BrandLoyalties Basic Concepts

By Richard Davis
BrandLoyalties, Inc.

In highly competitive marketplaces a primary predictor of the future success of corporations is the satisfaction and loyalty of their customer base. A highly loyal customer base can increase corporate margins by reducing marketing costs and enabling premium pricing. The loss of customer loyalty, on the other hand, can lead to a loss of market share and shrinking margins as corporations respond by cutting prices and/or increasing marketing efforts. The BrandLoyalties.com website offers purely quantitative metrics for the loyalty of on-line customers to the various brand names of a large number of widely traded equities.

In a highly competitive (non-monopolistic) marketplace a corporation is best able to maintain or expand its net margins when it has a loyal customer base. For example, when that customer base is exceptionally loyal a corporation may even be able to maintain revenue levels without expending any significant resources on the expensive marketing programs or high cost advertising campaigns that can erode net margins.

Furthermore, a highly loyal customer base may enable a corporation to price their products at a premium relative to their competitors. If the exceptionally high customer loyalty has been acquired and sustained without commensurate increases in the cost of the goods sold, the result of the premium pricing is increased operating margins for the corporation.

Examples of corporations with exceptionally loyal customer bases would include sports franchises or entertainers that can sell out venues by simply publishing their event schedules. In these cases margins can remain relatively high simply because there is no need for significant media buys. In other cases (e.g., Apple, Inc.) a fiercely loyal customer base can enable historically premium pricing even in the face of highly competitive feature sets.

Conversely, rapidly declining customer loyalty will likely lead to lost market share and plunging operating margins as prices erode and marketing expenses are ramped up. Contracting revenues and shrinking margins are never a good sign for any corporation, and early indications of such patterns should be welcomed by any investor.

Consumer “brand loyalty” was one of the first intangible assets recognized in academic literature. This asset is of key interest to investors because of the value that "brand loyalty" generates to companies in terms of:

– a substantial entry barrier to competitors;

– an increase in the firm's ability to respond to competitive threats;

– greater sales and revenue; and

– a customer base less sensitive to the marketing efforts of competitors.

Jack Treynor (2002) studied the concept in more blatantly economic terms, arguing that "Brand loyalty manifests itself in consumers' willingness to pay a higher price for the brand they prefer."

Obviously, customer brand loyalty is not the only necessary (or sufficient) criteria for successful corporations. Corporations can fail in spite of high customer loyalty for any number of reasons, including non-competitive cost structures, balance sheet challenges, and systemic market collapses.

Additionally there is a flip side to the non-customer-loyalty related characteristics of corporations: those characteristics can help them maintain operating margins and mask eroding market share – even as their traditional customers flee them. Among those corporate characteristics are aggressive cost reduction opportunities, rapidly expanding alternative (e.g., global) markets, and access to financing at extremely low rates.

And over longer time spans customer loyalty can deteriorate even within a fanatical customer base if the products and services of the corporation fail to evolve in a competitive manner – as customers begrudgingly move to new paradigms when the old loyalties become obsolete.

Brand Loyalties as Investment Tool

One person to recognize the value of brand loyalties as a tool that could be used during the selection process for equity portfolios was Peter Lynch. He managed the Magellan Fund at Fidelity Investments during 1977-1990, when the average annual return experienced by Magellan’s shareholders was over 29%.

One of Peter's key insights involved using “the power of common knowledge” to select consumer equities that were worthy of further rigorous research and fundamental analysis. He selected only those companies that he knew were selling products again and again to his family, colleagues and acquaintances.

At BrandLoyalties, Inc. we took a look at his results and asked several basic questions: What has changed since 1977? How can we harness evolving markets and 21st century technologies to improve the scope, timeliness and quantitative accuracy of Peter's methodologies?

In an effort to bring Peter’s investment strategy into the 21st Century we have developed consumer behavior metrics that data-mine “Big Data” to measure the fondness of on-line and social media consumers for certain brand names – and (more critically) how that fondness (or consumer loyalty) shifts over time. Those shifts in the feelings that consumers have for the brands that we track are then translated into the "brand loyalties" of on-line consumers for the products of well over 2,200 publicly traded US corporations.

Our "brand loyalties" equity metrics track actual on-line consumer behavior, and for that reason they are purely quantitative. We are using state-of-the-art internet data involving massive "Big Data" data-mining – our collection processes capture over 100,000,000 on-line consumer brand choices daily. This methodology has vastly greater sample sizes than conventional questionnaire surveys or boutique focus groups. The resulting metrics are also orders of magnitude more current than other sources, with daily updates posted on the subscriber portion of our website every morning.

We have vastly expanded Peter Lynch's "friends-and-neighbors informal-focus-group" concept by over 6 orders of magnitude by capturing data from nearly everyone shopping on the web, and doing the sampling in "real-time." As a consequence we have a twenty-first century version of his "brand loyalties" metric that has:

– The ultimate granularity, with results that are actionable at the equity-by-equity level.

– Extensive scope and coverage, with over 3,200 US, 2,150 Asian and 1,300 European equities tracked (a number that is growing daily).

– Daily updates to each equity's "brand loyalty," reflecting a mix of year-over-year and trailing 90 day shifts in customer affections.

-- Signals containing a material and consistent alpha that has a horizon (or latency) of 14 to 90 days.

How Do We Do It?

It all starts with people connected to the internet. Nearly everything people do on the web gets captured in one form or another. For example, advertisers utilize data captured by search engines to display the annoying (but highly targeted) ads that are constantly showing up in your browser when you surf the web. And at other times people consciously put data on the internet – in social media postings, when tweeting, while writing product reviews or recommendations at on-line forums, or when using auction sites.

In fact, our data comes from any and all publicly facing portions of the web. As part of our process, our web-crawling bots start in "link-rich" web environments and crawl throughout a specified time window each night. Critically, we do not access data behind firewalls or that require passwords or credentials of any sort.

As a consequence of this simple process, our bots collect data from all of the data sources commonly found on the public web, including online product forums, crowd sourced reviews, blogs and any social media content that has been re-posted to the web.

We then derive our published metrics by building vocabulary lists (i.e., "dictionaries" or lexicographies) of words encountered online, by deconstructing or disassembling publicly available internet material into its component words. We then accumulate, sort and count the frequency of the occurrences of those words.

The resulting summary “dictionaries” (and occurrence counts) of words used on the internet during a given time span are then scanned for registered brand and trade names owned by major corporations. The list of targeted brand and trade names is dynamic, and it can contain tens of thousands of entries.

