Causal Inference in Word-of-Mouth Research: Methods and Results*

In preparation for
Customer Analytics for Maximum Impact:
Academic Insights and Business Use Cases,
Taylor & Francis (CRC Press), edited by S. Seetharaman

Stephan Seiler
Stanford University

Song Yao
University of Minnesota

Georgios Zervas
Boston University

This draft: January 11, 2018

*Contact: Stephan Seiler, sseiler@stanford.edu; Song Yao, syao@umn.edu; Georgios Zervas, zg@bu.edu

1 Introduction

One of the biggest changes in the marketing landscape in recent years has been a shift toward fostering word-of-mouth (WOM) to let consumers advocate on a brand's behalf. Many marketing executivs consider online WOM, which has increased dramatically in volume in recent years, one of the most effective forms of marketing. Every second, 6,000 tweets are posted on Twitter, and that volume is growing at around 30% per year.[1] Similar patterns apply to other social media platforms such as Facebook, as well as to platforms that host costumer reviews. TripAdvisor and Yelp host 570 million and 142 million reviews, respectively, and are visited by 455 million and 188 million users each month.[2] However, reliably measuring the impact of WOM on demand is subject to some unique challenges, and many marketing managers admit that measuring WOM effectiveness remains difficult.[3]

In this chapter, we outline the current state of the academic literature regarding the impact of online WOM on demand. We first outline measurement challenges in the realm of WOM and how they can be resolved. We then summarize recent findings on the effectiveness of WOM in two domains: customer reviews and online conversations about brands on platforms such as Twitter or other social media. The former are a type of activity that typically occur after consumption and that impose a specific structure (often a rating scale) on consumers' WOM. The latter instead are less structured and can take place before and/or after consumption. These two areas are sufficiently different and shall be treated separately.

2 Measuring the Impact of WOM

To understand the measurement challenges and possible solutions when estimating the impact of WOM, it is instructive to draw parallels to the measurement of advertising effectiveness. The approaches to estimation in both realms share some common features, but also diverge in several ways. Importantly, methods for measuring advertising effectiveness, which are fairly well developed, cannot be directly applied to measuring the impact of WOM. Hence, marketing managers need to develop new frameworks for measuring the impact of WOM. We outline several methods below and discuss the situations in which they can be applied.

As an illustration for the general measurement framework, we consider the case of a firm offering multiple products that are characterized by different advertising and WOM intensities. The firm has historical data on marketing activities, namely, WOM and advertising, and demand (e.g., quantity sold, click-through rates, conversions) at the product level and wishes to understand the impact of WOM and advertising on demand. The general logic of our arguments below easily extends to situations in which data are available at the consumer rather than the product level. For the

[1] see http://www.internetlivestats.com/twitter-statistics/

[2] see https://tripadvisor.mediaroom.com/ and https://www.yelp.com/factsheet.

[3] The CMO Survey Report, February 2016, p.25, https://cmosurvey.org/wp-content/uploads/sites/15/2017/04/The_CMO_Survey-Topline_Report-Feb-2016.pdf.

remainder of this section, we refer generically to "WOM intensity" and do not distinguish between specific types of WOM, such as the number of tweets, number of consumer reviews, and so on.

2.1 Measurement challenges

Cross-sectional inference

A simple approach for measurement using a standard linear regression framework would be to regress some measure of demand on each type of marketing activity,

$$Demand_j = \alpha + \beta \times Marketing_j + \varepsilon_j,$$

where the index j refers to a product. One would then run a separate regression using either advertising or WOM as the explanatory marketing variable. This regression effectively compares products with high and low advertising / WOM and asks whether demand increases or decreases as a function of the marketing activity of interest. For the effect of marketing on demand, captured by the coefficient β, to be interpreted as a *causal* effect, $Marketing_j$ needs to be uncorrelated with the regression error term ε_j. Hence, any factor that might influence demand, such as the quality of the product, has to be uncorrelated with advertising / WOM. Random assignment of marketing in an A/B-test setting does assure no such correlation exists. Here, however, we consider the situation in which marketing activity is not necessarily randomly assigned. We start with a discussion of (the more familiar case of) measuring the advertising impact as a reference point, and then highlight similarities and differences regarding the WOM impact. In each case, we need to assess which underlying factors lead to a given level of the relevant activity in the historical data used for estimation.

With regards to advertising, we need to ask ourselves whether marketing activity is correlated with any factors that also affect demand directly (and that are therefore captured in ε_j when they are not controled in the regression.). A likely scenario in many settings is the case in which popular products that sold well in the past are advertised more heavily. Hence, "product popularity" is correlated with advertising intensity and also influences demand. In other words, products with high advertising levels would have enjoyed high demand even in the absence of advertising, due to their inherent popularity. Such a correlation will lead to an overestimation of the impact of advertising on demand. As we show in more detail below, an upward bias typically does seem to occur in cross-sectional studies of advertising impact.[4]

In the case of WOM, a similar type of bias can occur for slightly different reasons. In particular, a scenario of reverse causality seems likely in the case of WOM; that is, high-demand products that a lot of consumers buy are perhaps talked about more frequently. This relationship does indeed seem fairly mechanical, and it leads to a positive correlation between WOM and demand. Therefore,

[4] One notable exception in which the bias can go in the opposite direction is so-called "activity bias" in studies of online advertising (see Lewis and Reiley (2014)).

similar to the case of advertising, a simple cross-sectional analysis of the impact of WOM tends to overstate its effectiveness.

Inference with panel data

An alternative approach that deals with some of the problems outlined above is one that relies on panel data with fixed effects at the level of the cross-sectional unit. Such an analysis requires data on demand and marketing for various products over time, which is often available to firms. The fixed-effects regression takes the following form:

$$Demand_{jt} = \xi_j + \gamma \times Marketing_{jt} + \nu_{jt},$$

where ξ_j denotes the fixed effect for product j. Note also that demand and marketing are now indexed by (jt) to indicate we have product-/time-period-specific data on those variables.

