriptive statistics and the correlation matrix for our variables.

In Table 2, we present models 1–7, which correspond to our analysis using multiple fixed effects. Model 1 includes only our control variables. As expected, industry volatility, industry returns, VC backing, offer price revision, market hotness and stock market volatility (VIX) are all positively and significantly associated with *underpricing* (as well as the number of days elapsed since the outlier's emergence) while age is negatively and significantly associated with *underpricing*. Model 2 introduces prototype similarity as an independent variable and does not show a significant association between prototype similarity and underpricing. Hence, H1 does not receive direct support. Model 3 introduces outlier similarity as an independent variable and reveals a positive and significant association between outlier similarity and underpricing ($p = 0.001$), which supports H2. In model 3, an issuing firm with an outlier similarity one standard deviation above the sample mean (i.e. + 0.11, Table 1) experiences a level of underpricing 5.6% higher than that of an

Table 1. Correlations and descriptive statistics.

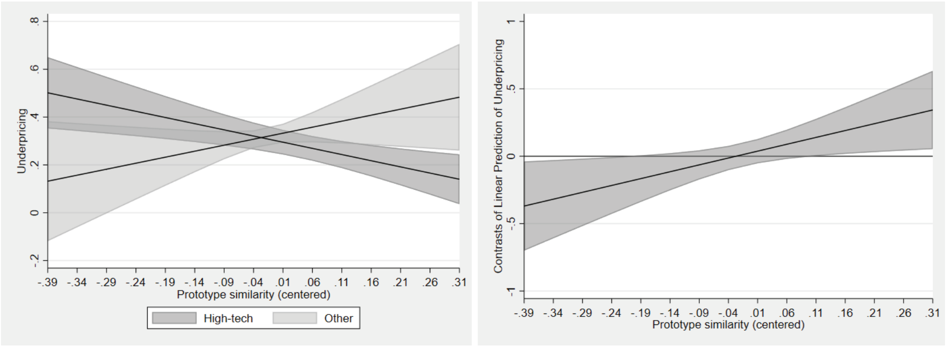
Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	Min	Max	Mean	p50	Std. Dev.
(1) Underpricing	1.000													−0.636	6.056	0.288	0.125	0.549
(2) Prototype similarity (centred)	−0.057	1.000												−0.394	0.328	0.000	0.004	0.104
(3) Outlier similarity (centred)	0.260	−0.100	1.000											−0.272	0.340	0.000	−0.015	0.108
(4) Tech firm	0.165	−0.092	0.312	1.000										0.000	1.000	0.565	1.000	0.496
(5) Temporal dist. from outlier (centred)	−0.108	−0.014	0.027	−0.027	1.000									−2.471	0.851	0.000	0.085	0.498
(6) Offer price revision	0.125	−0.040	0.078	0.075	−0.014	1.000								−0.861	1.800	−0.006	0.000	0.116
(7) Market hotness	0.362	−0.038	0.183	0.088	−0.138	0.151	1.000							−44.520	55.480	0.000	0.484	22.030
(8) VC backing	0.219	−0.032	0.288	0.497	−0.009	0.032	0.115	1.000						0.000	1.000	0.525	1.000	0.499
(9) Average VIX over past quarter	0.199	−0.003	0.073	0.036	0.026	0.022	0.197	0.070	1.000					10.980	52.660	19.790	19.990	5.120
(10) Av. volatility in industry past quarter	0.341	−0.123	0.170	0.225	−0.215	0.052	0.312	0.192	0.572	1.000				0.005	0.373	0.040	0.038	0.017
(11) Average return in industry past quarter	0.301	0.015	0.066	0.088	−0.228	0.108	0.543	0.078	0.052	0.189	1.000			−0.927	1.304	0.070	0.050	0.167
(12) Log of N of days since start of year	−0.030	−0.027	0.030	−0.024	0.702	0.012	−0.027	−0.005	0.045	−0.093	−0.237	1.000		2.079	5.869	5.068	5.298	0.724
(13) Log of assets	−0.158	0.068	−0.176	−0.414	0.126	0.037	−0.134	−0.413	−0.167	−0.342	−0.078	0.054	1.000	0.384	11.830	4.315	4.017	1.675
(14) Log of age	−0.183	0.028	−0.215	−0.248	−0.003	−0.033	−0.140	−0.321	−0.147	−0.212	−0.031	−0.031	0.350	0.000	5.075	2.325	2.197	0.954

Table 2. Underpricing: effects of prototype similarity (centred) and outlier similarity (centred).

Variables	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Prototype similarity (centred)		−0.006		−0.010	−0.516**		−0.457**
		(0.203)		(0.192)	(0.165)		(0.164)
Outlier similarity (centred)			0.510***	0.509***		0.526***	0.477**
			(0.142)	(0.143)		(0.141)	(0.156)
Prototype similarity#Tech firm					1.018*		0.912*
					(0.425)		(0.417)
Outlier similarity#Temp. dist. from outlier						−0.629*	−0.587*
						(0.274)	(0.274)
Tech firm	−0.012	−0.011	−0.017	−0.015	0.027	−0.016	0.020
	(0.049)	(0.052)	(0.045)	(0.048)	(0.044)	(0.044)	(0.041)
Temporal distance from outlier (centred)	0.111*	0.112*	0.105*	0.105*	0.110*	0.073	0.075
	(0.050)	(0.051)	(0.051)	(0.052)	(0.050)	(0.060)	(0.061)
Offer price revision	0.189**	0.189**	0.171**	0.171**	0.181**	0.161*	0.156*
	(0.055)	(0.055)	(0.057)	(0.057)	(0.056)	(0.061)	(0.060)
Market hotness	0.003**	0.002**	0.002**	0.002**	0.003***	0.002**	0.002***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
VC backing	0.094**	0.093**	0.087**	0.087**	0.084**	0.088**	0.080**
	(0.031)	(0.031)	(0.031)	(0.030)	(0.029)	(0.031)	(0.029)
Average VIX over past quarter	0.009*	0.009*	0.010*	0.010*	0.009*	0.009*	0.009*
	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)
Average volatility in industry over past quarter	5.203***	5.220***	5.020***	5.037***	5.236***	4.897***	4.953***
	(0.956)	(0.938)	(1.027)	(1.011)	(0.927)	(1.133)	(1.116)
Average return in industry over past quarter	0.627**	0.629**	0.649**	0.651**	0.626**	0.634**	0.632**
	(0.205)	(0.201)	(0.201)	(0.197)	(0.200)	(0.197)	(0.192)
Log of number of days since beginning of year	−0.040	−0.040	−0.036	−0.036	−0.039	−0.025	−0.026
	(0.026)	(0.026)	(0.028)	(0.028)	(0.025)	(0.028)	(0.028)
Log of age	−0.034*	−0.034*	−0.031*	−0.031*	−0.035*	−0.030*	−0.031*
	(0.013)	(0.014)	(0.013)	(0.014)	(0.015)	(0.013)	(0.015)
Log of assets	−0.018+	−0.018+	−0.014	−0.014	−0.012	−0.014	−0.009
	(0.010)	(0.010)	(0.009)	(0.009)	(0.010)	(0.009)	(0.009)
Constant	0.185	0.183	0.143	0.141	0.147	0.113	0.085
	(0.186)	(0.183)	(0.207)	(0.204)	(0.178)	(0.200)	(0.190)
Industry, Year, Lead underwriter, and Stock exchange FE	*YES*	*YES*	*YES*	*YES*	*YES*	*YES*	*YES*
Observations	2,050	2,046	2,050	2,046	2,046	2,050	2,046
Adjusted R-squared	0.286	0.285	0.290	0.289	0.290	0.292	0.295

Robust standard errors in parentheses, ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, $^{+}p < 0.1$.

issuing firm with the average level of outlier similarity, all else equal. In model 4, which includes both independent variables, the association between outlier similarity on underpricing remains positive and significant ($p = 0.001$), and its magnitude remains the same.



PANEL A

PANEL B

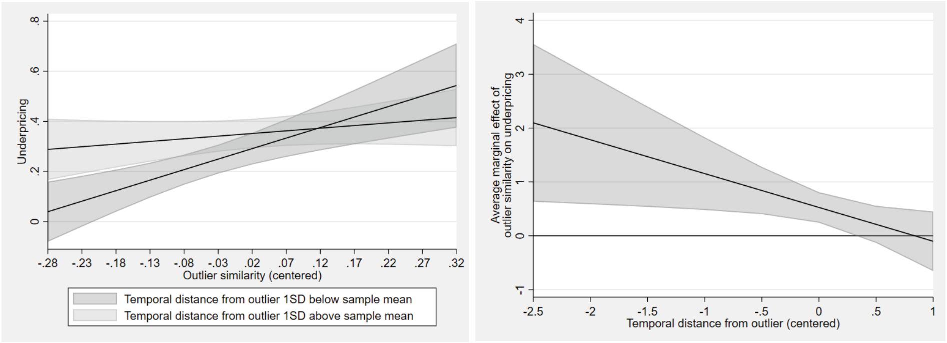
Panel A of Figure 2 represents the predicted level of underpricing as a function of prototype similarity (centered) using Model 5 from Table 2 by industry category. The negative relationship between prototype similarity and underpricing in non-high-tech industries is almost reversed in high-tech industries. This is consistent with the findings presented in Panel B where we contrast the average marginal effects of prototype similarity on underpricing in other vs high-tech industries.

Figure 2. Marginal effect of prototype similarity (centred) on underpricing by industry.
Panel A of Figure 2 represents the predicted level of underpricing as a function of prototype similarity (centred) using model 5 from Table 2 by industry category. The negative relationship between prototype similarity and underpricing in non-high-tech industries is almost reversed in high-tech industries. This is consistent with the findings presented in Panel B where we contrast the average marginal effects of prototype similarity on underpricing in other vs high-tech industries.