Examples would include Jetta, Camry and Corvette, all of which are clearly automotive models. Other brand names are more ambiguous – their words can also be part of the common lexicon. Examples of automotive brand names with ambiguous meanings include Malibu, Fit and Ram. In these cases it is necessary to expand the “brand name” to less ambiguous short phrases such as “Honda Fit.”

Care is taken to separate brand names from the investor focused names of the corporations themselves. For example, “Apple” is first of all ambiguous as a word, and when specifically intended as a corporate identity it is used in both consumer commerce and investment contexts. For that reason “iPhone” or “iPad” would be better brand names for AAPL (Apple Inc.).

Each time a brand name is mentioned in the “Big Data” cloud, it constitutes a “citation” of a brand owned by the respective corporation. The number of such citations found over any given time span is considered to be that corporation’s “citation rate.” From those rates the critical year-over-year “citation growth rate” and other metrics can be developed.

Corporate citation rates generally depend on both the number of customers it has and the frequency of its interactions with those customers. Some corporations (e.g., AMGN or LMT) simply do not interact directly with consumers at all, and have very low citation rates. Other corporations (e.g., MCD or JACK) have large scale consumer operations and transact relatively frequently with those customers – creating much higher citation rates as a by-product of those interactions.

Corporations are often classified and grouped by both the industry in which they are engaged and the scale of their market capitalization. It is important to understand how such classifications impact the availability and usefulness of the brand name citations that can be found in “Big Data.”

When comparing the number of times that consumers interact with MCD and AMGN, it is obvious that their respective industries are a critical factor in the vast differences in their citation rates. Similarly, when comparing citation rates for MCD and JACK, it is clear that the scale of consumer operations (which can be crudely proxied by market capitalization) is likely the major factor in MCD’s substantially higher citation rates.

Developing Metrics

We then measure the relative presence of a company's products or services within the entire collection of data from all sources, comparing the relative share of brand name citations with similar year-ago data and calculating the year-over-year growth (or contraction) in the share during each of the trailing 90 days. Those data points are then used to create regression slopes that are used to rank each equity relative to all others in the "universe" of equities that we track. The final result is a ranking based on percentiles, with lower percentiles representing companies with increasing popularity among on-line consumers.

Those relative rankings are expressed as "top X%" percentiles -- with lower numbers representing the better loyalties (e.g., the best 10% = 10th percentile and lower) and higher numbers the poorer loyalties (e.g. the worst 20% = 80th percentile and higher). Thus at any time the best half of the "full universe" of our tracked equities would be equities with percentile numbers of 50% or lower.

And when constructing a portfolio, a portfolio manager could include only those equities with current percentile rankings at or below any given threshold. For example, a threshold of 20% would include on the best fifth of all ranked equities, while a threshold of 10% would similarly only include the highest ranked 10% of the more than 2,200 US equities we track.

It is interesting to note that even our "full universe" of tracked equities substantially out-performs both the S&P 500 and the S&P Consumer Discretionary Indexes – primarily because our "full universe" generally includes equities with a significant presence on the web (an area of significant growth at the present time) or products/brands with a lot of current "buzz" on the web (usually meaning products and brands in high demand).

The equities that offer sufficient web luminosity to provide statistically rigorous signal to noise ratios are also generally companies that employ the forward looking consumer marketing strategies and/or transaction platforms – making them somewhat better suited for commerce in the 21st century than many of their peers.

Thus our "full universe" of over 3,200 US equities is actually highly selective by virtue of our sampling methodology (or sampling "bias" if you prefer), making it a form of positive alpha equity metric in its own right.

And for this reason even the poorest performing equities in our "full universe" may not be good candidates for shorting. Equity lists extracted exclusively from our "full universe" can present challenges to asset managers that need to construct market or sector neutral hedges – since there might not be many members of our "full universe" that are suitable for the short basket at any given time. Successful hedging strategies generally need to look beyond our "full universe" for the short side of the portfolio.

Source Data Scale

The scale of "Big Data" is a moving target in several ways. The amount of data being created has increased exponentially over the past two decades. And the primary data creation has migrated from desktop browsers to mobile "apps." That growth has been accompanied by a diffusion of citation purpose: citations were once focused tightly on searches conducted immediately prior to consumer purchases; citations now have broader purposes when embedded in social exchanges, reviews and commentary.

The demographics of the people creating the bulk of the data have also changed. The sources are now more egalitarian. Access to online resources has become nearly universal – and socially obligatory. Two decades ago the online "user" profile was biased towards well-educated, middle-class consumers in the developed world – often at the peak of their spending power. That profile has now shifted dramatically. "Users" now include texting, media watching and social media connected young people living anywhere on the globe – with most of them many years away from their peak earning powers. The data sources may still be biased, but towards a different demographic.

Meanwhile, the scale of data growth is staggering. "Big Data" content growth has been exponential: it has been growing at an average of ~60% per year. And it is now estimated to be about 8 zettabytes (10^{21} bytes) – roughly 2.5 terabytes for each and every global online user, or enough data to fill $1 trillion worth of hard drives.

For anyone processing the data there is a secondary scale problem – one associated with the scale of the resources deployed to sample the "Big Data." Not even the resources purchased by the NSA's "black budget" can collect and record zettabytes of content. Every real-world sample size is ultimately limited by the bandwidth, storage and processing resources available – meaning that even the most dedicated web "crawling" or "scraping" can effectively examine only about a billionth of the 8 zettabytes per day.

However, the application of more resources or more bandwidth inevitably creates larger sample sizes – making contemporary citation rate samples significantly greater than the historical rates gathered with more constrained resources.

All of these scale changes simply mean that the business intelligence value of each brand name "citation" is being diluted, and absolute citation rates sampled from today's internet using today's resources and bandwidths cannot be meaningfully compared to those recorded in 2006.

In fact, raw citation rates are generally meaningless without the context of the highly variable sample size – requiring comprehensive and proprietary normalization based on our internal bandwidth and latency statistics before any analytics can be performed.

Our bandwidth and latency normalizations are based in part on carefully instrumenting our collection processes and on proprietary "standard candle" approaches. In astrophysics, "standard candle" designates a star which has a known and constant absolute luminosity. We use the term "standard candle" to refer to non-brand related words that should have on-line citation rates (i.e., luminosities) which are largely invariant over time. We have built a number of large "baskets" of these "standard candles" that, as a group, should return roughly the same aggregate citation counts on a daily basis.