In contrast to the cross-sectional approach, we are now relating changes in marketing over time to changes in demand; that is, we are asking whether in time periods with relatively higher marketing intensity for a given product, demand for that product increases. Importantly, any product characteristics that stay stable over time do not pose a problem when using this approach, because we control for the influence of such factors through product fixed effects (ξ_j). For instance, products that are more popular might have higher *levels* of marketing and demand, but this issue is unproblematic, because the regression is relating changes over time in both marketing and demand to each other, while controlling for time-invariant popularity levels through the fixed effects. Hence, in terms of threats to a causal interpretation of the impact of marketing, we need to consider whether any *time-varying* factors correlate with marketing and also have a direct effect on demand.

With regards to advertising, we need to ask whether time periods in which advertising levels are higher are systematically different from other time periods in terms of demand. Whether this is the case again depends on the process by which advertising is set and that causes it to vary over time. In many markets, relatively high-frequency variation exists in advertising, such as in consumer packaged goods, where supermarkets advertise different products through feature-advertising leaflets in different weeks. In this market, advertising is often "locked in" in advance through longer-term advertising plans between manufacturers and advertisers. Consequently, the specific weeks in which advertising is run are not chosen in a particularly strategic way (Rossi (2014), Seiler and Yao (2017)). Hence, in such cases, the panel approach is likely to yield a causal estimate. Some empirical evidence also points toward fixed-effects regressions working well in this context. Shapiro (2016) finds that estimates change little when moving from fixed effects to more rigorous controls.[5] Gordon and Hartmann (2013) find that in the case of political advertising, controlling for (market-level) fixed effects leads to a substantial reduction in effect size, whereas further controls have a smaller impact on the effect magnitude.

Unfortunately, the panel approach is less likely to work in the context of WOM, because of

[5] The paper employs a regression discontinuity design at media market boundaries together with a fixed approach.

an extension of the argument made in the cross-sectional case above. If demand increases for a given product over time, WOM intensity is likely to increase as a result of the higher demand. For instance, consider a setting where we analyze data from a TV show over time and find that when viewership of the show increased, WOM also tended to be higher. Higher viewership was likely causing higher WOM levels rather than the other way around. In the case of advertising, such changes over time could also happen. For instance, a firm might shift advertising dollars to products that experienced sales increases. However, the allocation of advertising dollars tends to adapt more slowly, whereas WOM is likely to react very rapidly to changes in demand. Hence, the panel approach is less promising in the case of WOM than in the case of advertising.

A/B tests

A/B tests (also referred to as randomized field experiments) allow for clean causal inference and are increasingly used in many academic studies of advertising effectiveness (Lewis and Reiley 2014, Blake, Nosko, and Tadelis 2015, Sahni 2015, Gordon, Zettelmeyer, Bhargava, and Chapsky 2016), as well as by practitioners Shaoolian (2017). A/B tests involve randomly allocating different amounts of advertising to different products or groups of consumers. For example, in the context of social media advertising, Gong, Zhang, Zhao, and Jiang (2017) used the company's official account to randomly send tweets for a subset of TV shows and then compare viewership of these shows with the viewership of shows about which they did not send tweets (the control group).

Such an approach circumvents the concerns raised above in the context of cross-sectional and panel regressions. If tweets (or any other type of marketing activity) are randomly assigned, they are, by construction, not correlated with the regression error, and a simple regression of the relevant outcome measure can be used to recover the causal impact. In this example, the relevant regression would be

$$Viewership_j = \alpha + \beta \times Tweet_j + \varepsilon_j,$$

where $Tweet_j$ is equal to 1 if a tweet was sent out regarding show j; that is, show j was in the treatment group.

We can see this research paradigm cannot be easily applied to the realm of WOM. Running an A/B test requires direct control of the variable of interests, for example, the company's tweets in the example above, in order to implement a random assignment of that variable. In the case of WOM, however, firms will generally not have the ability to randomly vary WOM across products. This restriction leads to one of the most effective and simple tools of causal measurement not being available in the case of measuring WOM effectiveness.

We note that situations could exist in which an experimental manipulation of WOM is feasible, but the scope for such manipulations is likely to be limited. For instance, in principle, one could imagine that a social media platform could decide not to show tweets pertaining to a randomly selected group of products. A firm that was interested in studying the impact of WOM and was able to convince the platform to cooperate in the experiment could achieve an experimental manipulation of WOM due to the random elimination of tweets across products. Our sense is that platforms will be unlikely to cooperate in such experiments. Because a backlash could arise from users whose tweets do not appear due to the A/B test. A further option would be to randomly generate rather than eliminate WOM, by generating tweets from fake accounts or incentivizing users to tweet on the firm's behalf. Such activity is, however, likely to be something the platform would oppose, and might violate the FTC's truth-in-advertising standard.

2.2 Correctly measuring causal effects

In the previous section, we argued that WOM effectiveness is difficult to measure reliably, and methods that can be applied to studies of the impact of advertising are either not available (in the case of A/B tests) or unlikely to work (in the case of fixed-effects regressions). Having outlined the obstacles to correct measurement, we now turn to discussing methods that can be used to estimate causal WOM effects.

Natural experiments

Natural experiments are a useful tool for causal inference that are sometimes available to researchers interested in measuring the causal impact of WOM on demand. Natural experiments rely on historical data that contain events that affect WOM but are otherwise unrelated to demand.

We begin by discussing a concrete example from Seiler, Yao, and Wang (2017), who study the impact of user-generated microblogging content on the viewership of Chinese TV shows. The authors have access to data on the daily number of tweets related to each show, as well as the viewership numbers for different episodes of the same show. During the time period covered by the data, Sina Weibo – the most popular Chinese microblogging platform – was unexpectedly shut down for a brief period of time. The shutdown qualifies as a natural experiment, because it was driven by political events that are unlikely to have been related to TV viewership rates. Therefore, the sudden change in microblogging activity due to the shutdown can be used to estimate the causal impact of WOM on demand.