Additional analysis probing the relationship between the two similarities and underpricing

Probing the mechanism supporting the effect of prototype similarity. Per our preceding argument, the negative relationship between prototype similarity and underpricing rests on investors' ability to both make converging inferences about firms' unobserved attributes and use shared interpretive schemas associated with categories when firms have high prototype similarity. However, categorical knowledge and interpretive schemas may not be equally distributed across industries. For instance, issuing firms in high-tech industry categories tend to have idiosyncratic features (Ozmel et al., 2017; Wu & Reuer, 2021), be young, develop new products and face considerable risk (Carpenter et al., 2003). All these factors complicate the abstraction of their features into category prototypes, limiting the quality of categorical knowledge and the kind of inference it permits. Hence, if our theory is correct, the negative association between prototype similarity and underpricing could be weakened in these industries.

In model 5, we thus include an interaction between prototype similarity and the high-tech industry dummy. We find that prototype similarity is negatively associated with underpricing as per H1's expectations ($p=0.003$), and as expected, the coefficient of the interaction between prototype similarity and the tech dummy is positive and significant ($p=0.02$). More precisely, when prototype similarity is low, the difference between the average predicted levels of underpricing in high-tech and low-tech industries is negative and significant, while it is positive and significant when prototype similarity is high (see Figure 2). Overall, these results show that prototype similarity has a



PANEL A

PANEL B

Panel A of Figure 3 represents the predicted level of underpricing as a function of outlier similarity (centered) using Model 6 from Table 2 by temporal distance from the outlier emergence. Panel B represents the average marginal effects of outlier similarity on underpricing as a function of temporal distance from the outlier.

Figure 3. Marginal effects of outlier similarity (centred) on underpricing by temporal distance from outlier emergence.
Panel A of Figure 3 represents the predicted level of underpricing as a function of outlier similarity (centred) using model 6 from Table 2 by temporal distance from the outlier emergence. Panel B represents the average marginal effects of outlier similarity on underpricing as a function of temporal distance from the outlier.

negative and significant association with underpricing in categories where prototype similarity provides better information (i.e. in low-tech industries) and a positive but not significant association with underpricing in less mature categories where prototypes are less established (i.e. in high-tech industries), which is in line with our theory. This explains why we do not find support for H1 in model 2 since the two effects (in low- vs. high-tech industries) cancel each other out (as in Figure 2A).

Probing the mechanism supporting the effect of outlier similarity. The mechanism driving the positive association between outlier similarity and underpricing rests on the ambiguity of salient outliers' extreme performance, which leads some investors toward divergent and possibly misguided valuations. If this mechanism is at play, we expect that as time passes and more information is revealed about outliers, investors converge on a common interpretation of the causes and implications of outliers' past extreme performance. Thus, as time passes, investors' valuation should diverge less and the positive association between outlier similarity and underpricing should be weaker.

In model 6, we thus include an interaction between outlier similarity and temporal distance from the outlier. The interaction between outlier similarity and temporal distance is negative and significant ($p=0.026$). As illustrated in Panel A of Figure 3, the slope of the relationship between outlier similarity and underpricing is thus contingent on the temporal distance from the outlier. Panel B of Figure 3 further shows the average marginal effects of outlier similarity on underpricing as a function of the temporal distance from the outlier. As expected, for low values of temporal distance, the average marginal effect of outlier similarity on underpricing is positive and significant but not significant for higher values of temporal distance.

Full model. In model 7 (full model), the main and interaction terms are in the expected direction and significant ($p < .01$ and $p < .05$, respectively), further supporting our theorizing.

Robustness checks

Results from CEM. One limitation of our approach is the potential covariate imbalance between firms with high vs. low prototype or outlier similarity. To partially address this limitation, we applied coarsened exact matching (CEM) (Blackwell et al., 2009) to assess whether being more vs. less similar to prototypes or outliers while being identical in as many other dimensions as possible is associated with more or less information asymmetry (see Corritore, Goldberg, & Srivastava, 2020 for a comparable approach). We created one discrete measure for both 'treatments', which takes the value of 1 if prototype (outlier) similarity is greater than the issuing firm's industry median in a given year and 0 otherwise. We then performed two separate CEMs using the command cem in Stata (Blackwell et al., 2009) – one for each of our treatment variables – using *VC backing*, *log of assets* and *log of age* to create strata of comparable observations. To determine the variables to use to create strata, we ran a logistic regression of our treatment variables on all the controls used in our main OLS regression. We found that *tech firm*, *VC backing* and *log of assets* are significant predictors of having prototype similarity above the industry median in a given year, while *tech firm*, *VC backing* and *log of age* are significant predictors of having outlier similarity above the industry median in a given year. As we interact *tech firm* with *prototype similarity* in our analysis, we do not use it for matching to ensure that it varies within strata.

For both treatments, we used exact matching on *VC backing* and used the default binning strategy implemented in *CEM* in Stata for *log of age* and *log of assets*. The upper part of Table 3 shows that we successfully reduced imbalances in the data along the selected covariates based on the reduction in the multivariate L_1 statistic, which measures the extent to which the distribution of covariates of interest in the treated group mirrors that of the control group (Blackwell et al., 2009). Additionally, differences in the mean values of the covariates used for matching purposes between the treated and control groups are not significant after matching.

After matching, we ran the OLS underpricing regression on our continuous measure of the treatment used for matching. The lower part of Table 3 summarizes our results, which provide strong support for our hypotheses.

Other robustness checks. We measured prototype similarity using 4- and 2-digit SIC codes to determine whether the coarseness of the industry classification used to identify peers influenced our results. As investors may consider several outliers when assessing new offerings, we tested models where outlier similarity refers to an issuing firm's average similarity to the two, three, four and five firms with the highest underpricing in the preceding year. Our results are robust to these alternative measures of the two similarities.[20]

We tried different specifications of Doc2Vec with a window size of 5 (emphasis placed on similarities between neighbouring words) or 10 (emphasis placed on broader topical similarities) and embeddings dimensions of 100, 200 or 300. Results are robust to these alternative specifications (see Appendix B).

Another limitation of our main results is that prototype and outlier similarity are not randomly assigned across firms and may be endogenous. To address this endogeneity concern, we identified two exogeneous instruments that should be correlated with prototype and outlier similarity but not with the error term. Results are robust to this approach (see Appendix B).

Table 3. OLS regression of underpricing on prototype similarity (centred) and outlier similarity (centred) after CEM.

	Treatment: Prototype similarity				Treatment: Outlier similarity			
Number of strata/matched strata	170/111				170/119			
Number of treated units/controls	768/1199				1030/973			
L_1 statistic before matching/after matching	.39/.29				.39/.27			
VC backing (diff. in mean)	.223 ($p=.000$)		.000 (exact matching)		.207 ($p=.000$)		.000 (exact matching)	
Log of age (diff. in mean)	−.189 ($p=.000$)		.000 ($p=.990$)		−.263 ($p=.000$)		.008 ($p=.823$)	
Log of assets (diff. in mean)	−.628 ($p=.000$)		−.010 ($p=.885$)		−.518 ($p=.000$)		.019 ($p=.765$)	
Variables	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14	Model 15
Prototype similarity	−0.291	−0.128	0.215	0.227				
	(0.178)	(0.174)	(0.175)	(0.165)				
Outlier similarity					1.443***	1.106***	0.394*	0.475**
					(0.254)	(0.211)	(0.192)	(0.159)
Controls	*NO*	*NO*	*NO*	*YES*	*NO*	*NO*	*NO*	*YES*
CEM Strata FE	*NO*	*YES*	*YES*	*YES*	*NO*	*YES*	*YES*	*YES*
Industry, Year, Lead underwriter, and Stock exchange FE	*NO*	*NO*	*YES*	*YES*	*NO*	*NO*	*YES*	*YES*
Observations	1,967	1,967	1,967	1,967	2,003	2,003	2,003	2,003
Adjusted R-squared	0.015	0.083	0.204	0.278	0.071	0.128	0.235	0.299
Variables	Model 16	Model 17	Model 18	Model 19	Model 20	Model 21	Model 22	Model 23
Prototype similarity	−0.933***	−0.759***	−0.813***	−0.741***				
	(0.234)	(0.221)	(0.232)	(0.203)				
Prototype sim.#Tech firm	1.364***	1.239***	1.649***	1.557***				
	(0.234)	(0.255)	(0.291)	(0.304)				
Outlier similarity					1.511***	1.183***	0.401*	0.478**
					(0.235)	(0.205)	(0.201)	(0.167)
Outlier sim.#Temp. dist.					−2.107***	−1.786***	−1.001***	−0.668*
					(0.331)	(0.329)	(0.251)	(0.256)
Controls	*NO*	*NO*	*NO*	*YES*	*NO*	*NO*	*NO*	*YES*
CEM Strata FE	*NO*	*YES*	*YES*	*YES*	*NO*	*YES*	*YES*	*YES*
Industry, Year, Lead underwriter, and Stock exchange FE	*NO*	*NO*	*YES*	*YES*	*NO*	*NO*	*YES*	*YES*
Observations	1,967	1,967	1,967	1,967	2,003	2,003	2,003	2,003
Adjusted R-squared	0.027	0.092	0.215	0.288	0.100	0.149	0.240	0.301

The upper part of Table 3 presents a summary of our coarsened exact matching strategy. The intermediate part presents OLS regressions of underpricing on the main independent variables after matching observations using CEM. Models 8 to 11 present results after matching on prototype similarity. Models 12 to 15 present results after matching on outlier similarity. The lower part presents the same models but including the interactions between prototype similarity and high-tech industry category and between outlier similarity and temporal distance. Robust standard errors in parentheses.

***$p<0.001$, **$p<0.01$, *$p<0.05$.

We find in models including an interaction term between prototype similarity and the high-tech dummy that the association of prototype similarity with information asymmetry is negative and significant, and the interaction coefficient between prototype similarity and the high-tech dummy is positive and significant ($p<0.000$). In models testing for the effect of outlier similarity on underpricing, we further find an increase of information asymmetry for higher levels of outlier similarity, attenuated as time passed.