Significant variations from the historical norms for the citation rates of these standard candles (especially if widely observed within the group even as intra-group share ratios are preserved) can provide quantitative normalization factors to compensate for the daily bandwidth and latency fluctuations that we experience, both globally and regionally. To assist with compensating for inadvertent sampling biases, the words are deliberately chosen to cover a wide range of luminosities. A few examples of standard candles from the North America basket would include the words cancer, Bruckner, diabetes and Euripides.

Additionally, new sources of brand name citations that are being created every day. Furthermore, as more scanning resources are brought online, older citations are constantly being discovered within the 8 zettabyte maze. This previously un-harvested back history makes even “historic” citation rates a moving target.

One way to deal with both the vast differences in sample size and the staggering scale of the data is to focus on the proportional composition of the samples. Our key metrics normalize for sample size noise by converting the raw citation rates to “citation share” – a term analogous to the use of “market share” within the investment community. In theory, citation share and citation share growth rates should be largely invariant over simple resource driven scale changes.

But this approach introduces yet another moving target, since citation share is itself a function of the scale or depth of the chosen corporate coverage. Expanding the brand name targets from those owned by the S&P 500 to those owned by the Wilshire 5000 dilutes the citation share metrics materially. To address this additional scale issue, we have chosen to use citation share growth ranking percentiles. Our historical simulations indicate that the significance of being in the top decile of covered corporations at any given time is generally invariant to simple resource driven sample size growth and industry/capitalization neutral changes to the depth of corporate coverage.

Social Media and Sentiment

The advent of social media has made sentiment measurements an area of active research. Useful “Big Data” can be generated by a wide variety of different sources, ranging from conventional search engines (e.g., Yahoo!, Bing and Google), social media (e.g., Twitter, Instagram and Facebook), and crowd sourcing platforms (e.g., Yelp and TripAdvisor).

We have found that the “business intelligence” available to investors from each of these sources depends to a large extent on the frequency with which that investor turns his or her portfolio. For high frequency or day trading investors (that typically hold equity positions for time spans measured in seconds to hours), the value of a breaking news “tweet” that front-runs the wire services by minutes can be significant.

However, for investors that hold equity positions for time spans measured in weeks or months, the intelligence available from conventional consumer search activities can be shown quantitatively to have substantially greater alpha potential. The reasons for this are three-fold:

– First, parsing “mood” from internet postings is, at best, technically challenging. Natural Language Processing (NLP) includes double or triple negatives and implied sarcasm that can play havoc with sentiment parsing algorithms – which, frankly, have trouble correctly deciphering snarky comments.

Additionally, social media is replete with especially cryptic forms of language and emoticons that are not standardized from platform to platform (e.g. how would a sentiment parsing algorithm deal with the following product commentary: “Although everyone thinks that Product-X stinks, I remain a true believer”?).

Ultimately each portfolio manager has to ask themselves whether they are willing to bet the portfolio on the accuracy of someone else’s sentiment parsing algorithm.

– Secondly, a conventional search is generally in closer proximity to a revenue generating transaction than a similar “like” in social media space. A search for store hours or product prices has a higher probability of a closely following purchase than a brand “like” in social media.

– Thirdly, social media demographics are strongly biased towards internet users still years away from their peak earnings and spending power.

The demographic and parsing issues alone explain the difficulties encountered when using social media to predict future human behavior – particularly election outcomes. Such predictions have, in fact, been nothing short of disastrous.

Portfolio simulations have shown that simple search derived citation share growth metrics materially outperform social media derived “sentiment” metrics in quantitatively managed portfolios.

Events

But even simple citation rate based metrics can be significantly impacted by “public relations” events or nightmares. These nightmares can be created by especially visible or spectacular product incidents (e.g., Bank of America’s 2011 debit card fee fiasco), service issues (Blackberry/RIM 2012 outages) , safety recalls (GM 2014), corporate fiscal implosions or massive layoffs (RadioShack 2014), liquidation sales (Borders Group 2011), mergers and acquisitions (Men’s Wearhouse / Joseph A Banks, 2014), or human resources litigation (Walmart 2011). Even margin-depressing (or desperation/suicidal) sales can skew citation rates over short terms. Such events generate citations for reasons unrelated to ongoing or sustainable consumer commerce.

We have found that over the past decade such events tend to be transient and provide statistically identifiable increases in citation rates. For each of the examples mentioned above we observed at least a 4-sigma deviation from baseline citation rates over a short-term time span (optimally 7 calendar days, which compensates for weekly citation cycles). We actually report the “event risk” for each equity as a peer-relative percentile ranking, with 4-sigma events generally occurring in or above the 98th percentile.

We make no attempt to characterize the cause for the sudden surge in citations or project its consequences for investors. We have found that those kinds of judgements are best left to analysts who are intimately familiar with the corporation in question.

Web Luminosity and Relevance

Corporate citation rates vary over time due to a number of factors: the organic growth (or contraction) of their consumer businesses, various seasonal factors, and often unplanned events that catch the public's attention. At any point in time, each corporation has a characteristic level of citations that we call its "web luminosity."

For corporations in industries that are closely engaged with consumers (e.g., apparel retail chains or Southwest Airlines Co. [LUV]), that web luminosity will often be positively correlated to revenues. For many other corporations in non-consumer oriented industries (e.g., agricultural chemicals producers or Lockheed Martin Corporation [LMT]) there is very little correlation between web luminosity and revenues. We refer to the correlation of citation rates to revenue as the "signal relevance" for a given corporation.

Those characteristics are best understood when considered on an industry by industry basis. For example, although computer and auto manufacturers have substantial citation luminosity, most of that luminosity is support related and only a fraction of the citations are created by revenue generating activities. As a more specific example, most MSFT brand citations are directly related to product support activities, with many citations for products (e.g., "Windows XP") that are no longer producing any corporate revenues. And although "Corvette" is a highly luminous brand name, in the bigger corporate picture the product line has marginal impact on either revenues or earnings for GM (not to mention the fact that a significant portion of "Corvette" citations are from enthusiasts looking for the parts needed to restore models originally purchased decades ago).

In contrast, apparel retailers have very few post-purchase support activities and most on-line consumer citations are related to potential future transactions.

Each corporation has a unique web luminosity and signal relevance profile. We have found that, in general, the corporations with the highest signal relevance are engaged in the production or marketing of consumer discretionary durable goods that require minimal post-purchase support.