To isolate the variation in WOM driven by the natural experiment, the authors implement an instrumental variable (IV) approach in a panel fixed-effects-regression framework (similar to the framework outlined above):

$$Viewership_{jt} = \xi_j + \gamma \times WOM_{jt} + \nu_{jt}$$

where $Viewership_{jt}$ denotes demand for the episode of show j that aired on day t and WOM_{jt} denotes the number of tweets about the show on day t. To isolate exogenous variation in WOM, the WOM measure is instrumented with a dummy variable that is equal to 1 for episodes that aired during the shutdown, and zero otherwise. This regression recovers the causal impact of WOM on demand if the instrument and the error term are uncorrelated, even if WOM itself is correlated with the error term.

In general, any event that affects WOM without directly affecting demand qualifies as a natural experiment. Another example of such an event is the merger of two review platforms studied by Lewis and Zervas (2016). Prior to the merger, the two platforms were collecting reviews separately, each platform computing average ratings using its own database of reviews. The merger of the review databases led to a sudden change in average ratings. Similar to the shutdown of Sina Weibo, the merger of the two review databases was unlikely to have been related to a change in demand.

The main difficulty in using natural experiments as a tool for causal inference is identifying such experiments. Events that cause exogenous changes in the status quo are rare, and pinpointing them often requires expert domain knowledge. Moreover, arguing for the suitability of different events as a natural experiment has to be done on a case-by-case basis. This requirement is in contrast to methods such as A/B testing and fixed-effects regressions, which can be readily implemented by firms to provide a standardized analytical toolkit. Despite this limitation, because natural experiments are usually broad in scope, once identified, they can be leveraged to analyze different product markets and outcomes of interest. For example, both the shutdown and the merger described above constitute platform-wide shocks that can be used to study WOM effects for any set of products represented on the respective platform.

A second limitation of natural experiments is that they often yield context-specific estimates. For instance, Lewis and Zervas (2016) can estimate the causal impact of ratings in July 2013, the date of the review platform merger. This natural experiment cannot provide insight as to the magnitude of the treatment effect at other points in time.

A final concern with natural experiments, especially those causing large changes to the status quo, is that firms or individuals may anticipate them and strategically adjust their behavior. In this case, comparisons of outcomes shortly before and shortly after the natural experiment can result in biased estimates. Thus, natural experiments are more convincing when evidence can be furnished that they were unpredictable. An indicative piece of evidence is the absence of pre-treatment trends. To the extent that natural experiments are unpredictable, patterns of the outcome variable just before the natural experiment should look similar to historical patterns. Unexplained systematic variation prior to the natural experiment may suggest the natural experiment is invalid.

A/B tests as instruments

A method that is related to the natural experiment approach just outlined, but more readily available, is an *indirect* manipulation of WOM.[6] The key idea is to use information on any activity a firm might be able to leverage to foster WOM among its users. For instance, a firm might be running a targeted advertising campaign to users that provided positive WOM in the past. In a setting where the firm sells different products, it could randomly assign products to such a

[6] As discussed earlier in section 2.1, A/B tests that *directly* manipulate WOM are often infeasible.

targeted advertising campaigns. Hence, for some randomly selected set of products, active users are targeted with advertising and start generating additional WOM. Relative to a standard A/B test, the random assignment has been applied not to the variable of interest directly, but to an activity (targeted advertising in this case) that is under the firm's control and correlated with the variable of interest.

An example of such an approach is the "peer encouragement" design used by Eckles, Kizilcec, and Bashky (2016). The authors randomly alter the Facebook interface of some users such that they become more likely to provide feedback (e.g., likes) to other users in their network. They then study the impact of the altered feedback behavior of the affected users on the behavior of their peers. Although this study does not measure the impact of WOM on demand but rather on user behavior within Facebook, the general framework is identical to the one described above. In this case, the assignment to the platform interface, which makes feedback more likely, serves as an instrument for the actual feedback behavior.[7]

Such an approach based on A/B tests has the advantage of being scalable. In contrast to natural experiments, we are not constrained to leveraging existing variation that might occur infrequently, but can instead generate useful (i.e., random) variation ourselves. Furthermore, we circumvent the issue of WOM not being directly under the firm's control, by manipulating a variable over which the firm does have control.

The main downside of this approach is that it might be hard to find interventions that affect WOM without also directly affecting demand. In our first example above, the targeted advertising campaign likely not only affects the amount of WOM, but also makes the users receiving advertising more likely to buy the product. Hence, the instrument will affect demand directly and not only via its impact on WOM. In this particular case, the problem could be solved if one can track the sales of users who were targeted with advertising and only study the impact of WOM on the demand of users who did not receive any advertising.

Regression Discontinuity

A final approach that can be useful to measure the impact of a specific type of WOM is a regression discontinuity (RD) approach. Although this approach is more narrow in its application, it does provide a broadly applicable approach to measuring the impact of review valence, that is, the impact of a higher average review rating on demand (e.g., Luca (2016), Anderson and Magruder (2012), and Luca and Vats (2013)).

The approach is most easily explained by focusing on the particular application to review valence. A common way of presenting the average rating score of a particular product to users is through a star rating that is rounded to the nearest half star. Yelp (as well as many other platforms) presents average ratings in this particular way. The rounding generates a discontinuous jump in the perceived rating, and two products with very similar underlying scores might take on

[7] An experiment of this kind requires the cooperation of the platform and hence might be hard to implement for individual firms.

different values on the half-star scale. For instance, two products with average ratings of 4.24 and 4.26 will be presented as 4- and 4.5-star products, respectively. We can compare such products that are characterized by similar continuous scores but different star ratings to isolate the impact of the star rating.

Note this comparison is different from comparing all products with, say, 4 versus 4.5 stars. We might reasonably believe products with higher star ratings are systematically different along other dimensions that correlate with demand. Instead, those products near the rounding threshold are likely to be similar except for their rounded star ratings. Hence, the causal impact of the star ratings can be obtained by comparing demand for products marginally above and marginally below the rounding threshold.