Discussion

This paper first contributes to the literature on market valuation (Lamont, 2012), and specifically to research on categories and audience evaluation (Delmestri, Wezel, Goodrick, & Washington, 2020; Hannan et al., 2019). Recent research in this vein focused on between-audience member variations in valuation of the same firm or product, tying it to audience members' reliance on different evaluation modes or theory of value (Gouvard & Durand, 2023; Paolella & Durand, 2016). This paper continues this vein of research, introducing salient outliers as a source of between-audience member variations in valuation. Due to the ambiguity that surrounds salient outliers, they do not help audiences converge in their valuations but rather beget confusion. In the IPO market, this fuels information asymmetry concerns and results in a value discount. Importantly, rather than presenting audiences' reliance on outliers as an alternative mode of evaluation relative to prototype-based evaluation, we argue that the two similarities *concomitantly* affect valuation.

Similarity to others is often presented as a source of legitimacy, facilitating valuation and helping audiences reach an agreement on the value of a particular object – an advantage which comes at the expense of reduced distinctiveness and attention (Askin & Mauskapf, 2017; Slavich & Castellucci, 2016; Zuckerman, 2016). Outlier similarity seems to work relatively differently from this 'baseline' picture: it may be conducive to both greater attention being received and greater divergence among audiences regarding the value of a particular object. One potential implication is that outlier similarity may be used to counteract the influence of one's similarity to other reference points as one strives to achieve optimal distinctiveness (Zhao & Glynn, 2022).

We further find evidence that the influence of salient outliers on audiences' valuations co-exist with that of pre-existing category prototypes. This is particularly interesting given the different nature of category prototypes and salient outliers. Models of audience members' valuations in markets often take preexisting categories as given (Hannan et al., 2019; Zuckerman, 1999). Although categories may change over time (Gollnhofer & Bhatnagar, 2021; Innis, 2022; Pedeliento, Andreini, & Dalli, 2020), they are generally presented as relating to the enduring features of organizations (Hannan et al., 2019; Kim & Jensen, 2011). Our findings suggest that, in addition to this relatively stable, slow-changing component of audience members' valuations, we should consider a more transient, temporary component that may correspond to sudden hype, often epitomized by salient outliers, that may or may not subsequently move on to provide the foundation for new categories (Durand & Khaire, 2017).

The respective strength of these two influences on audience members' valuations remains a relatively open question. In our setting, it seems that outlier similarity has a stronger and steadier association with prototype similarity, whose influence is washed out in high-tech categories.[21] However, this might be due to boundary conditions such as IPO investors' generally lower interest in firms with high prototype similarity and greater openness to firms engaging in atypical activities (Pontikes, 2012). In any case, the influence of salient outliers on audiences' valuations may suggest that audiences' valuations regularly fluctuate in response to the rise and fall of salient outliers, which periodically attract their attention. This may contribute to explaining why, in prior studies that did not account for the influence of salient outliers, the premium that theoretically accrues to firms with high prototype similarity is limited or not observed in markets where outliers lie far away from prototypes. The meddling influence of salient outliers in audience members' valuations would thus join the rank of possible explanations for the limited benefit of prototype similarity in some markets, alongside the use of different evaluation modes (Glaser, Atkinson, & Fiss, 2020; Gouvard & Durand, 2023), market growth (Pozner, DeSoucey, Verhaal, & Sikavica, 2022), categorical contrast (Kovács & Hannan, 2010) or status (Sharkey, 2014).

One aspect that we could not explore in depth in this paper is salient outliers' own relationship to prototypes. Theoretically, salient outliers could either have high or low prototype similarity.[22] However, we could expect salient outliers which are also categorical anomalies to appear particularly ambiguous. While in our setting we did not find evidence of this, we believe that future research could benefit from further exploring how category prototypes relate to salient outliers. For instance, it could be that salient outliers that do not belong to pre-existing categories and eventually impose a positive interpretation of their performance succeed in becoming atypical exemplars that audience members may use as a benchmark when confronted with unexpected offerings (Gouvard & Durand, 2023).

We further contribute to organizational research on meaning and culture using computational methods (Aceves & Evans, 2023; Kozlowski et al., 2019; Poschmann et al., 2023). A recent trend in research on optimal positioning has leveraged natural language processing to represent documents produced by organizations in a semantic space as a proxy for their positioning (Haans, 2019; Majzoubi, Zhao, Zuzul, & Fisher, 2024; Vossen & Ihl, 2020). The motivation for doing so is that constructs such as prototype similarity or outlier similarity are difficult to measure on a large scale. For instance, it is generally not possible to ask market participants to rate firms in terms of their grade of membership into market categories or of their similarity to salient others when using a sample of thousands of firms. Natural language processing-based methods allow us to proxy for audiences' perceptions of such similarities, under the assumption that semantic similarities between texts produced by organizations correlate with perceived similarities between them. As part of this paper's contribution, we provide in Appendix A a detailed discussion of existing alternatives to measure constructs such as the two similarities to scale using organizational documents which we hope will help interested readers to navigate their way into this emerging field. We further detail the advantages of the method that we propose relative to other methods, namely its ability to capture semantic content that is unique to a specific document and hence representative of a focal organization's specific positioning. The method that we propose can be applied to other settings with few modifications, offering the opportunity to standardize our measurements across studies and fields.

This study has certain limitations that suggest additional avenues for future research. First, we acknowledge that our models cannot fully address unobserved heterogeneity. Therefore, we do not present our findings as causal. However, the correlations that we unveil are robust to many different specifications, which warrants further exploration. Another limitation is that the dimensions of the semantic space in which we represent documents are not readily interpretable. In future work, researchers may delve into the dimensions along which organizations tend to be similar to prototypes or outliers using, for example, a sample of hand-coded texts. To simplify our theory, we did not address investor heterogeneity in the IPO market. Yet, investors may differ in several ways – for instance, institutional and retail investors, as well as long-term and short-term investors, are known to value firms differently (DesJardine et al., 2021; Jenkinson & Jones, 2009; Martens et al., 2007). This means that the associations between the two similarities and underpricing that we observe may be driven by specific types of investors – possibly by less sophisticated retail investors and more opportunistic short-term investors. Another possible caveat related to investor heterogeneity is that investors may switch from prototype-based evaluation to alternative evaluation modes (Gouvard & Durand, 2023) when confronted with firms with high outlier similarity, which may contribute to explaining part of our result. We hope that future research will be able to disentangle these effects using additional data on investors.

Conclusion

This paper shows that salient outliers influence audiences' valuation in markets. However, unlike other reference points used by audience members to support their evaluation, most notably category prototypes, which generally help audience members converge in their valuations, outliers beget confusion. While identifying outliers and measuring outlier similarity is not easy, natural language processing methods applied to large corpuses of corporate documents can help us to do so, as this paper illustrates. They can thus help researchers clear out confusion around audiences' valuations in markets. More broadly, natural language processing methods constitute a unique tool to study meaning making in markets and organizations as well as its consequences for collectives which, we hope, organizational researchers will seize to bring organizational research on meaning and culture across new frontiers.

Acknowledgements

We would like to warmly thank the editorial team of the Special Issue, Jan Goldenstein, Dennis Jancsary, Stine Grodal, Bernard Forgues, and P. Devereaux Jennings, editor-in-chief, Tammar Zilber, and three anonymous reviewers who were crucial to shaping this paper. We are also grateful for the friendly comments from participants to the AOM 2021 on an earlier version of this paper. We would also like to thank the members of the SNO Institute for feedback on early versions of this paper. Finally, we thank Emanuele Bettinazzi for his friendly feedback during informal discussions.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Paul Gouvard https://orcid.org/0000-0001-8730-8005

Notes

1. As we focus on investors valuing issuing firms at a given point in time, we take categories as given within the context of a single evaluation (Hannan et al., 2019). However, through repeated interactions between audiences and producers – and notably repeated evaluations – market categories may change as audiences and producers co-construct the meanings and values associated with them (e.g. Khaire & Wadhwani, 2010; Krabbe & Grodal, 2023).
2. See e.g. www.wsj.com/finance/stocks/stocks-magnificent-seven-tesla-ea645bc0
3. Importantly, that Tesla is a salient outlier in financial markets does not imply that it is not also an exemplar of the electric vehicle manufacturer category. We focus on its salient outlier status for the purpose of our example. Salient outliers are defined at the market level as they attract the attention of all market participants. In this respect, they differ from the exemplars studied in Barlow et al. (2019) and the failures and successes studied in Soublière and Gehman (2020), which are defined within existing categories.
4. See e.g. https://fortune.com/2024/01/10/2024-ipo-outlook-blockbuster-is-mostly-dead/ ; https://www.bloomberg.com/view/articles/2020-12-16/doordash-airbnb-wish-ipo-pops-show-wall-street-works-just-fine?embedded-checkout=true ; https://www.ft.com/content/f29db698-1c69-4a03-b27f-f6c8d0a0bddd ; https://www.ft.com/content/ffe2e371-cc9e-41cc-94d4-b6d3446b6ed7
5. For instance, one commentator stated: 'Cava now trades at more than five times its theoretical annualized sales based on the first quarter, much higher than the 1.4 times Sweetgreen commands. That's closer to Chipotle Mexican Grill's (CMG.N), opens new tab multiple, but Chipotle, unlike Cava, has been profitable for years and operates more than 10 times as many restaurants' (www.reuters.com/breakingviews/cava-ipo-comes-little-too-hot-2023-06-16/) ; Another indicated : 'By one measure, the company's opening pop of 91% marked the best debut since July 2021 for a firm listing on a US exchange that raised more than $100 million, data compiled by Bloomberg show' (https://fortune.com/2023/06/15/restaurant-cava-shares-ipo-trading/). For other interpretations of Cava's first-day pop and speculations on its drivers in the financial press, see e.g. https://slate.com/business/2023/06/cava-stock-ipo-investors-fast-casual-mediterranean.html ; https://fortune.com/2023/07/07/

cava-shares-deemed-terribly-overpriced-by-skeptics-following-big-ipo-and-sell-side-analysts-are-about-to-weigh-in/

[6] www.nytimes.com/2019/05/02/technology/beyond-meat-ipo-stock-price.html#:~:text=Beyond%20Meat%2C%20which%20makes%20vegetarian,history

[7] www.ft.com/content/a1c5cc26-b224-470a-84fe-8a6575fd33dc

[8] www.reuters.com/article/us-poshmark-ipo-idINKBN29J2B6/

[9] See e.g. https://techcrunch.com/2023/05/23/cava-ipo-analysis/

[10] Importantly the influence of outlier similarity on investors' valuations may be conscious or unconscious, depending on whether investors actively compare a target issuing firms to salient outliers or are passively influenced by similarities between a target issuing firm and salient outliers.