In order to avoid spurious "false positive" correlations between citation rates and revenue, it is important to understand a plausible causality between the two variables. Industry groupings provide a macro way to think about that causality: a high correlation between consumer citations and revenue is more plausible for a fast food chain than it is for an agricultural chemical producer. For this reason corporate "industry" classifications are helpful to determine not only the availability of "Big Data" (web luminosity), but for the usefulness (relevance) of that data as well.

Revenue Proxy

Quantitatively, we have found that about 40% of the Russell 3000 has citation share-to-revenue correlations that are materially positive.

In these cases, the real-time citation rate metrics can be a reasonable proxy for as-yet unreported revenues. Citation metrics are anticipatory for two reasons:

– The activity being captured in "Big Data" is at the leading edge of the distribution channel (and in many cases even prior to consumer transactions);

– Formal earnings reports necessarily lag revenue transactions.

For research driven tactical managers, daily revenue proxies can provide a considerable edge.

The Price Causality Chain

It is also important to understand the integrity of the entire causality chain from citations to revenue, then revenue to earnings, and finally earnings to equity price.

Each of those steps may from time to time fail to correlate. Revenue may not correlate to citation share for any of the reasons mentioned above. Earnings may decouple from revenues for any number of operational and non-operational causes. And prices may decouple from earnings because of M&A activities, adverse news or systemic market movements.

For that reason we also monitor the integrity of the citation to price causality chain for each tracked equity by calculating the correlation between trailing citation share growth and equity price movements. In a sense this is our "bottom line" correlation data, and it has proven over time to be one of the more powerful alpha generating metrics.

The process of identifying the integrity of the causality chain also provides metrics on the historical latency time (or "lag") between citation share changes and subsequent equity price movements. This lag time will be dependent upon each equity's quarterly reporting schedule and the length and complexity of the corporation's product distribution channel. In turn that "lag" time can provide useful information to portfolio managers wishing to anticipate earnings surprises.

A new Fundamental for "Smart Beta"?

"Smart Beta" is an investment community catch-phrase that covers a wide range of passive and semi-passive formula driven investment strategies that vary at least to some degree from the traditional market capitalization weighting of the constituent equities in an index – which arguably over-weights over valued equities and under-weights those with relatively smaller capitalization levels. The new "Smart Beta" weighting formulas can be as simple as reallocations to equally valued positions, or weighting constituents according to more "fundamental" metrics – such as earnings, dividends or momentum.

"Big Data" resources provide yet another innovative source of such weighting criteria for formulaic investment models, with citation share and citation share growth being particularly attractive as alternative portfolio selection and weighting criteria, especially if slavish index tracking and composition matching are not primary investment goals.

In fact, since many equities included in the major indexes have neither the “web luminosity” nor the “relevance” necessary for inclusion in our “full universe” (see below), passive strategies utilizing the new “Big Data” fundamentals will start with a broad selection bias towards large-cap and mid-cap equities in the consumer discretionary sector. These selection criteria and weighting strategies can then be tuned in various ways for significant “excess” returns.

Proof-of-Concept Indexes

We have created a number of “proof-of-concept” indexes that demonstrate potential uses of our metrics in quantitative portfolios. These indexes are shown near the top of our BrandLoyalties.com home page under the heading “Examples of BrandLoyalties.com Metrics in 'Smart Beta' Indices (Proforma Performance)”.

Most of the indexes share a simple and basic set of investment rules:

– The constituent equities are included in the BrandLoyalties published list of covered equities (and by virtue of that have a mean daily on-line brand name citation rate that is greater than three times the standard deviation of their daily citation rates).

– The constituents generally have mid and large market capitalizations (>= $2 billion).

– The constituents have a materially positive trailing year BrandLoyalties citation share growth ranking to trailing price correlation.

– The corporations included in the index at any reconstitution are in the best 20% of BrandLoyalties ranked corporations.

– Each index is completely rebalanced and reconstituted quarterly to equal valuations.

The differences between the various indexes shown on the BrandLoyalties.com home page result primarily from the application of sector and industry selection criteria to the list of BrandLoyalties covered equities.

Upon request we can provide quarter-by-quarter allocation information for each of the indexes shown on our home page. Additional information can also be found in our “Data Usage Example” paper.

What are the caveats to using our data?

Because of the on-line consumer data behind our BrandLoyalties metrics:

– They only work for companies with a strong on-line visibility – something that we call "web luminosity."

– They work best for companies whose principle operations are in the consumer sector – a factor that we refer to as "signal relevance."

– The accuracy of the "brand loyalties" signal generated is a function of both the "web luminosity" and "signal relevance" of any equity.

(For example, some equities have no presence in the consumer sector of the

economy (e.g., LMT); others sell a lot of goods to consumer, but no transactions or significant portions of that commerce can be found on the web (e.g., XOM); still other equities have highly luminous sub-units operating on the web (e.g., BRK-A), but whether those units are good proxies for the health of the entire corporation requires much further analysis; in yet other cases where the distribution channel is vertically segregated (e.g., KO/COKE or HOT/HST) the importance of brand loyalties has to be understood in the context of where that brand loyalty most favorably impacts the operating units along the distribution channel; in some cases (e.g., SHLD or BGP) no amount of brand loyalty can overcome other structural, financial or management issues; and lastly some equities (e.g. BAC or LULU) can become momentarily brilliant on the web for all of the wrong reasons.)

– A positive Brand Loyalty ranking can mask suicidal pricing, a temporary novelty factor, product quality or service deterioration that haven't yet harmed consumers, or other fundamentally unsound business practices.

– Real-world portfolios should only use Brand Loyalty metrics in combination with other equity selection criteria and common sense.

– In fact, the quality of your additional equity metrics is what will separate your performance from the other portfolio managers who are using our data.

– To provide a robust signal our metrics require a relatively high level of “web luminosity" – making even our "full universe" of tracked equities highly selective. It typically includes only companies with "forward looking" distribution models and products or services that are in demand. Because of this, our "full universe" significantly out-performs the S&P Consumer Discretionary Index.

– The lag-time between changes in BrandLoyalties and the consequential changes in earnings will depend on equity-specific details – e.g., fiscal calendars and the length, complexity and inertia of the corporation's distribution channels.

– We presently track well over 3,200 US NYSE, AMEX and Nasdaq traded stocks, although that number is increasing over time. A separate Global “universe” is under construction, with an Asian collection of over 2,150 equities and 1,300 European equities already active.