This approach leverages the specific institutional feature of rounded star ratings. Because of the ubiquitous usage of rounding in star ratings across platforms, the RD design is a widely applicable approach when studying the valence of reviews. Importantly, any firm that sells multiple products through a platform that uses rounded rating scores can find the subset of products in their assortment that have ratings close to the rounding cutoffs and implement this approach.

Contrary to natural experiments, RD thus provides a scalable and broadly applicable approach to measuring WOM effectiveness. The approach is, however, limited in scope to the analysis of review valence, and cannot be easily applied to study the impact of other aspects of WOM, such as the number or content of reviews.

2.3 Measurement challenges and solutions: Summary

In summary, despite unique measurement challenges in the realm of WOM, three prominent approaches to causal inference are available: natural experiments, A/B tests that are used as instruments, and a regression discontinuity (RD) approach. Our review of the existing evidence regarding the impact of WOM in the next two sections predominantly relies on studies based on these approaches.

3 Consumer Reviews

Academic interest in consumer reviews goes nearly as far back as the appearance of the first online reputation systems on e-commerce platforms such as eBay and Amazon. Reviews – the information units of reputation systems – typically consist of a numeric score (generally represented as a rating between 1 and 5 stars), the review submission date, and some open-ended text that allows consumers to evaluate a business (or a product) in their own words. Some reputation systems allow consumers to attach pictures to their reviews, and to separately rate businesses on dimensions such as service and price. Below, we discuss findings from the existing academic literature around three important aspects of consumer reviews: (1) review valence, (2) review volume, (3) and review content. We note the literature on reviews is large and we do not cover it exhaustively.

3.1 Review valence

The most common way reputation systems aggregate the plethora of information they collect is to compute an average rating for each business or product.[8] Average ratings, which are meant to capture overall quality, are prominently displayed and used to rank products and businesses in response to user queries. For instance, the query “hotels in San Francisco” on TripAdvisor is likely to return higher-rated hotels as the top search results. Maybe due to their wide use and simplicity, average ratings have been well studied. By now, average ratings are well known to have a substantial causal impact on demand, though the magnitude of this effect varies by timing and context.

eBay’s reputation system, which allows buyers and sellers on the platform to rate each other, was among the first to be studied and has been the focus of many studies (e.g., Ba and Pavlou (2002), Houser and Wooders (2006), Lucking-Reiley, Bryan, Prasad, and Reeves (2007), Eaton (2002), Bajari and Hortacsu (2003), Kalyanam and McIntyre (2001), McDonald and Slawson (2002), Cabral and Hortacsu (2010), Dewally and Ederington (2006), Jin and Kato (2006)). The findings of these papers, which rely on different methods and study different eBay product categories, are broadly consistent: highly rated sellers attract more bidders in their auctions, fetch higher prices, and sell their items with higher probability.

Similar effects for review valence have been demonstrated on other review and e-commerce platforms. Chevalier and Mayzlin (2006) examine the impact of review valence on book sales, and find economically significant effects. Looking at Yelp, Luca (2016) and Anderson and Magruder (2012) estimate the impact of average ratings on restaurant demand and respectively find that a one-star increase in ratings causes a 5% increase in revenue and a 50% increase in the probability of being sold out. To establish causality, both of these studies rely on the RD design described at the end of section 2.2. Using the same RD strategy, Luca and Vats (2013) study ZocDoc, a platform that allows consumers to rate doctors and make appointments with them, and find that a half-star increase in ratings is associated with a 10% increase in the likelihood that an appointment will be filled.

More recently, reviews and ratings have become important components of peer-to-peer markets such as Airbnb and Uber. These markets connect consumers with individual suppliers, most of whom have no outside reputation. To inform consumers about the quality of suppliers, peer-to-peer markets rely on reviews. Proserpio, Xu, and Zervas (2016) estimate the impact of Airbnb ratings on Airbnb properties using a natural experiment: the average rating of each Airbnb property is only disclosed to consumers once the property has accumulated three reviews. These disclosures can be used to estimate a causal effect under the assumption that the arrival of the third review is unpredictable, and thus uncorrelated with unobserved variation in demand. The study finds the disclosure of a perfect 5-star rating raises prices by approximately 2%.

Although many studies have investigated the impact of average ratings, less is known about

[8] An average rating is simply the *unweighted* mean of all individual ratings a business or product has received, often rounded to the nearest decimal point or half-star.

other aggregate measures. One exception is the work of Sun (2012), who studies the variance of ratings. Sun (2012) presents empirical evidence that increased variance in ratings has a positive impact on the demand of low-rated products. The key intuition behind this result is that low-rated products with some high individual ratings (i.e., high variance) may be appealing to at least a certain subset of consumers, whereas products with consistently low ratings (i.e., low variance) are less likely to be appealing to anyone.

3.2 Review volume

Another salient and well-studied feature of reputation systems is review volume. Review counts are typically displayed prominently next to average ratings, and consumers can also use them to infer quality. All else equal, we may expect that consumers will have less uncertainty about the quality of products with more reviews. Therefore, increases in review volume could lead to increased demand. Although the hypothesis that changes in review volume affect demand is plausible, testing it raises an identification challenge of the type we discussed in section 2.1: although more reviews may cause sales, the reverse is also true; that is, increased sales can result in more reviews.

Chevalier and Mayzlin (2006) resolve this identification challenge by combining data from Amazon and Barnes & Noble. They implement an identification strategy that relates changes in book sales over time on one retailer relative to the other to differential changes in reviews across the two retailers. The study finds that review volume has a significant impact on demand: a 10% increase in review volume decreases a book's sales rank by 2%.[9]

As discussed in section 2.1, WOM experiments are not common. A notable exception is experiments that randomize the assignment of reviews and subsequently compare the outcomes of sellers treated with a review to sellers that were randomly assigned to a no-review condition. Even though these experiments do not manipulate review volume at the margin, they yield interesting insights into the economic benefits of establishing a reputation, for example, by having at least one review versus not being reviewed at all. Resnick, Zeckhauser, Swanson, and Lockwood (2006) and Pallais (2014) implement experiments of this nature.