[11] A fourth possible mechanism is that in front of ambiguity, investors may adopt goal-based evaluation (Boulongne & Durand, 2021; Gouvard & Durand, 2023) and evaluate issuing firms based on their alignment with investors' idiosyncratic goals. If investors have different goals, this will generate greater divergence.

[12] For simplicity we do not distinguish between different types of IPO investors in the development of hypothesis 1 and hypothesis 2. However, it could be that specific types of investors are more sensible to specific similarities. For instance, relatively less sophisticated retail investors could be relying more heavily on the two similarities than more sophisticated institutional investors. We return to this eventuality in the Discussion section.

[13] Note that this effect might be stronger when salient outliers are also categorical anomalies and hence particularly ambiguous, although we do not find evidence of such an interaction effect in the IPO setting.

[14] Results are robust to measures of prototype similarity based on 2- and 4-digit SIC codes (cf. robustness checks).

[15] To test the validity of our measure of prototype similarity, we created a measure of 'coverage mismatch' for the fiscal year following a firm's IPO, which has been used as a measure of atypicality (Bowers, 2015; Zuckerman, 1999, 2004) using data on analysts' coverage from I/B/E/S. Our measure of prototype similarity has a negative and significant correlation with coverage mismatch (−0.19, $p < 0.00001$). In an unreported OLS regression analysis, we found that prototype similarity is negatively and significantly (−0.17, $p = 0.015$) associated with coverage mismatch post-IPO when controlling for firm age, size, industry and IPO year. Taken together these findings suggest that issuing firms with high prototype similarity tend to receive more coherent coverage by analysts since they appear more typical of their industry category.

[16] More precisely, $Average\,return\,n\,industry = \frac{1}{N}\sum_{i=1}^{N} R_i$ and $Average\,volatility\,in\,industry = \frac{1}{N}\sum_{i=1}^{N} V_i$, where N is the number of publicly listed firms in an issuing firm's industry, i indexes publicly listed firms in that industry, R_i is the stock return and V_i is the volatility over the three-month period that precedes an issuing firm's IPO of publicly listed firm i . V_i is computed as the standard deviation of the residual of a regression of firm i 's daily returns on market returns.

[17] The VIX is provided by CBOE, measures the expected 30-day volatility of the S&P 500 and is widely used to measure the level of 'fear' or uncertainty in financial markets. For more details see https://cdn.cboe.com/api/global/us_indices/governance/Volatility_Index_Methodology_Cboe_Volatility_Index.pdf

[18] We use Kile and Phillips' (2009) list of 3-digit SIC codes high-tech industries.

[19] We use this control in an interaction in additional analyses, so we mean-centred it to facilitate interpretation.

[20] The results are not included due to space limitations but are available on request.

[21] Based on model 3, firms whose outlier similarity is one standard deviation above the sample mean experience a level of underpricing 5.6% higher than that of firms with an average level of outlier similarity, all else equal. Thus, if we consider two 'average' firms A and B with a market capitalization on the IPO date of $700 million (the sample average), with A having outlier similarity one standard deviation above that of B, at the end of the first day of trading, A will have a market capitalization $39 million above that of B. Prototype similarity has an effect comparable in magnitude in low-tech categories. Indeed, based on model 5, low-tech issuing firms whose prototype similarity is one standard deviation above the sample

mean (i.e. +0.1, Table 1) experience a level of underpricing which is 5.2% lower than that of a firm with an average level of prototype similarity, all else equal. However, prototype similarity does not have a significant effect on underpricing in high-tech categories (56.5% of issuing firms in our sample) and has no significant effect when considering the entire sample.

[22] The correlation between prototype similarity and outlier similarity is weakly negative (−0.1, p-value < 0.001).

[23] Note that there are two types of document embedding models: distributed memory models, the kind we use in this article and that have this property, and distributed bag-of-words models, which do not have this property.

[24] Note however that one could interpret the meaning of semantic dimensions in embedding models by looking at the words that are most similar to each dimension.

[25] The magnitude of an embedding is irrelevant to interpreting its meaning. Two embeddings with very different magnitudes might have similar meanings if they point to the same direction within the semantic space.

References

Aceves, Pedro, & Evans, James A. (2023). Mobilizing conceptual spaces: How word embedding models can inform measurement and theory within organization science. *Organization Science*, *35*, 769–1202.

Akerlof, George A. (1970). The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, *84*, 488–500.

Arthurs, Jonathan D., Hoskisson, Robert E., Busenitz, Lowell W., & Johnson, Richard A. (2008). Managerial agents watching other agents: Multiple agency conflicts regarding underpricing in IPO firms. *Academy of Management Journal*, *51*, 277–294.

Askin, Noah, & Mauskapf, Michael (2017). What makes popular culture popular? Product features and optimal differentiation in music. *American Sociological Review*, *82*, 910–944.

Barlow, Matthew A., Verhaal, J. Cameron, & Angus, Ryan W. (2019). Optimal distinctiveness, strategic categorization, and product market entry on the Google Play app platform. *Strategic Management Journal*, *40*, 1219–1242.

Biais, Bruno, & Faugeron-Crouzet, Anne Marie (2002). IPO auctions: English, Dutch, . . . French, and Internet. *Journal of Financial Intermediation*, *11*(1), 9–36.

Blackwell, Matthew, Iacus, Stefano, King, Gary, & Porro, Giuseppe (2009). CEM: Coarsened exact matching in Stata. *The Stata Journal*, *9*, 524–546.

Boulongne, Romain, & Durand, Rodolphe (2021). Evaluating ambiguous offerings. *Organization Science*, *32*, 257–272.

Bowers, Anne (2015). Relative comparison and category membership: The case of equity analysts. *Organization Science*, *26*, 571–583.

Carpenter, Mason A., Pollock, Timothy G., & Leary, Myleen M. (2003). Testing a model of reasoned risk-taking: Governance, the experience of principals and agents, and global strategy in high-technology IPO firms. *Strategic Management Journal*, *24*, 803–820.

Carter, Richard, & Manaster, Steven (1990). Initial public offerings and underwriter reputation. *Journal of Finance*, *45*, 1045–1067.

Chang, Jonathan, Gerrish, Sean, Wang, Chong, Boyd-Graber, Jordan L., & Blei, David M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, *32*, 288–296.

Cohen, Boyd D., & Dean, Thomas J. (2005). Information asymmetry and investor valuation of IPOs: Top management team legitimacy as a capital market signal. *Strategic Management Journal*, *26*, 683–690.

Corritore, Matthew, Goldberg, Amir, & Srivastava, Sameer B. (2020). Duality in diversity: How intrapersonal and interpersonal cultural heterogeneity relate to firm performance. *Administrative Science Quarterly*, *65*, 359–394.

Dai, Andrew M., Olah, Christopher, & Le, Quoc V. (2015). Document Embedding with Paragraph Vectors. NIPS 2014 Deep learning workshop. http://arxiv.org/abs/1507.07998

Delmestri, Giuseppe, Wezel, Filippo Carlo, Goodrick, Elizabeth, & Washington, Marvin (2020). The hidden paths of category research: Climbing new heights and slippery slopes. *Organization Studies*, *41*, 909–920.

Demers, Elizabeth, & Lewellen, Katharina (2003). The marketing role of IPOs: Evidence from internet stocks. *Journal of Financial Economics*, *68*, 413–437.

DesJardine, Mark R., Marti, Emilio, & Durand, Rodolphe (2021). Why activist hedge funds target socially responsible firms: The reaction costs of signaling corporate social responsibility. *Academy of Management Journal*, *64*, 851–872.

Durand, Rodolphe, & Khaire, Mukti (2017). Where do market categories come from and how? Distinguishing category creation from category emergence. *Journal of Management*, *43*, 87–110.

Durand, Rodolphe, & Paolella, Lionel (2013). Category stretching: Reorienting research on categories in strategy, entrepreneurship, and organization theory: Reorienting research on categories. *Journal of Management Studies*, *50*, 1100–1123.

Durand, Rodolphe, & Thornton, Patricia H. (2018). Categorizing institutional logics, institutionalizing categories: A review of two literatures. *Academy of Management Annals*, *12*, 631–658.

Etzion, Dror, & Ferraro, Fabrizio (2010). The role of analogy in the institutionalization of sustainability reporting. *Organization Science*, *21*, 1092–1107.

Glaser, Vern L., Atkinson, Mariam Krikorian, & Fiss, Peer C. (2020). Goal-based categorization: Dynamic classification in the display advertising industry. *Organization Studies*, *41*, 921–943.

Goldstein, Michael A., Irvine, Paul, & Puckett, Andy (2011). Purchasing IPOs with commissions. *Journal of Financial and Quantitative Analysis*, *46*, 1193–1225.

Gollnhofer, Johanna, & Bhatnagar, Kushagra (2021). Investigating category dynamics: An archival study of the German food market. *Organization Studies*, *42*, 245–268.

Gondat-Larralde, Céline, & James, Kevin R. (2008). IPO pricing and share allocation: The importance of being ignorant. *Journal of Finance*, *63*, 449–478.