– And unlike other retail surveying technologies, the strength and accuracy of our signals will only improve as the web's share of total commerce inexorably increases.

Simply put: the "signal relevance" of any web-based data will change over time as consumers react to brands in varying ways – and for varying reasons.

How do we quantify and monitor “Signal Relevance”?

For each equity that we track we are able to calculate the correlation between consumer BrandLoyalties for the products of that corporation and subsequent movements of the equity's price. If corporate earnings (and therefore equity prices) increases with increasing brand loyalty, we will observe a positive mathematical correlation between equity prices and our BrandLoyalties rankings.

On the other hand, if prices do not correlate at all to our BrandLoyalties rankings, we will observe a low or negative mathematical correlation between equity prices and our rankings.

Thus our correlation data (for the trailing year and updated each day along with the rankings themselves) serves as a quantitative measure of the "Signal Relevance" for each of the equities that we track.

In addition to the raw correlation data, for our more sophisticated clients we also measure the time lag between changing BrandLoyalties and the consequential equity price movements. We do this by finding the time offset between the two series of data that demonstrate the best correlation. That data is also provided in our downloadable data sets.

Brand Name Mapping

The mapping matrix that associates brand names to their respective corporate owners is both critical and highly dynamic. New brand names are constantly being created, while brand names can also fall into disuse – or at least cease to generate meaningful revenue for their corporate owners (e.g., “Windows XP”) even as they remain highly luminous on the web. As a practical matter we attempt to map brands that represent in aggregate 95% of the current revenue of any given corporation.

We also take care to separate when possible corporate identifiers (e.g., “Apple” or “Apple, Inc.”) from the product identifiers owned by that corporation (e.g., “iPhone”, “iTunes” and “MacIntosh”). In some cases (particularly retail) that can become difficult, since in retail settings the brand and corporate names are often commingled (e.g., “McDonalds”, “Starbucks” and “Kohls”).

For non-US equities we have a global network of multi-lingual analysts that maintain the mapping matrix for equities domiciled in their region. And for some multi-national corporations the matrix reflects the multiple languages that provide the most material sources of their revenue.

Additionally, brand names can be functionally and simultaneously co-owned by multiple corporations (e.g., KO and COKE, or HOT and HST). Additionally, mergers and acquisitions can cause differences in historical ownerships when compared to the current “live” mappings (e.g., “Sealy Posturepedic”).

Our brand mapping matrix is one of our key proprietary technologies. It represents our single largest capital investment, our most important intellectual property, and our single largest ongoing maintenance expense. And it is not for sale.

The dynamic nature of the mapping matrix can present conundrums when looking at historical data. Live portfolio management requires history that is “pro-forma” so that the current quarter’s product citation share can be compared to the appropriate and corresponding product citation shares from earlier quarters – even if a merger or acquisition has complicated the comparison. Back testing data, on the other hand, requires historic files that remain "as published." On request we can provide data in either or both contexts.

What Data Is Provided?

For each market day we provide clients with:

– The date of publication, 14 days after data collection (Note: Our citation data is published for every calendar day, although our legacy data is published exclusively for market days.);

– The ticker symbol used as corporate identifier (Note: the number of equities covered increases over time as additional brand names achieve sufficient web luminosity that they can be included in our "universe" of covered equities. The number increases from slightly over 200 in January 2006 to well over 3,200 US by January 2024.);

– For data sets containing non-US listed equities, we provide the geographical scope of the IP addresses used to collect the citation data. In those data sets US listed equities have a default global geographical scope. Some equites (e.g., Honda or Toyota) may have more than one ticker symbol listed, with a different geographical scope (and language) for each of their citation counts – often with a global coverage under their US ticker symbol and country specific scopes for their symbols on other exchanges;

– The percentile ranking (0-100, to four decimal places) of the brand name(s) citation share growth for each equity on that given day among our entire tracked universe – calculated via an algorithm that weights both YOY change and 91-day slope. Best performing equities have lowest percentiles;

– The optimal correlation coefficient between the equity's citation share growth percentile rankings and adjusted closing equity price, at "lag-days" (see below) offset between rankings and (lagged) market prices, and over a 250 market day time span – approximately one calendar year (Note: this series cannot be calculated or included during the first year of our historic data);

– The number of market days of lag between percentile ranking series and equity price series that generated the optimal correlation. (Note: The correlation data exists only when 250 market days of consecutive/contiguous prior citation and pricing data is available [and the correlations and lags series cannot be reported for the first market year for any given equity]. Correlation and lag data will also be suppressed if the daily citation rate data for a given equity is statistically noisy [>3 sigma] – typically as a result of relatively low web luminosity. Additionally there may be several "sweet spots" of materially similar high correlation within a 250 market day time span [generally corresponding to quarterly or annual resonances], and the optimal lag may occasionally flip between those materially similar resonances – giving the appearance of discontinuities in the data series. In such cases the shortest recent lag period is operationally the most significant.);

– The cumulative YOY change in citation share growth ranking during trailing quarter, again as a percentile ranking relative to our entire tracked universe (Note: This YOY change percentile ranking is based on the cumulative YOY differences of this year's ranking –vs– the prior year's ranking for the same date, summed over the trailing 91 calendar days. Once the cumulative YOY differences are calculated for each equity, those differences are ranked into percentiles. Lower percentiles represent higher cumulative positive YOY differences. Since this metric requires a prior year's quarter of data for its calculation, this field cannot be populated for the first year of the data set.);

– How the citation share growth rankings have correlated to revenue changes over the trailing 8 quarters – as a percentile ranking (Note: This metric is derived from a simple correlation coefficient using quarterly revenue as one variable and inverse average quarterly citation share growth percentiles as the second variable. This series requires 8 prior quarters of both citation share and revenue data, and therefore is not calculated or included during the first two years of our historic data. It is also not available for many equities [principally non-US domiciled corporations] where the revenue histories are less readily accessible.);

– Similarly, the peer-relative percentile ranking of how tightly the citation share growth rank changes have correlated to equity price movements over the trailing 250 market day time span. (Note: This peer-relative percentile ranking is based on the previously mentioned price correlation coefficient. This metric essentially ranks the integrity of each equity's citation share → revenue → earnings → equity price causality chain. This field cannot be populated for times spans where the price correlation coefficient is not available.);

There is an additional data field that is available in the "live" metrics suite that is **not available** in the historic data files:

– An event risk percentile – warning of any potential revenue damaging event-related publicity (usually created by some sort of PR nightmare). These nightmares could be created either by especially visible or spectacular product or service failures, safety recalls, corporate fiscal implosions, M&A activities and/or executive suite turnover/turmoil, scandals, or malfeasance/criminal behavior. This metric captures very short-term (7 calendar days) multi-sigma deviations from base-line citation rates, and is reported as a peer-relative percentile ranking (lower percentile = lower event risk).