In the Resnick, Zeckhauser, Swanson, and Lockwood (2006) study, an eBay postcard seller with an established reputation was asked to create a set of new identities. The new identities had no reviews associated with them, and from the perspective of consumers, they were distinct from the seller's established identity. The seller randomly assigned items to be sold under the various identities while making sure to offer the same quality of service to all customers. Consumers' willingness to pay was 8% higher for purchasing from the established seller even though the items sold across the various identities were similar. Pallais (2014) studies the value of reputation on oDesk, an online marketplace where employers can hire workers to perform various tasks. The oDesk experiment involved hiring approximately 1,000 workers to perform the same data-entry task. Workers were randomly assigned to the treatment or control group. Treated workers received detailed and

[9] Sales rank is negatively related to sales, and hence a decrease in rank (i.e., a higher rank) is associated with a larger sales volume.

objective feedback on their performance, whereas workers in the control group received equally objective but less detailed feedback. Treated workers had better future employment outcomes, such as higher wages and earnings, compared to workers in the control group who performed as well on the data-entry task but received coarser feedback.

Overall, the evidence in the literature points to economically significant effects of establishing a reputation: consumers tend to trust products (or sellers) with few or no reviews less than well-reviewed products. Therefore, additional reviews can improve sales even if they do not affect a product's average rating.

3.3 Review content

Review text is inherently high dimensional and typically not amenable to analysis by the same methods researchers use to study review valence and volume. Therefore, most analyses of review text rely on a pre-processing step that transforms text into a small number of variables that capture variation along dimensions of interest. These variables can subsequently be analyzed using standard econometric approaches.

At a high level, the literature has employed two approaches to analyze specific aspects of review content. First, researchers have used statistical approaches that rely on dimensionality-reduction algorithms to map text onto low-dimensional measures. Statistical approaches vary in their sophistication from the application of simple formulas to compute metrics, such as readability, to complex methods, such as topic modeling (Blei, Ng, and Jordan, 2003), that can extract latent structures and meaning from unstructured data. Another common approach to reduce the dimensionality of text is the use of pre-built and validated dictionaries, such as LIWC (Tausczik and Pennebaker, 2010), that measure the incidence of psychometric attributes such as affect and emotion in text. Moe, Netzer, and Schweidel (2017) provide a comprehensive review of the statistical approaches used to analyze text from reputation systems and social media.

These approaches allow the researcher to characterize existing text along different dimensions, and these content characteristics can then be correlated with sales. The obstacles to obtaining causal estimates described earlier also apply to this setting, and any analysis of existing text does not allow the researcher to conclude which characteristics of review content have a positive *causal* impact on demand.

Several papers have used statistical methods to uncover interesting patterns in review text. Ghose and Ipeirotis (2011) study the text of travel reviews and find that objectivity, readability, and lack of spelling errors are correlated with higher product sales. Archak, Ghose, and Ipeirotis (2011) extract product features from review text, and use these features to predict future product sales. Ludwig, De Ruyter, Friedman, Brüggen, Wetzels, and Pfann (2013) show that positive and negative affective content (e.g., words such as "love" and "hate," which are detected using LIWC) are positively associated with conversion rates.

The second approach is grounded in consumer behavior theory and starts by forming a hypothesis regarding the effects of specific text constructs. These constructs are typically detected manually (e.g,, using human coders) or with simple pattern-matching rules. Sometimes, researchers combine this technique with machine learning: after manually labeling a small set of reviews, a classifier is trained to predict the presence of specific text constructs in larger unlabeled review corpora. An advantage of using theory to form hypotheses is that one can design experiments to test these hypotheses, leading to internally valid estimates. Packard and Berger (2017) and Kupor, Giblin, and Morewedge (2017) offer examples of this approach.

Packard and Berger (2017) combine lab experiments and field data to investigate how the language consumers use to endorse products affects how persuasive their reviews are. The authors find explicit endorsements ("I recommended this book") are more persuasive than implicit endorsements ("I enjoyed this book"). Kupor, Giblin, and Morewedge (2017) also use lab experiments to investigate the effects of endorsement authenticity. The study finds that endorsements that are perceived to be more authentic are more convincing than endorsements that are perceived to be more thoughtful. For example, product endorsements that contain typographical errors are perceived as more authentic, thus enhancing their persuasiveness. Interestingly, these results contradict the results in Ghose, Ipeirotis, and Li (2012), who find that reviews with spelling errors are associated with lower demand.

Both approaches for studying the effects of review text are valuable. Statistical approaches can uncover patterns in text whose importance might not be evident a priori. At the same time, statistical approaches offer no direct connection to theory, and linking patterns uncovered by statistical methods to theoretical constructs ex post can be difficult. Consumer behavior theory can help resolve these problems by guiding research in the high-dimensional space of language patterns, and informing the design of experiments to estimate causal treatment effects.

3.4 Moderators of the relationship between reviews and demand

In addition to studying the impact of reviews on demand, a number of papers have also examined moderators of the relationship between reviews and demand. A consistent theme that has emerged is that the impact of reviews depends on consumers' prior information. Reviews are not the only information source available to consumers, and thus they affect demand only to the extent that they provide new information. Below, we discuss how other sources of information available to consumers can moderate the impact of reviews on demand.

A brand is a traditional marketing mechanism that can be used to convey useful information about quality to consumers. The evidence in the literature suggests that even in the presence of reviews, consumers continue to rely on brands to learn about product quality. For instance, Luca (2016) finds that Yelp ratings do not impact the revenue of chain restaurants, because less uncertainty exists about their quality than about the quality of independent restaurants. Similarly, Lewis and Zervas (2016) show the impact of TripAdvisor ratings is much smaller for chain-affiliated hotels than for independently operated properties. At the same time, recent evidence suggests the impact of brands is declining due to the increased adoption of consumer reviews Hollenbeck (2016). Overall, the literature suggests that although consumers continue to rely on brands to learn about quality, review platforms can act as a substitute source of information.