Gouvard, Paul, & Durand, Rodolphe (2023). To be or not to be (typical): Evaluation-mode heterogeneity and its consequences for organizations. *Academy of Management Review*, *48*, 659–680.

Gouvard, Paul, Goldberg, Amir, & Srivastava, Sameer (2023). Doing organizational identity: Earnings surprises and the performative atypicality premium. *Administrative Science Quarterly*, 68, 781–823.

Haans, Richard F. J. (2019). What's the value of being different when everyone is? The effects of distinctiveness on performance in homogeneous versus heterogeneous categories. *Strategic Management Journal*, *40*, 3–27.

Hanley, Kathleen Weiss (1993). The underpricing of initial public offerings and the partial adjustment phenomenon. *Journal of Financial Economics*, *34*, 231–250.

Hannan, Michael T., Mens, Gaël Le, Hsu, Greta, Kovács, Balázs, Negro, Giacomo, Pólos, László, et al. (2019). *Concepts and categories: Foundations for sociological and cultural analysis* (1st ed.). New York, NY: Columbia University Press.

Hannigan, Timothy R., Haans, Richard F. J., Vakili, Keyvan, Tchalian, Hovig, Glaser, Vern L., Wang, Milo Shaoqing, et al. (2019). Topic modeling in management research: Rendering new theory from textual data. *Academy of Management Annals*, *13*, 586–632.

Hoberg, Gerard, & Phillips, Gordon (2010). Product market synergies and competition in mergers and acquisitions: A text-based analysis. *Review of Financial Studies*, *23*, 3773–3811.

Hoberg, Gerard, & Phillips, Gordon (2018). Conglomerate industry choice and product language. *Management Science*, *64*, 3735–3755.

Hsu, Greta (2006). Jacks of all trades and masters of none: Audiences' reactions to spanning genres in feature film production. *Administrative Science Quarterly*, *51*, 420–450.

Hsu, Greta, & Grodal, Stine (2015). Category taken-for-grantedness as a strategic opportunity: The case of light cigarettes, 1964 to 1993. *American Sociological Review*, *80*, 28–62.

Hsu, Greta, Koçak, Özgecan, & Hannan, Michael T. (2009). Multiple category memberships in markets: An integrative theory and two empirical tests. *American Sociological Review*, *74*, 150–169.

Hsu, Greta, Roberts, Peter W., & Swaminathan, Anand (2012). Evaluative schemas and the mediating role of critics. *Organization Science*, *23*, 83–97.

Ibbotson, Roger G., Sindelar, Jody L., & Ritter, Jay R. (1994). The market's problems with the pricing of initial public offerings. *Journal of Applied Corporate Finance*, *7*(1), 66–74.

Innis, Benjamin D. (2022). Category change in cultural fields: Practice deviation and the discursive maintenance of category meanings in jazz music. *Organization Studies*, *43*, 1745–1767.

Jenkinson, Tim, & Jones, Howard (2009). IPO pricing and allocation: A survey of the views of institutional investors. *Review of Financial Studies*, *22*, 1477–1504.

Kennedy, Mark Thomas, Lo, Jade Yu-Chieh, & Lounsbury, Michael (2010). Category currency: The changing value of conformity as a function of ongoing meaning construction. In Greta Hsu, Giacomo Negro, & Özgecan Koçak (Eds.), Categories in markets: Origins and evolution. *Research in the Sociology of Organizations* (Vol. 31, pp. 369–397). Bingley, UK: Emerald Group Publishing Limited.

Ketokivi, Mikko, Mantere, Saku, & Cornelissen, Joep (2017). Reasoning by analogy and the progress of theory. *Academy of Management Review*, *42*, 637–658.

Khaire, Mukti, & Wadhwani, R. Daniel (2010). Changing landscapes: The construction of meaning and value in a new market category—Modern Indian art. *Academy of Management Journal*, *53*, 1281–1304.

Kile, Charles O., & Phillips, Mary E. (2009). Using industry classification codes to sample high-technology firms: Analysis and recommendations. *Journal of Accounting, Auditing & Finance*, *24*, 35–58.

Kim, Bo Kyung, & Jensen, Michael (2011). How product order affects market identity: Repertoire ordering in the US opera market. *Administrative Science Quarterly*, *56*, 238–256.

Kovács, Balázs, & Hannan, Michael T. (2010). The consequences of category spanning depends on contrast. In Greta Hsu, Giacomo Negro, & Özgecan Koçak (Eds.), Categories in markets: Origins and evolution. *Research in the Sociology of Organizations* (Vol. 31, pp. 175–201). Bingley, UK: Emerald Group Publishing Limited.

Kozlowski, Austin C., Taddy, Matt, & Evans, James A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, *84*, 905–949.

Krabbe, Anders Dahl, & Grodal, Stine (2023). The aesthetic evolution of product categories. *Administrative Science Quarterly*, *68*, 734–780.

Lamont, Michèle (2012). Toward a comparative sociology of valuation and evaluation. *Annual Review of Sociology*, *38*, 201–221.

Le, Quoc, & Mikolov, Tomas (2014). Distributed representations of sentences and documents. *Proceedings of Machine Learning Research*, *32*, 1188–1196. http://proceedings.mlr.press/v32/le14.html

Lau, Jey Han, & Baldwin, Timothy (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. In Association for Computational Linguistics (Ed.), Proceedings of the 1st Workshop on Representation Learning for NLP, Berlin, Germany (pp, 78–86).

Lee, Peggy M., & Wahal, Sunil (2004). Grandstanding, certification and the underpricing of venture capital backed IPOs. *Journal of Financial Economics*, *73*, 375–407.

Leung, Ming D., & Sharkey, Amanda J. (2014). Out of sight, out of mind? Evidence of perceptual factors in the multiple-category discount. *Organization Science*, *25*, 171–184.

Lix, Katharina, Goldberg, Amir, Srivastava, Sameer B., & Valentine, Melissa A. (2022). Aligning differences: Discursive diversity and team performance. *Management Science*, *68*, 8430–8448.

Loughran, Tim, & McDonald, Bill (2013). IPO first-day returns, offer price revisions, volatility, and form S-1 language. *Journal of Financial Economics*, *109*, 307–326.

Loughran, Tim, & McDonald, Bill (2017). The use of EDGAR filings by investors. Journal of Behavioral Finance, 18, 231–248.

Loughran, Tim, & Ritter, Jay R. (2002). Why don't issuers get upset about leaving money on the table in IPOs? *Review of Financial Studies*, *15*, 413–443.

Loughran, Tim, & Ritter, Jay. (2004). Why has IPO underpricing changed over time? Financial Management, 5–37.

Majzoubi, Majid, Zhao, Eric Yanfei, Zuzul, Tiona, & Fisher, Greg (2024). The double-edged sword of exemplar similarity. *Organization Science*. https://doi.org/10.1287/orsc.2022.16855

Martens, Martin L., Jennings, Jennifer E., & Jennings, P. Devereaux (2007). Do the stories they tell get them the money they need? The role of entrepreneurial narratives in resource acquisition. *Academy of Management Journal*, *50*, 1107–1132.

Mervis, Carolyn B., & Rosch, Eleanor (1981). Categorization of natural objects. *Annual Review of Psychology*, *32*, 89–115.

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S., & Dean, Jeff (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling,

Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26* (pp. 3111–3119). New York, NY: Curran Associates, Inc. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

Murphy, Gregory L., & Ross, Brian H. (2005). The two faces of typicality in category-based induction. *Cognition*, *95*, 175–200.

Naumovska, Ivana, & Zajac, Edward J. (2022). How inductive and deductive generalization shape the guilt-by-association phenomenon among firms: Theory and evidence. *Organization Science*, *33*, 373–392.

Negro, Giacomo, & Leung, Ming D. (2013). ‘Actual’ and perceptual effects of category spanning. *Organization Science*, *24*, 684–696.

Ozmel, Umit, Reuer, Jeffrey J., & Wu, Cheng-Wei (2017). Interorganizational imitation and acquisitions of high-tech ventures. *Strategic Management Journal*, *38*, 2647–2665.

Paolella, Lionel, & Durand, Rodolphe (2016). Category spanning, evaluation, and performance: Revised theory and test on the corporate law market. *Academy of Management Journal*, *59*, 330–351.

Park, Haemin Dennis, & Patel, Pankaj C. (2015). How does ambiguity influence IPO underpricing? The role of the signalling environment. *Journal of Management Studies*, *52*, 796–818.

Park, U. David, Borah, Abhishek, & Kotha, Suresh (2016). Signaling revisited: The use of signals in the market for IPOs. *Strategic Management Journal*, *37*, 2362–2377.

Pedeliento, Giuseppe, Andreini, Daniela, & Dalli, Daniele (2020). From Mother’s Ruin to ginaissance: Emergence, settlement and resettlement of the gin category. *Organization Studies*, *41*, 969–992.

Pollock, Timothy G., & Rindova, Violina P. (2003). Media legitimation effects in the market for initial public offerings. *Academy of Management Journal*, *46*, 631–642.

Pollock, Timothy G., Rindova, Violina P., & Maggitti, Patrick G. (2008). Market watch: Information and availability cascades among the media and investors in the U.S. IPO market. *Academy of Management Journal*, *51*, 335–358.

Pontikes, Elizabeth G. (2012). Two sides of the same coin: How ambiguous classification affects multiple audiences’ evaluations. *Administrative Science Quarterly*, *57*, 81–118.

Pontikes, Elizabeth G., & Barnett, William P. (2017). The non-consensus entrepreneur: Organizational responses to vital events. *Administrative Science Quarterly*, *62*, 140–178.

Poschmann, Philipp, Goldenstein, Jan, Büchel, Sven, & Hahn, Udo (2023). A vector space approach for measuring relationality and multidimensionality of meaning in large text collections. *Organizational Research Methods*. https://doi.org/10.1177/10944281231213068.