Data Formats

The BrandLoyalties metrics are available as an HTML table, as well as in a CSV formatted file (or on special request, XML and SQL dump file formats can be provided).

And the historic values for most of our BrandLoyalties data can be downloaded for analysis in the CSV file format (or on request XML and SQL data formats). All subscribers can access all of our historic data. Sample files are also available from our website for non-subscribers, but they cover only an historic two year period.

Daily Alerts

Subscribers also receive daily access to updated data tables and daily alerts of significant changes in the BrandLoyalties of the equities that you have chosen to follow. For example, subscribers would receive alerts if an equity:

– had its "brand loyalties" rise into the top 20% of all of the equities that we track;

– had its "brand loyalties" rise even further, into the top 10% of that same equity "universe";

– or, conversely, had its "brand loyalties" fall into the bottom 20% of our "universe";

– or had it drop even further into the bottom 10% of all of the equities that we track.

Practical Portfolio Management

We provide a number of metrics that can be used to construct portfolios that are likely to out-perform looking forward. Those metrics can provide a comprehensive and nuanced understanding of how corporations have been interacting with their customers over the past several weeks and the trailing quarter. The full suite of metrics is the ideal complement to the traditional research used by fundamental portfolio managers and the analysts that closely follow consumer oriented corporations.

However, many asset managers prefer to take a more quantitative approach to portfolio construction. To that end we have modeled a number of quantitative indices that reflect industry and/or capitalization based slices of the equity markets.

One of the goals of our index modeling was to identify the metrics within our suite that provide the best alpha contributions when used in simple quantitative approaches. Our modeling of quantitative strategies using various combinations of all of our metrics has consistently indicated that the highest first-order alpha contributions come from a combination of two of our metrics: the trailing quarter share growth percentile rankings and the correlation metric for the full price causality chain.

This makes qualitative sense since the “percentile ranking” identifies the equities with the greatest share growth in their web luminosity, while the “price correlation” metric identifies the equities where those citations are most relevant – to revenue, and then to earnings and finally to subsequent stock price movements.

In general we have found that screening our equities first by “price correlation” (e.g., by selecting only the best third at any given time) and then applying a second screen based on “percentile rankings” (e.g., by selecting only the top ranked 20% of the price correlation qualified equities) leads to a robust first-order long term alpha that can then be tweaked using additional quantitative tools.

Compliance Friendly

From a compliance perspective, information gathered from “Big Data” has the potential for both good news and bad news. The good news is that a collection of publicly available consumer brand name citations does not contain any “insider” or other privileged information. The data collection focus is far removed from potentially sensitive corporate information, and the "publicly available" aspect means that anyone sufficiently clever and resourceful has full access to the same information.

The bad news is that nearly everything on the internet is the intellectual property of somebody, and as such is likely to be copyrighted material. The storage and reproduction of copyrighted material in any manner without the consent of the material's owner would represent a serious compliance issue.

But it is important to understand what constitutes the intellectual property covered by the copyright. For example, if Thomas Jefferson were alive and writing today, he could both copyright and register as his trademarked tagline: "Life, Liberty and the pursuit of Happiness."

But the stand-alone reproduction of any of those individual words – or even a simple alphabetized list of all of the words contained in that famous catch-phrase ("and, happiness, liberty, life, of, pursuit, the") – does not convey the same intellectual concept as Jefferson's tagline and therefore would not infringe on his copyrighted intellectual property. Similarly, a dictionary of the words found on the internet does not violate US copyright law, even if those words have been observed occurring within copyrighted material.

All of our metrics are original statistical derivations from our own lexicographic studies of material that is publicly available on the internet. We do not cross any firewalls. We do not buy any data; every source document we analyze is publicly available to anyone with an internet connection and a common browser. And we do not aggregate or derive in any way from research done by others.

In simple terms, we deconstruct or disassemble publicly available internet material into its component words, then accumulate, sort and count the frequency of the occurrences of those words across all of the material encountered on any given day. The only material that we store is a cumulative vocabulary list of words being used on the internet on that day.

Note that all of the data collected is totally anonymous, making the resulting lexicon fully compliant with the GDPR, CCPA and any other similar privacy regulations.

Compliance departments often ask if we have received permission to use any copyrighted material we may have encountered on the web. Because of our lexicographic approach we are in full compliance with the United States Copyright Act of 1976, 17 U.S.C. § 107. In the context of that code section a statistical or lexicographic study of vocabularies found on the web constitutes a "fair use" research of the distribution of common words that are themselves in the public domain.

We do not replicate, distribute or store any of the raw materials we encounter. The only data that we store are disassembled, aggregated, sorted, and tallied lists of common language words that are themselves in the public domain – a compiled dictionary from which the individual original raw web materials (and their intellectual property content) cannot be reconstructed.

Our original research is transformative, lexicographic, statistical, word centric (i.e., thought, context and/or idea agnostic), without "substantiality" to the original raw material as a whole, and without impact on the "potential market for or value of" any original material encountered on the web – all in full compliance with 17 U.S.C. § 107.

Continuity Risks

Additionally, the continuity of our services is assured because we utilize no third party data and require no third party approvals, permissions or tolerance.

And our IT operational continuity is assured due to the fact that all of our processing and data resides in AWS and S3 environments – allowing for quick migration to alternate availability zones if necessary.

Summary

The BrandLoyalties.com data provides Portfolio Managers, Research Analysts, Risk Managers, Fundamentalists, Quants, Private Equity Analysts with the following benefits:

– Early Signals: Often weeks before corporate guidance on shifting consumer loyalty

– Proven Stock Selection Metric: ~30% annualized alpha (top 10% ranked equities since 2006)

– Down-side Risk Protection: Email alerts provide early warnings of waning brand loyalty

– Persistence in Market Cycles: Positive performance in each market cycle segment since 2006

– Robust "Big Data" Data-Mining: Over 100 million online and social media consumer citations captured daily

– Expanding Market Breadth: Coverage of over 3,200 U.S., 2,150 Asian & 1,300 European listed equities, and growing

– Compliance Friendly Data: All data derived originally by BrandLoyalties.com using lexicographic analysis of the public domain language components of publicly available internet material

– "Smart Beta" Potential: A new "Big Data" sourced set of fundamental metrics that could be applied in passive or semi-passive formulaic investment strategies to provide return premiums.