The amount of information available about a product typically increases over its life cycle. Therefore, the timing of reviews with respect to the product life cycle can moderate the impact of reviews on demand. Babić, Sotgiu, De Valck, and Bijmolt (2016) conduct a meta analysis of existing studies to show that early reviews are more important than recent reviews in driving the sales of new products. Because product quality remains relatively stable, early reviews can resolve quality uncertainty, and subsequent reviews provide little additional information to consumers. A notable exception are services, whose quality can vary significantly over time, and hence recent reviews that provide up-to-date information about service quality are the main driver of demand (Babić, Sotgiu, De Valck, and Bijmolt (2016)).

In addition to providing information about quality, review platforms can help consumers discover new products and services they were not aware of. Berger, Sorensen, and Rasmussen (2010) show that even negative reviews can have a positive impact on demand when they cause a sufficient increase in consumer awareness. For instance, compared to not being reviewed at all, a negative book review may increase the sales of a lesser-known author by driving awareness. By contrast, a negative review for a well-known author has a negative impact on sales.

Finally, consumers' adoption of review platforms moderates the impact of reviews on demand. As more consumers turn to review platforms to learn about products and services, and as review platforms accumulate more reviews, the influence of reviews on demand will also likely become stronger. Looking at the impact of TripAdvisor on hotel demand over time, Lewis and Zervas (2016) find the impact of a one-star increase in a hotel's average rating increased from zero to 25% between 2004 and 2014.

4 Online Conversations

We now turn to online conversations. This type of WOM involves consumers discussing topics on social media platforms such as Twitter or Facebook. We lay out what is known about the impact of online conversations on demand, following a similar structure as in the previous section on reviews. In particular, we discuss the impact of the volume as well as the content of online conversations. Note that "valence," which was discussed in the context of reviews, is not a characteristic that can describe online conversations, due to the absence of an aggregation mechanism such as the average ratings score. We then turn to a discussion of the moderators of WOM effects in the realm of online conversations.

We also note the academic literature on online conversations is small, at least in terms of studies that use research designs aimed at uncovering causal effects. The discussion below is therefore shorter than the discussion pertaining to reviews.

4.1 Volume of online conversations

Can the volume of online conversations affect sales? Based on the extant research, the answer is a qualified yes. However, this assessment of effect size is based on only two studies with fairly distinct research designs.

Seiler, Yao, and Wang (2017) study the effect of customer microblogging on TV viewership. Due to a political scandal, the Chinese government temporarily shut down the microblogging platform Sina Weibo in 2012, which was the most important and popular social media website in China at the time. Relying on this exogenous shock, the authors show the reduction in conversation volume did lead to lower consumption of TV shows. Accordingly, a higher volume of online conversations may increase demand. However, the authors find that WOM only has a moderate effect in this setting. Increasing the volume of online conversations by 10% leads to an increase in demand of only 0.16%.

In a related paper, Gong, Zhang, Zhao, and Jiang (2017) study the effect of microblogging on TV viewership. The authors execute a field experiment in China that is primarily focused on understanding the impact of a firm's tweets about its own TV shows. However, the paper also investigates one specific form of WOM, by having influential users retweet the companies' tweets for a random subset of TV shows.[10] They find that an influential user's retweet boosts viewership (demand) by 33%. This effect is large in magnitude and (although not directly comparable) is contrary to the modest effect size found in Seiler, Yao, and Wang (2017). However, the nature of the WOM is very different. A retweet by an influential user will likely be more impactful than a tweet by the average consumer. We return to the influence of the sender's identity on WOM effectiveness when discussing moderators of conversational WOM below.

4.2 Content of conversations

Similar to online reviews, the content of online conversation is also high dimensional, and a preprocessing step is required to convert text into a smaller set of variables of interest (e.g., Liu, Singh, and Srinivasan (2016)). The two dimensions that studies have focused on so far are (1) positive or negative sentiment that is expressed in online conversations and (2) whether content is informative in the sense that it contains factual information about a product. This focus is in contrast to the literature on reviews discussed earlier, where a broader set of content-related constructs has been analyzed.

To analyze the impact of sentiment, Seiler, Yao, and Wang (2017) use human coders to assign comments on Sina Weibo to either positive or negative sentiment (or neither). They find that shows with higher levels of either positive or negative comments (or both) experienced a larger decrease in viewership than other shows when the platform was disabled. Although the study does not experimentally manipulate content, these patterns across different shows suggest conversations

[10] In the United States, hiring influential users to promote a firm's products does not conform with FTC's truth-in-advertising standards (to conform with the standard, the retweet would have to indicate the user was paid for the retweet).

with a higher amount of sentiment have a greater impact on demand. Interestingly, even negative conversations seem to have a positive effect on demand. One possible explanation is that negative sentiment can occur as part of an engaged discussion pertaining to a particular TV show. We note that other studies (e.g., Sonnier, McAlister, and Rutz (2011)) find that negative content leads to lower demand.

Another potential channel by which conversations can affect demand is by providing information about the product and hence affecting consumers' consumption decisions. For example, Gong, Zhang, Zhao, and Jiang (2017) analyze the content of influential users' retweets. They find that when the retweets contain information about the TV shows, such as the show time and channel, the impact on increasing viewership is larger than from non-informative retweets. By contrast, Seiler, Yao, and Wang (2017) find no significant effect of informative content on the show's viewership.

In summary, the small amount of evidence on the content of conversations is very mixed, and no study has used a research design that can obtain causal effects. Hence, although the content of WOM is an interesting area for future research, no general findings have been established with regards to which content is more effective in increasing demand.