Pozner, Jo-Ellen, DeSoucey, Michaela, Verhaal, J. Cameron, & Sikavica, Katarina (2022). Watered down: Market growth, authenticity, and evaluation in craft beer. *Organization Studies*, *43*, 321–345.

Reed, Stephen K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 382–407.

Ritter, Jay R., & Welch, Ivo (2002). A review of IPO activity, pricing, and allocations. *The Journal of Finance*, *57*, 1795–1828.

Rock, Kevin (1986). Why new issues are underpriced. *Journal of Financial Economics*, *15*, 187–212.

Rothaermel, Frank T. (2020). *Tesla, Inc*. Harvard Business Publishing Education. https://hbsp.harvard.edu/product/MH0067-PDF-ENG

Sharkey, Amanda J. (2014). Categories and organizational status: The role of industry status in the response to organizational deviance. *American Journal of Sociology*, *119*, 1380–1433.

Slavich, Barbara, & Castellucci, Fabrizio (2016). Wishing upon a star: How apprentice-master similarity, status and career stage affect critics’ evaluations of former apprentices in the haute cuisine industry. *Organization Studies*, *37*, 823–843.

Smith, Edward Bishop (2011). Identities as lenses: How organizational identity affects audiences’ evaluation of organizational performance. *Administrative Science Quarterly*, *56*, 61–94.

Soublière, Jean-François, & Gehman, Joel (2020). The legitimacy threshold revisited: How prior successes and failures spill over to other endeavors on Kickstarter. *Academy of Management Journal*, *63*, 472–502.

Vossen, Alexander, & Ihl, Christoph (2020). More than words! How narrative anchoring and enrichment help to balance differentiation and conformity of entrepreneurial products. *Journal of Business Venturing*, *35*, 106050.

Wry, Tyler, Lounsbury, Michael, & Jennings, P. Devereaux (2014). Hybrid vigor: Securing venture capital by spanning categories in nanotechnology. *Academy of Management Journal*, *57*, 1309–1333.

Wu, Cheng-Wei, & Reuer, Jeffrey J. (2021). Acquirers' reception of signals in M&A markets: Effects of acquirer experiences on target selection. *Journal of Management Studies*, *58*, 1237–1266.

Zhao, Eric Yanfei, & Glynn, Mary Ann (2022). Optimal distinctiveness: On being the same and different. *Organization Theory*, *3*(1). https://doi.org/10.1177/26317877221079340

Zhao, Eric Yanfei, Ishihara, Masakazu, Jennings, P. Devereaux, & Lounsbury, Michael (2018). Optimal distinctiveness in the console video game industry: An exemplar-based model of proto-category evolution. *Organization Science*, *29*, 588–611.

Zuckerman, Ezra W. (1999). The categorical imperative: Securities analysts and the illegitimacy discount. *American Journal of Sociology*, *104*, 1398–1438.

Zuckerman, Ezra W. (2004). Structural incoherence and stock market activity. *American Sociological Review*, *69*, 405–432.

Zuckerman, Ezra W. (2016). Optimal distinctiveness revisited: An integrative framework for understanding the balance between differentiation and conformity in individual and organizational identities. In Michael G. Pratt, Majken Schultz, Blake E. Ashforth, & Davide Ravasi (Eds.), *The Oxford handbook of organizational identity* (pp. 183–189). Oxford: Oxford University Press.

Zuckerman, Ezra W. (2017). The categorical imperative revisited: Implications of categorization as a theoretical tool. In Rodolphe Durand, Nina Granqvist, & Anna Tyllström (Eds.), *From categories to categorization: Studies in sociology, organizations and strategy at the crossroads*. Research in the Sociology of Organizations (Vol. 51, pp. 31–68). Bingley, UK: Emerald Publishing Limited.

Author biographies

Paul Gouvard was an assistant professor at Università della Svizzera italiana and is an assistant professor at ESSEC Business School since September 2024. His research explores how the meanings conveyed by organizations and their representatives impact their performance and their (e)valuations by external audiences. His research often relies on advanced computational analysis of texts produced by organizations and their representatives to study the influence of the meanings that they convey at scale. His work has been published in journals such as *Academy of Management Review* and *Administrative Science Quarterly.*

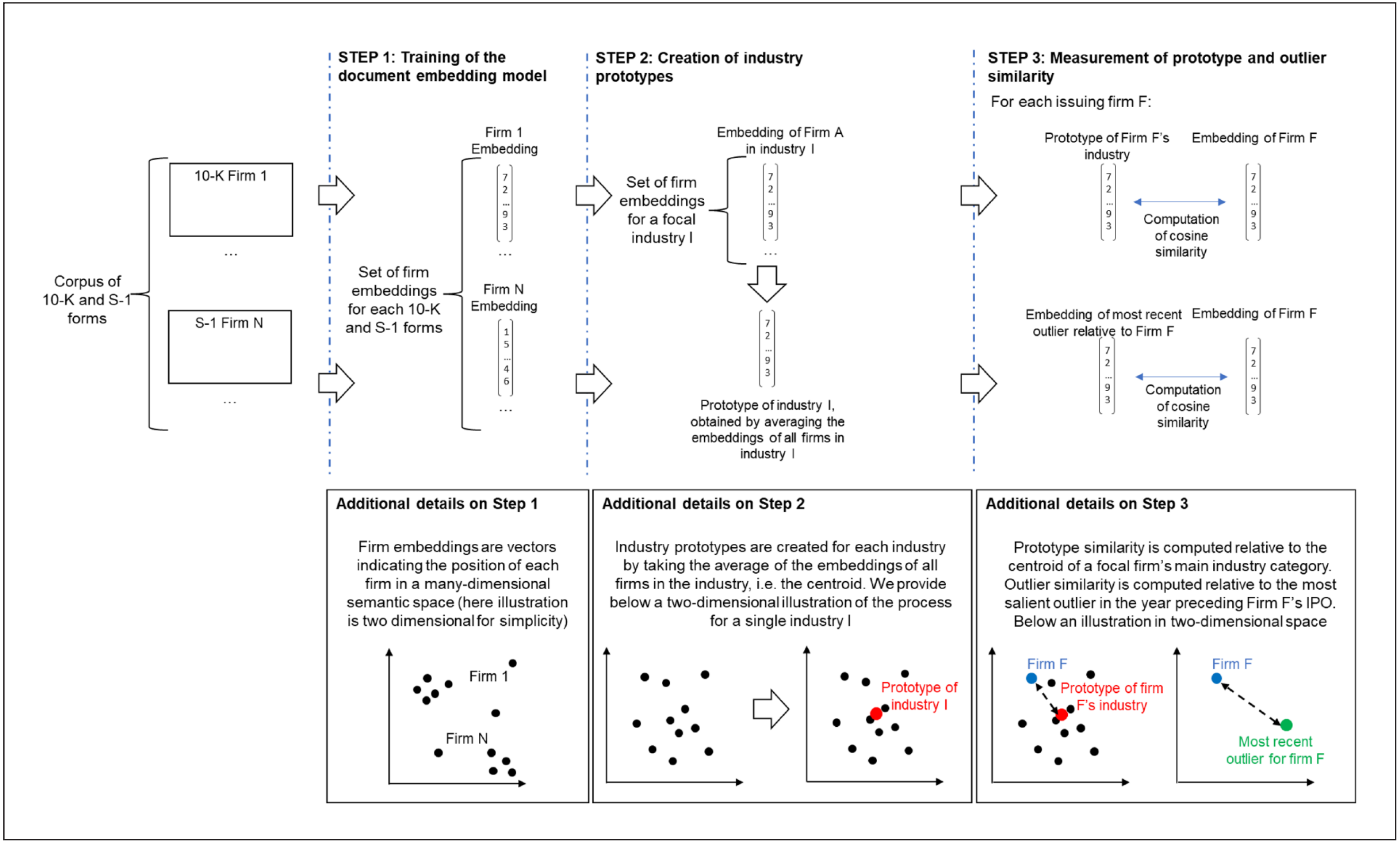
Rodolphe Durand is full professor, holder of the Joly Family Purposeful Leadership Chair at HEC Paris. His primary research interests concern the normative and cognitive dimensions of firms' performance, and especially the consequences for firms of identifying and coping with the current major environmental and social challenges. His research has been published in journals including *American Journal of Sociology, Academy of Management Review* and *Strategic Management Journal*. For his work, Rodolphe received the American Sociological Association's R. Scott Award in 2005, the European Academy of Management/Imagination Lab Award for Innovative Scholarship in 2010, and was inducted Fellow of the Strategic Management Society in 2014.

Appendix A: In-depth discussion of document embeddings

To capture similarities among firms, we relied on semantic similarities among documents produced by firms, which capture underlying industry category membership (Hoberg & Phillips, 2010, 2018) and pairwise similarities in market positioning (Barlow et al., 2019). Specifically, we (1) represented firms in a semantic space by applying a document embedding model to firms' annual reports and IPO prospectuses, (2) constructed prototypes and measured prototype similarity and (3) measured outlier similarity. We provided an overview of our entire approach in Figure A1.