– **Unique Metrics**: Unprecedented insight into consumers via synthesized quantitative metrics:

  – Brand name citation share growth ranking, seasonally adjusted, over trailing quarter

  – YOY change in citation share growth ranking during trailing quarter

  – Citation share growth rank changes correlation to revenue changes, trailing 8 quarters

  – Citation share growth rank changes correlation to equity price movements

  – Lag time between citation share growth changes and equity price movements

– Event risk percentile – warning of unwanted negative event-related publicity

– Easy to Interpret: Daily Percentile Rankings for each equity

– Dynamic Rankings: Methodology adjusts for seasonality and over-weights more recent activity

– Supporting Charts: Provides quick visual confirmation of unfolding consumer trends

– Flexible Delivery: Data can be provided in most common IT and desktop formats, push/pull

– Daily Email Alerts: Highlight critical positive and negative changes in rankings

– Interactive Tools: Sortable web interface with drill-down capability for each equity

– Relevance: Correlation data (to both revenue and equity price) augments rankings

– Optimal Lag Time: Each equity has its own time footprint between online activity and price

– Ahead of the "Insiders": Receive signals even before corporate insiders fully realize what's happening at the far end of their extended distribution channels.

– Real-Time Validation of Marketing Plans: Know how corporate marketing plans are really going

– Anticipate Revenue Surprises: Don't be in the dark heading into Earnings Season

---

Detailed Notes:

1. The now well over 2200 US (3800 including Asia & Europe) equities are chosen for the "web luminosity" of their brand names on the web and in social media, as well as for the "relevance" of brand name citations to the revenue streams of the corporation. Note that the number of equities covered increases over time as additional brand names achieve sufficient web luminosity that they can be included in our "universe" of covered equities.

One of the characteristics of our "Big Data" analytical methodologies is that at any given time a number of corporations will exhibit web luminosity (in the form of a brand name citation count series) with borderline or inadequate statistical robustness (i.e., the equity fails to maintain a mean luminosity > than 3 sigma [standard deviations] of a rolling 14 day sample). Those borderline equities may be included or excluded from our coverage based on their citation luminosity signal-to-noise levels during the trailing two weeks.

Additionally, we are always expanding our coverage and new equities may be added to the list at any time. Hence the list of reported equities may change slightly from day to day, and the number of "rows" in the data matrix can expand or contract slightly on a daily basis.

2. Many of our clients have found that some of the best quantitatively managed alphas can be achieved simply by screening the equities based on some simple combination of best citation share growth percentiles (i.e., lowest percentile rankings) and price correlation percentiles on any given day. This has been confirmed by our own simulations, which use only those two criteria (percentile rankings and materially [one sigma] positive citation to price correlation) to generate the performance data shown on our home web page. Our other metrics can be considered secondary or tertiary qualifying data for tactically managed fundamental portfolios.

3. The vast bulk of our pro-forma "proof of concept" portfolio work has been performed to ascertain if there is any alpha inherent in "Big Data" and if that alpha is to any extent a function of the signal selectivity used in assembling a portfolio at any point in time (i.e., comparing models using our full universe to ones that selectively pick only our top 50%, top 20%, top 10% etc. of our ranked and positively correlated equities at any given time). We have found both significant and consistent alpha in those long portfolios and a positive correlation between that alpha and the selectivity used in assembling portfolios. Again, all of that testing was done using long-only models.

Our testing has shown that our "full" universe has significant alpha relative to "the market" as represented by either the S&P 500 or S&P Consumer Discretionary benchmarks (i.e., candidate equities for inclusion in our tracked universe generally had high web luminosity to generate adequate signal/noise ratios, and were therefore likely to have aggressive on-line consumer marketing and distribution strategies). Even our "bottom 50%" has shown a consistent alpha relative to the S&P 500. Hence shorting strategies utilizing any portion of our universe are problematic.

Equity lists extracted exclusively from our "full universe" can present challenges to asset managers that need to construct market or sector neutral hedges – since there might not be many members of our "full universe" that are suitable for the short basket at any given time. Successful hedging strategies generally exclude from their short baskets any equities in our "full universe." For "market neutral" portfolios this would generally mean the S&P 500 less those S&P 500 equities in our universe, and "sector neutral" portfolios would require a more complex matching process that paired equities in our universe with sector-matched equities not in our universe. If "beta neutrality" is also required, it might be necessary to use either our full universe (or at least the top 50%) in the long basket to increase diversification and hold down beta.

4. One of the narrower measures of the "relevance" of a specific equity's brand loyalty percentile rankings is found in the correlation between the quarterly average of those rankings (inverted) and the quarterly revenues posted by the corporation. We measure that correlation over the trailing 8 quarters, and revise the percentile rankings whenever new revenue data becomes available. That data series is not available for many equities (principally non-US domiciled corporations) where quarterly revenue histories are less readily accessible on a mechanical basis.

5. A broader measure of the “relevance” of a specific equity’s brand loyalty rankings and the investment opportunity represented by that corporation can be found by measuring the correlation between those rankings and the equity’s subsequent price movement (which serves as a daily proxy for anticipated revenue growth). In effect this measurement (and the percentile ranking based on the measurement) quantifies the integrity of the causality chain from citation share growth to revenue growth to earnings growth and finally to equity price movement. Our daily correlation data ranges from +1.00 (perfect correlation) to -1.00 (perfect inverse correlation), with a 0.00 correlation indicating no statistical relationship at all. Values typically fall in the +0.70 to -0.70 range, and the correlation for any given equity will change over time.

The correlations will also change when a lag time is introduced between the brand loyalties rankings and the subsequent equity prices. The appropriate lag time between changes in consumer loyalties and the consequential changes in equity prices will depend on equity-specific details – e.g., fiscal calendars and the length, complexity and inertia of the equity’s product distribution channel. This lag provides one of the critical values in our data, since it allows our clients to be aware of changes in a corporation’s relationships with their customers some time before that change is fully reflected in the equity’s price (or in some cases even before the changes are fully recognized by the corporation’s “insiders”).