4.3 Moderators of the impact of online conversations

To organize the discussion of moderators in the context on online conversations, we first present a simple conceptual framework for the possible channels through which conversations can influence demand. We consider three broad channels, namely, informative effects, persuasion, and complementarity between product consumption and online conversations.[11] Informative effects can arise when online conversations among customers generate or disseminate product/brand information that resolves uncertainty about the product, enhances the awareness, or reminds customers about the product. Persuasive effects may enhance customers' preference toward the product without delivering information. Lastly, complementarity between online conversations and product consumption may occur when customers derive a higher utility from the product if they can participate in online conversations with others who share the same interest in the product. We note that in the context of reviews, informative effects are likely to be the most relevant channel, and we hence focused on effects relating to information provision and uncertainty when discussing moderators of reviews earlier. In the case of online conversations, all three channels are potentially relevant.

With regards to informative effects, one potential moderator of a conversation's impact on demand is consumers' uncertainty about the product. Product uncertainty has been found to be an important moderator in the context of reviews and could conceivably also play a role in the case of conversations. Moretti (2011) provides evidence to this effect. He analyzes how customers obtain information about the quality of movies from peers, and finds the impact of WOM is largest for movies about which consumers had little prior knowledge.[12]

[11] This taxonomy is based on Bagwell (2007).

[12] In this paper, no direct data on conversations are available, and conversations in principle can happen both online and offline.

A second possible moderator is the trustworthiness of participants in the conversation, especially if the conversation is persuasive in nature. The literature on network effects has well documented that specific individuals ("influential consumers") have a disproportionate impact on the behavior of other users in the network (e.g., Katz and Lazarsfeld (1955), Nair, Manchanda, and Bhatia (2010), and Iyengar, den Bulte, and Valente (2011)). We conjecture such patterns may also apply to the context of online conversations. One example for such an effect is Gong, Zhang, Zhao, and Jiang (2017), who show demand increased strongly when influential users retweeted a firm's tweets.

A final moderator is the timing of the conversations. Both the exposure to conversations prior to a purchase as well as the ability to engage in conversations after the purchase can in principle affect demand. Seiler, Yao, and Wang (2017) study the impact of conversations before or after consumption in the context of TV shows, and find the anticipation of the post-consumption conversations has a positive impact on demand. By contrast, conversations prior to the show airing do not impact show viewership. The authors interpret these findings as evidence of complementarity between (post-show) conversations and product consumption, because informative or persuasive effects would both entail an impact of pre-show conversations on demand.

5 Summary

In this chapter, we outlined the state of the academic literature with regards to the impact of WOM on demand. Measuring impact is difficult in the context of WOM because WOM is likely to react to changes in demand and cannot be directly manipulated by firms. Therefore, a set of alternative measurement methods is required. We described three such methods that are promising for WOM measurement: natural experiments, A/B tests that serve as instruments for WOM, and regression discontinuity designs. We then reviewed the literature that analyzes the impact of WOM, based on these three (as well as related) methods.

With regards to consumer reviews, the evidence points to review valence and volume having a substantial economic impact, though the magnitudes of these effects vary depending on context. The evidence regarding content is more preliminary and only partially causal. Across many studies, results are consistent with the idea that prior information about a product determines how impactful reviews are in terms of increasing demand. Hence, firms in high-uncertainty environments will tend to benefit more from positive customer reviews. In the case of online conversations, the amount of academic studies is much smaller, and findings should be considered preliminary. Some evidence suggests conversation can positively impact demand, but the magnitude of the effect differs greatly across studies. Similar to the context of reviews, evidence shows that greater uncertainty enhances the impact of conversations on demand. With regards to the content of conversations on online platforms, findings are mixed and do not yet offer any generalizable results.

References

Anderson, M., and J. Magruder (2012): “Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database,” *Economic Journal*, 122(563), 957–989.

Archak, N., A. Ghose, and P. G. Ipeirotis (2011): “Deriving the pricing power of product features by mining consumer reviews,” *Management science*, 57(8), 1485–1509.

Ba, S., and P. A. Pavlou (2002): “Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior,” *MIS quarterly*, pp. 243–268.

Babić, R. A., F. Sotgiu, K. De Valck, and T. H. Bijmolt (2016): “The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors,” *Journal of Marketing Research*, 53(3), 297–318.

Bagwell, K. (2007): “The Economic Analysis of Advertising,” in *Handbook of Industrial Organization*, ed. by M. Armstrong, and R. Porter, vol. 3, pp. 1701–1844. Elsevier Science.

Bajari, P., and A. Hortacsu (2003): “The winner’s curse, reserve prices, and endogenous entry: Empirical insights from eBay auctions,” *RAND Journal of Economics*, pp. 329–355.

Berger, J., A. T. Sorensen, and S. J. Rasmussen (2010): “Positive effects of negative publicity: When negative reviews increase sales,” *Marketing Science*, 29(5), 815–827.

Blake, T., C. Nosko, and S. Tadelis (2015): “Consumer Heterogeneity and Paid Search Effectiveness: A Large Scale Field Experiment,” *Econometrica*, 83(1), 155–174.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003): “Latent dirichlet allocation,” *Journal of machine Learning research*, 3(Jan), 993–1022.

Cabral, L., and A. Hortacsu (2010): “The dynamics of seller reputation: Evidence from eBay,” *The Journal of Industrial Economics*, 58(1), 54–78.

Chevalier, J. A., and D. Mayzlin (2006): “The Effect of Word of Mouth on Sales: Online Book Reviews,” *Journal of Marketing Research*, 43(3), 345–354.

Dewally, M., and L. Ederington (2006): “Reputation, certification, warranties, and information as remedies for seller-buyer information asymmetries: Lessons from the online comic book market,” *The Journal of Business*, 79(2), 693–729.

Eaton, D. (2002): “Value Information: Evidence from Guitar Auctions on eBay,” .

Eckles, D., R. F. Kizilcec, and E. Bashky (2016): “Estimating peer effects in networks with peer Encouragement designs,” *PNAS*.

Ghose, A., and P. G. Ipeirotis (2011): “Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics,” *IEEE Transactions on Knowledge and Data Engineering*, 23(10), 1498–1512.