Description of document embedding models and alternative methods

To represent firms in semantic space, we applied a document embedding model to IPO prospectuses and annual reports (10-Ks). Document embedding models learn to represent entire documents as vectors in a semantic space based on the words they contain (Le & Mikolov, 2014). A document embedding model creates for each document in the corpus a *document embedding* (i.e. a vector) that


STEP 1: Training of the document embedding model
STEP 2: Creation of industry prototypes
STEP 3: Measurement of prototype and outlier similarity
Corpus of 10-K and S-1 forms
10-K Firm 1
...
S-1 Firm N
...
Set of firm embeddings for each 10-K and S-1 forms
Firm 1 Embedding
7 2 ... 9 3
...
Firm N Embedding
1 5 ... 4 6
...
Set of firm embeddings for a focal industry I
Embedding of Firm A in industry I
7 2 ... 9 3
...
7 2 ... 9 3
Prototype of industry I, obtained by averaging the embeddings of all firms in industry I
For each issuing firm F:
Prototype of Firm F's industry
Embedding of Firm F
Computation of cosine similarity
Embedding of most recent outlier relative to Firm F
Embedding of Firm F
Computation of cosine similarity
Firm 1
Firm N
Prototype of industry I
Firm F
Prototype of firm F's industry
Firm F
Most recent outlier for firm F

Additional details on Step 1

Firm embeddings are vectors indicating the position of each firm in a many-dimensional semantic space (here illustration is two dimensional for simplicity)

Additional details on Step 2

Industry prototypes are created for each industry by taking the average of the embeddings of all firms in the industry, i.e. the centroid. We provide below a two-dimensional illustration of the process for a single industry I

Additional details on Step 3

Prototype similarity is computed relative to the centroid of a focal firm's main industry category. Outlier similarity is computed relative to the most salient outlier in the year preceding Firm F's IPO. Below an illustration in two-dimensional space

Figure A1. Summary of our measurement strategy for prototype and outlier similarity.

captures what is semantically specific about a particular document. The model learns document embeddings by training to predict the words that are likely to appear in a focal document given other words contained in the document.

Document embeddings are thus the appropriate tool to capture what is semantically specific about a particular document. A first alternative to this approach is to use bag-of-words representations of financial documents to locate them in semantic space (Hoberg & Phillips, 2010, 2018). A bag-of-words representation is a high-dimensional representation of a document in a semantic space where the 'dimensions' correspond to words in the vocabulary, and the coordinates of a document vector along one of these dimensions is the frequency of the word corresponding to this dimension in the document. However, bag-of-words representations are not sensitive to the semantic specificities of documents (Le & Mikolov, 2014) and thus capture their position in semantic space less accurately.

A second alternative would have consisted of using a word embedding model rather than a document embedding model. Word embedding models create *word embeddings* (i.e. vectors in a high-dimensional semantic space representing the meaning of a particular word). Word embeddings are learned by training the model to predict which word is most likely to appear given some context words (Mikolov et al., 2013). After training a word embedding model on the entire corpus, we could have located firms in the semantic space by taking the centroid of the word embeddings of the words contained in the firm's annual report or IPO prospectus (as in e.g. Gouvard, Goldberg, & Srivastava, 2023; Lix, Goldberg, Srivastava, & Valentine, 2022). However, such an approach captures less accurately what is semantically specific to a particular firm. Word embeddings summarize the semantic content of a particular word based on all of the words it co-occurs with across all documents in the corpus. Thus, they wash away document-specific variations through the use of individual words. Taking the centroid of the word embeddings of words contained in a document to represent it in semantic space does not address this problem. In contrast, since document embeddings are specifically trained to predict the words that are likely to occur in a particular document given other words contained in the document, they operate as a kind of 'distributed memory'[23] that summarizes document-specific semantics (Le & Mikolov, 2014). They thus capture more accurately what is semantically idiosyncratic to a particular firm than word embeddings and hence locate firms more precisely in the semantic space.

A third alternative would have been to use topic modelling (Hannigan et al., 2019), which has already been used to measure prototype similarity (Haans, 2019). Similar to document embedding models, topic models capture what is semantically specific to a document. Unlike document embedding models, they do so by representing documents as a probability distribution over a set of topics, which are themselves probability distributions over a vocabulary. An important difference between topics and the semantic dimensions of document embeddings is that topics are perceived as easier to interpret for humans,[24] provided that the number of topics remains relatively small; when topics are more numerous, they tend to appear nonsensical to humans (Chang, Gerrish, Wang, Boyd-Graber, & Blei, 2009). Representations of documents in topical space are thus usually less fine-grained (often anywhere from 10 to 50 topics) relative to their representations in semantic space obtained through document embedding models (usually approximately 300 semantic dimensions or more). It is possible, however, to forego topic interpretability in favour of higher-dimensional and more accurate topical representations (as in e.g. Corritore et al., 2020). Since the interpretability of topics is then lost, the two methods become relatively equivalent alternatives to capture the semantic content of documents. Ultimately, we picked document embeddings to represent firms in semantic space, as we were more interested in trying to capture accurately what is uniquely semantically specific to a particular firm than in the substantive content of topics on which firms may draw to structure their financial documents.

Measuring similarities in semantic space

After training a document embedding model on our corpus of 10-K forms and IPO prospectuses, we measured similarities between firms based on semantic similarities between the documents they produce. Two documents have similar meanings when their document embeddings point in the same general direction along many dimensions of the semantic space.[25] Thus, when the angle between two document embeddings is zero, such that they point in the same direction within the space, the implication is that their corresponding documents convey the same document-specific meanings. In contrast, when two document embeddings point in opposite directions within the space, the implication is that they convey opposite or contradictory meanings. Semantic similarities between document embeddings are captured using cosine similarity, which is a measure of the similarity between two vectors that ranges between 1 (identical meanings) and −1 (opposite meanings) commonly used in natural language processing.

Assessing model validity

One common method to establish the validity of document embeddings is to use most similar queries. Document embeddings capturing meaningful semantic similarities between texts are expected to return interpretable results when looking for documents similar to a target document. Thus, to establish the face validity of our document embedding model, we looked at the top 15 US firms in the Fortune 500 in terms of revenues and identified their five most similar peers based on similarities between document embeddings. We listed the results for 2015 in Table A1. As seen, semantic similarities between document embeddings reflected underlying similarities between firms in terms of product scope. For instance, Walmart was found to be similar to other well-known retail companies such as Costco, Home Depot or Target, while Berkshire Hathaway, which is a conglomerate notably present in insurance and railways, was found to be similar to BNSF, a railway company, or W.R. Berkley, an insurance company.

Table A1. Fifteen largest US Fortune 500 companies by revenues in 2015 along with the top 5 firms most similar to them based on cosine similarity between the document embeddings of their forms 10-K.

Company name	5 most similar firms based on cosine similarity between document embeddings
Walmart	Costco, Home Depot, Target, TJX, Wayfair
Exxon	ConocoPhillips, Chevron, Hess, Imperial Oil, Marathon Oil
Chevron	ConocoPhillips, Exxon, Hess, Marathon Oil, Murphy Oil
Berkshire Hathaway	BNSF, W.R. Berkley, Markel, Chubb, Alleghany
Apple	Citrix Systems, RealNetworks, Microsoft, Plantronics, Amazon
General Motors	Ford Motor, FCA US, Delphi Automotive, Tesla, American Axle and Manufacturing
Phillips 66	Valero, Shell, Holly Energy, Delek Logistics, MPLX
General Electric	MUFG Americas, Rockwell Automation, CIT Group, Caterpillar Financial Services, Capital One
Ford Motor	General Motors, TRW Automotive, Borgwarner, Dana Incorporated, Tesla
CVS Health	Catamaran Corp, Express Scripts, Rite Aid, PharMerica, Omnicare
McKesson	Cardinal Health, AmerisourceBergen, Omnicare, Omnicell, Walgreens Boots Alliance
AT&T	nTelos Inc., Verizon, Qwest Corporation, Shentel, CenturyLink
Valero	Alon USA Energy, PBF Energy, Phillips 66, NuStar Energy, Western Refining
UnitedHealth Group	Humana Inc., Aetna Inc., Molina Healthcare, Centene, Health Net
Verizon	AT&T, Qwest Corporation, Windstream, US Cellular, CenturyLink

Table A2. Sample IPOs along with the five established firms most similar to them.

Similarity rank	First Data Corp (2015)	Facebook Inc (2012)	General Motors Co (2010)
1	Global Payment Inc	Google Inc	Ford Motor Co
2	Mastercard Inc	CrowdGather Inc	American Axle and Manufacturing
3	Total System Services	Zynga Inc	TRW Automotive
4	Fidelity National Information Services	LinkedIn Corp	Dana Holding
5	Visa Inc	Quepasa Inc	GMAC Inc

We further provided in Table A2 a list of well-known IPOs along with the five established firms most similar to them. As seen, the firms most similar to a focal IPO engaged in similar activities: First Data was related mainly to other payment processing companies and credit card companies; Facebook was related to internet companies and social networks; and General Motors (GM) was related to other auto manufacturers and related companies (i.e. GMAC, a car financing company and former GM subsidiary).

Overall, the document embedding model successfully captured similarities among both established firms and issuing and established firms.

Using industry centroids as a proxy for prototypes

While outlier similarity can be directly measured based on the similarity between issuing firms' embeddings, prototypes are not directly observable. As shown in Figure A1, we used the centroid of established firms' embeddings within a focal industry category as a proxy for this industry's prototype. It is common in natural language processing to take the centroid of a set of semantic vectors to represent the overall meanings associated with this set. For instance, Lix et al. (2022) represent the overall meanings conveyed by web developers to their peers on Slack by taking the centroid of the embeddings of the words contained in their Slack messages. In a setting closer to our own, Haans (2019) uses the average topical distributions of firms' websites to measure prototype similarity in Dutch creative industries. The motivation for interpreting the centroid of a set of embeddings as summarizing their overall meanings comes from the fact mentioned above that in a semantic space, it is the direction of an embedding that indicates its meaning. By taking the average of a set of embeddings, a new embedding is created whose overall direction reflects the average direction of embeddings within the set – and hence captures the average meaning that they convey.

To showcase the face validity of using industry centroids as proxies for prototypes, in Table A3, we provide a list of the ten most populated three-digit SIC codes in 2015 in our sample, along with the five established firms whose firm embedding is the most similar to the centroids of these industries. As seen, established firms with high prototype similarity are specialized firms with a clear focus on the corresponding category, indicating that industry centroids are good proxies for industry prototypes.

Appendix B: Robustness checks

As part of our robustness checks, we replicated our results using alternative specifications of our main independent variables based on alternative specifications for Doc2Vec. We present these models in Tables B1 and B2.