Ghose, A., P. G. Ipeirotis, and B. Li (2012): “Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content,” *Marketing Science*, 31(3), 493–520.

Gong, S., J. Zhang, P. Zhao, and X. Jiang (2017): "Tweeting as a Marketing Tool - Field Experiment in the TV Industry," *The Journal of Marketing Research*, p. forthcoming.

Gordon, B., F. Zettelmeyer, N. Bhargava, and D. Chapsky (2016): "A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook," *Working Paper.*

Gordon, B. R., and W. R. Hartmann (2013): "Advertising Effects in Presidential Elections," *Marketing Science*, 32(1), pp. 19–35.

Hollenbeck, B. (2016): "Online Reputation Mechanisms and the Decreasing Value of Brands," .

Houser, D., and J. Wooders (2006): "Reputation in auctions: Theory, and evidence from eBay," *Journal of Economics & Management Strategy*, 15(2), 353–369.

Iyengar, R., C. V. den Bulte, and T. W. Valente (2011): "Opinion Leadership and Social Contagion in New Product Diffusion," *Marketing Science*, 30(2), 195–212.

Jin, G. Z., and A. Kato (2006): "Price, quality, and reputation: Evidence from an online field experiment," *The RAND Journal of Economics*, 37(4), 983–1005.

Kalyanam, K., and S. H. McIntyre (2001): "Return on reputation in online auction markets," .

Katz, E., and P. F. Lazarsfeld (1955): *Personal Influence*. Free Press, New York, NY.

Kupor, D. M., C. E. Giblin, and C. K. Morewedge (2017): "Spontaneous Influence: Persuasion through Perceived Authenticity," Discussion paper.

Lewis, G., and G. Zervas (2016): "The Welfare Impact of Consumer Reviews: A Case Study of the Hotel Industry," *Working Paper.*

Lewis, R. A., and D. H. Reiley (2014): "Online ads and offline sales: measuring the effect of retail advertising via a controlled experiment on Yahoo!," *Quantitative Marketing and Economics*, 12(3), 235–266.

Liu, X., P. V. Singh, and K. Srinivasan (2016): "A Structured Analysis of Unstructured Big Data by Leveraging Cloud Computing," *Marketing Science*, 35(3), 363–388.

Luca, M. (2016): "Reviews, Reputation, and Revenue: The Case of Yelp.com," *Working Paper.*

Luca, M., and S. Vats (2013): "Digitizing Doctor Demand: The Impact of Online Reviews on Doctor Choice," *Cambridge, MA: Harvard Business School.*

Lucking-Reiley, D., D. Bryan, N. Prasad, and D. Reeves (2007): "Pennies from eBay: The determinants of price in online auctions," *The journal of industrial economics*, 55(2), 223–233.

Ludwig, S., K. De Ruyter, M. Friedman, E. C. Brüggen, M. Wetzels, and G. Pfann (2013): "More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates," *Journal of Marketing*, 77(1), 87–103.

McDonald, C. G., and V. C. Slawson (2002): "Reputation in an internet auction market," *Economic Inquiry*, 40(4), 633–650.

Moe, W. W., O. Netzer, and D. A. Schweidel (2017): "Social Media Analytics," in *Handbook of Marketing Decision Models*, pp. 483–504. Springer.

Moretti, E. (2011): "Social Learning and Peer Effects in Consumption: Evidence from Movie Sales," *The Review of Economic Studies*, 78(1), 356–393.

Nair, H. S., P. Manchanda, and T. Bhatia (2010): "Asymmetric Social Interactions in Physician Prescription Behavior: The Role of Opinion Leaders," *Journal of Marketing Research*, 47(5), 883–895.

Packard, G., and J. Berger (2017): "How language shapes word of mouthÕs impact," *Journal of Marketing Research*, 54(4), 572–588.

Pallais, A. (2014): "Inefficient hiring in entry-level labor markets," *The American Economic Review*, 104(11), 3565–3599.

Proserpio, D. M., W. Xu, and G. Zervas (2016): "You Get What You Give: Theory and Evidence of Reciprocity in the Sharing Economy," *Working Paper*.

Resnick, P., R. Zeckhauser, J. Swanson, and K. Lockwood (2006): "The value of reputation on eBay: A controlled experiment," *Experimental economics*, 9(2), 79–101.

Rossi, P. E. (2014): "Invited Paper – Even the Rich Can Make Themselves Poor: A Critical Examination of IV Methods in Marketing Applications," *Marketing Science*, 33(5), 655–672.

Sahni, N. (2015): "Effect of Temporal Spacing between Advertising Exposures: Evidence from an Online Field Experiment," *Quantitative Marketing and Economics*, 13(3), pp. 203–247.

Seiler, S., and S. Yao (2017): "The impact of advertising along the conversion funnel," *Quantitative Marketing and Economics*, 15(3), 241–278.

Seiler, S., S. Yao, and W. Wang (2017): "Does Online Word of Mouth Increase Demand? (And How?) Evidence from a Natural Experiment," *Marketing Science (forthcoming)*.

Shaoolian, G. (2017): "5 Digital Marketing Tips To Increase Your Brand's Growth Online And Improve Ad Results," *Forbes*.

Shapiro, B. T. (2016): "Positive Spillovers and Free Riding in Advertising of Prescription Pharmaceuticals: The Case of Antidepressants," *The Journal of Political Economy*, p. forthcoming.

Sonnier, G. P., L. McAlister, and O. J. Rutz (2011): "A Dynamic Model of the Effect of Online Communications on Firm Sales," *Marketing Science*, 30(4), 702–716.

Sun, M. (2012): "How Does the Variance of Product Ratings Matter?," *Management Science*, 58(4), 696–707.

Tausczik, Y. R., and J. W. Pennebaker (2010): "The psychological meaning of words: LIWC and computerized text analysis methods," *Journal of language and social psychology*, 29(1), 24–54.