Table A3. Ten most populated three-digit SIC codes along with the five established firms with the highest prototype similarity.

3-digits SIC code	Description of SIC code	Top 5 established companies in terms of prototype similarity
283	Drugs	Rexahn Pharmaceuticals, Rigel Pharmaceuticals, ContraVir Pharmaceuticals, Evoke Pharma, Chimerix
737	Computer and data processing services	Apigee, Jive Software, ChannelAdvisor, Linkedln, Mobilelron
131	Crude petroleum and natural gas	SM Energy, Carrizo Oil & Gas, Concho Resources, Lynden Energy, Mexco Energy
384	Medical instruments and supplies	Merit Medical, Endologix, Misonix, Vascular Solutions, Angiodynamics
367	Electronic components and accessories	Monolithic Power Systems, Micrel, Anadigics, Diodes, Semtech
491	Electric services	Northern States Power Company, Sierra Pacific Power Company, Southwestern Electric Power Company, Southern Power Company, Georgia Power Company
382	Measuring and controlling devices	Keysight Technologies, MKS Instruments, MTS Systems, Faro Technologies, Nanometrics
357	Computer and office equipment	Infoblox, Fortinet, Cavium, NetApp, Ruckus Wireless
581	Eating and drinking places	Red Robin Gourmet Burgers, Zoe's Kitchen, Buffalo Wild Wings, Chuy's, Diversified Restaurant Holdings
371	Motor vehicles and equipment	Dana Corporation, Tenneco, BorgWarner, Delphi Automotive, Allison Transmission

To address endogeneity concerns, we further re-estimate models 2 and 3 using two-stage least squares regressions. To do so, we create two exogenous instrumental variables that should be correlated with prototype and outlier similarity but not with the error term in models 2 and 3, respectively.

The instrumental variable that we use for prototype similarity is the prototype similarity of established firms that are similar to a focal issuing firm. Indeed, part of the issuing firm's prototype similarity is determined by the baseline level of prototype similarity among similar firms. Issuing firms need to adhere to this baseline tendency to be recognized as part of this group of similar firms (Zuckerman, 1999, 2016). However, there is no reason to expect that the prototype similarity of firms similar to a focal issuing firm are related to its underpricing, except through its association with the issuing firm's own prototype similarity. Formally, we compute the average prototype similarity of the five most similar peers to a focal issuing firm.

The instrumental variable that we use for outlier similarity is the outlier similarity of established firms that are similar to a focal issuing firm. We would expect a mechanical association between one and the other as, if a focal issuing firm is most similar to firms that are also similar to a salient outlier, then the issuing firm should be similar to the salient outlier as well. However, there is no reason to expect that the outlier similarity of firms similar to a focal issuing firm are related to its underpricing, except through its association with the issuing firm's own outlier similarity. Formally, we compute the average outlier similarity of the five most similar peers to a focal issuing firm.

Table B3 shows the results of our instrumental variable approach, which are consistent with our main analysis. The first-stage model for prototype similarity (model B21) shows a significant effect of the instrument on prototype similarity. The second-stage model for prototype similarity (model B22) does not show a significant association between prototype similarity and underpricing, as in model 2. The first-stage model for outlier similarity (model B23) shows a significant

Table B1. Ordinary least squares regressions of underpricing on prototype similarity under different specifications of Doc2Vec.

Variables	Model B1	Model B2	Model B3	Model B4	Model B5	Model B6	Model B7	Model B8	Model B9	Model B10
Prot. sim. (w:5, d: 100)	0.032 (0.174)	−0.402** (0.132)								
Prot. sim. (w:5, d: 100)#Tech firm		0.860* (0.345)								
Prot. sim. (w:5, d: 200)			−0.008 (0.198)	−0.508** (0.159)						
Prot. sim. (w:5, d: 200)#Tech firm				1.006* (0.413)						
Prot. sim. (w:10, d: 100)					0.004 (0.183)	−0.350** (0.124)				
Prot. sim. (w:10, d: 100)#Tech firm						0.685* (0.300)				
Prot. sim. (w:10, d: 200)							−0.000 (0.199)	−0.420** (0.141)		
Prot. sim. (w:10, d: 200)#Tech firm								0.814* (0.387)		
Prot. sim. (w:10, d: 300)									0.016 (0.220)	−0.433** (0.147)
Prot. sim. (w:10, d: 300)#Tech firm										0.859* (0.415)
Tech firm	−0.012 (0.053)	0.025 (0.044)	−0.011 (0.052)	0.028 (0.044)	−0.011 (0.053)	0.017 (0.044)	−0.011 (0.053)	0.018 (0.045)	−0.011 (0.053)	0.018 (0.045)
Controls	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Industry FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Lead underwriter FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Stock exchange FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Year Effects	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Observations	2,046	2,046	2,046	2,046	2,046	2,046	2,046	2,046	2,046	2,046
Adjusted R-squared	0.285	0.290	0.285	0.290	0.285	0.288	0.285	0.288	0.285	0.289

Robust standard errors in parentheses, ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

Table B2. Ordinary least squares regressions of underpricing on outlier similarity under different specifications of Doc2Vec.

Variables	Model B11	Model B12	Model B13	Model B14	Model B15	Model B16	Model B17	Model B18	Model B19	Model B20
Out. sim. (w:5, d: 100)	0.363***	0.372***								
	(0.090)	(0.087)								
Out. sim. (w:5, d: 100)#Temp. dist. from outlier		−0.509*								
		(0.203)								
Out. sim. (w:5, d: 200)			0.421***	0.431***						
			(0.117)	(0.116)						
Out. sim. (w:5, d: 200)#Temp. dist. from outlier				−0.589*						
				(0.242)						
Out. sim. (w:10, d: 100)					0.356***	0.374***				
					(0.101)	(0.102)				
Out. sim. (w:10, d: 100)#Temp. dist. from outlier						−0.468*				
						(0.226)				
Out. sim. (w:10, d: 200)							0.532***	0.551***		
							(0.143)	(0.145)		
Out. sim. (w:10, d: 200)#Temp. dist. from outlier								−0.578*		
								(0.248)		
Out. sim. (w:10, d: 300)									0.581***	0.609***
									(0.149)	(0.151)
Out. sim. (w:10, d: 300)#Temp. dist. from outlier										−0.638*
										(0.267)
Temporal distance from outlier (centred)	0.103+	0.061	0.106*	0.070	0.105*	0.073	0.105*	0.075	0.107*	0.081
	(0.052)	(0.060)	(0.051)	(0.059)	(0.051)	(0.060)	(0.050)	(0.059)	(0.050)	(0.059)
Controls	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Industry FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Lead underwriter FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Stock exchange FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Year Effects	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Observations	2,050	2,050	2,050	2,050	2,050	2,050	2,050	2,050	2,050	2,050
Adjusted R-squared	0.290	0.293	0.289	0.291	0.290	0.292	0.291	0.294	0.291	0.293

Robust standard errors in parentheses, $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$.

effect of the instrument on outlier similarity. The second-stage model for outlier similarity (model B24) shows a positive and significant association between outlier similarity and underpricing, as in model 3. Overall, these results are consistent with those obtained using OLS regression.

Table B3. Two-stage least square regression of underpricing on prototype and outlier similarity.

Variables	Model B21	Model B22	Model B23	Model B24
	First-Stage DV: Prototype similarity	Second-Stage DV: Underpricing	First-Stage DV: Outlier similarity	Second-Stage DV: Underpricing
Av. prototype similarity of established peers	0.602*** (0.031)			
Prototype similarity (centred)		0.284 (0.547)		
Av. outlier similarity of established peers			0.216*** (0.050)	
Outlier similarity (centred)				5.489*** (0.951)
Tech firm	0.022+ (0.011)	−0.019 (0.059)	0.005 (0.011)	−0.061 (0.055)
Temporal distance from outlier (centred)	−0.006 (0.010)	0.117* (0.053)	0.015 (0.013)	0.042 (0.092)
Offer price revision	0.001 (0.016)	0.188** (0.056)	0.034* (0.014)	−0.001 (0.088)
Market hotness	−0.000+ (0.000)	0.003*** (0.001)	0.000+ (0.000)	0.002* (0.001)
VC backing	0.007 (0.005)	0.090** (0.027)	0.011** (0.004)	0.023 (0.032)
Average VIX over past quarter	0.001 (0.001)	0.009* (0.004)	−0.001 (0.001)	0.015* (0.007)
Average volatility in industry over past quarter	0.164 (0.193)	5.174*** (0.951)	0.215 (0.230)	3.229 (2.140)
Average return in industry over past quarter	0.001 (0.010)	0.626** (0.199)	−0.035** (0.011)	0.860*** (0.165)
Log of number of days since beginning of year	−0.002 (0.005)	−0.041 (0.027)	−0.009 (0.007)	−0.002 (0.058)
Log of age	0.001 (0.004)	−0.035* (0.016)	−0.006+ (0.003)	0.001 (0.017)
Log of assets	−0.002 (0.004)	−0.017 (0.011)	−0.006*** (0.002)	0.022 (0.013)
Industry, Year, Lead underwriter, and Stock exchange FE	*YES*	*YES*	*YES*	*YES*
Observations	2,046	2,046	2,050	2,050

Robust standard errors in parentheses, $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$.

Av. prototype similarity of established peers: The underidentification test leads to rejecting the null hypothesis that the instrument is not correlated with prototype similarity (Kleibergen-Paap rk LM statistic = 7.137, $p = 0.0076$). The weak identification test suggests that the instrument is not weakly correlated with prototype similarity (Kleibergen-Paap Wald rk F statistic = 386.685).

Av. outlier similarity of established peers: The underidentification test leads to rejecting the null hypothesis that the instrument is not correlated with outlier similarity (Kleibergen-Paap rk LM statistic = 5.042, $p = 0.0247$). The weak identification test suggests that the instrument is not weakly correlated with outlier similarity (Kleibergen-Paap Wald rk F statistic = 18.396).