We find the appropriate lag time by determining which lag generates the highest ranking/price correlation factor. This specific lag is used for our published “Best Correlation” data, and the lag itself is provided as the “Best Correlation Lag” data points. Clients should be aware that the computed lag times will naturally expand and compress on either side of regularly scheduled earnings and/or guidance reports, and that on occasion there may be several lag times that produce very similar correlations (often at quarterly, semi-annual or annual multiples). These natural “resonances” may result in a smoothly progressing series of “Best Correlation Lag” data points to suddenly flip from one resonance to another, giving the appearance of a discontinuity in the data. The key thing to remember is that the published “best” lag time is not the only appropriate lag time, and the lowest recently published lag time is probably the most important operationally – since it represents the worst case lead time that is available to adjust portfolios before the customer loyalty changes we have measured become fully priced into the market via subsequent earnings reports or guidance.

Very high correlations can mean that our insights into the behavior of an equity’s customer base are already understood by the market – either through ongoing transparency on the part of management or a particularly resourceful following by research analysts. For this reason they may not necessarily represent better “signal” opportunities than equities with much lower (or even negative) correlations. Also note that lag times for “Best Correlations” that are near zero are by definition meaningless, while “Best Correlation Lag” times that are less than ~20 days or more than ~150 days indicate that non-cyclical factors may be driving either our rankings or the equity’s prices.

The correlation data exists only for equities when sufficient consecutive/contiguous daily citation data and publicly published trading prices are available (and therefore correlations and lags do not exist even for the first market year for any publicly traded equity). Correlation and lag data will also be suppressed if the daily citation rate data is too noisy to be statistically reliable.

6. The YOY change percentile compares each equity's daily citation share growth percentile ranking with that same percentile ranking one year prior. The sum of the day-by-day differences over the trailing quarter (91 days) is then ranked against all other equities in our coverage universe, with greater positive cumulative differences resulting in a higher ranking (i.e., lower percentile). This metric (when coupled with the revenue correlation metric) can be used to anticipate either positive or negative revenue surprises during a pending "earnings season."

7. Our data is delivered daily on US market days before market open. Note however that there is a 14-day delay built into our data series (e.g., data set rows dated August 21st represent brand citations collected during the North American evening of August 7th, and are first published before market open on the next market day after August 21st). The 14-day delay addresses several business continuity compliance issues (e.g., guaranteeing uninterrupted data flows despite real-time data collection platform outages); allows for our internal quality assurance processes; and provides low luminosity equities an extended multi-week sampling period to meet our 3-sigma signal-to-noise threshold. The delay means that our data is not suited for high frequency "day traders." It also increases the latency in our alpha to 2 to 14 weeks when using our key metrics based on citation growth rates over the trailing quarter. All of our clients receive the same data at the same time with exactly the same offset.

8. Data delivery timing: Our data is generally available by about 6am US ET. We generally provide data by FTP "pull" from our servers. The data is provided in ZIP compressed .CSV files, and can be delivered either for just the most recent publication date or in a "forever-to-date" file. We could address other delivery formats should you so request.

9. When we use multiple brand names for an equity (which is the case for most of the equities that we track) the different brand names are weighted according to the ratios of the relative observed citation rates. Using CCL as an example, "Carnival" is weighted relative to "Holland America" and "Princess" according to their absolute citation rates in each and every given time period.

10. For our US universe of tracked equities, we restrict the IP addresses of the consumers making the citations to those assigned to US based ISPs. We do that for several reasons: a) most of our tracked retailers have their principal operations in the US; b) the vast majority of the web luminosity for those brand names came from US consumers; c) for multinational equities we felt that US demand might be a good first approximation (i.e., proxy) for global demand; and d) we were initially using English language key words that could be represented in a standard Western (Latin) ISO character set.

We can optionally capture global or foreign national IP addresses. The "global" citations that we capture are for the English rendition of the brand names in the Western Latin ISO character set. The "global" citations that we capture are for the English rendition of the brand names in the Western Latin ISO character set. European brands are captured using the appropriate local language as encoded in Western Latin ISO characters.

Asian national citation rates require matching the scope of the citation coverage with the appropriate brand name variations caused by local languages, customs, cultures and character sets. For example, collecting Japanese home market national citation rates involves collecting citations made in the Unicode 3.0 Kanji character set. Similarly for Korea the Unicode 3.0 Hangul character set is used, for Thailand the ISO 15924 Thai script is used, for citations in the Chinese home market at least Simplified Chinese GB2312 characters encoded in the Unicode 3.0 set are used, etc.

In data sets that contain non-US listed equities, we provide the geographical scope of the IP addresses used to collect the citation data. In those data sets US listed equities have a default global geographical scope. Some equites (e.g., Honda or Toyota) may have more than one ticker symbol listed, with a different geographical scope (and language) for each of their citation counts – often with a global coverage under their US ticker symbol and country specific scopes for their symbols on other exchanges;

11. Each of our equities is ranked among our "universe" of tracked equities in a zero-sum manner, which can amplify modest changes in relative citation rate shares into dramatic swings from the lowest decile into the highest decile in a matter of a few weeks. This is by design to capture citation rate share changes in an actionable manner prior to earnings reports or the publication of corporate guidance.

12. Although the pure first-order seasonality is successfully removed from our data through the use of year-over-year metrics, our "peer ranking" methodology within our universe of tracked equities ultimately amplifies modest citation share (i.e., market share) shifts over the trailing quarter. As a consequence, companies that consistently gain market share on a seasonal basis because of the nature of their products, markets or customers may show significant second-order seasonality in our ranking data. In such cases it is useful to compare our rankings on a year-over-year basis, in order to determine whether the upcoming earnings report is likely to provide a year-over-year revenue surprise one way or another.

13. We have also found some periodicity that correlates to the cyclical product introduction pipelines of some corporations (e.g. AAPL), and in such cases comparing our rankings on a year-over-year basis can provide useful information about consumer interest in the current year's crop of products.

14. We are also dealing with significant changes in raw citation rates over time as the internet, social media and mobile apps constantly evolve. Current citation rates are several orders of *magnitude greater for any brand name than they were in 2006 -- because of rapidly changing mobile technologies, internet accessibility, social practices, product marketing strategies, demographics and consumer cultures. Clearly the two orders of magnitude increase in on-line "Big Mac" citations in the US since 2006 doesn't mean that McDonald's is actually selling two orders of magnitude more Big Macs in the US – the increase comes mostly from an order of magnitude more people having moment-to-moment access to the internet and an order of magnitude greater inclination to tell others about even their most mundane food choices. By ranking each company against peer equities (also living in the same expanding citation rate ecosystem), our methodology effectively "normalizes" for the radically shifting technologies and cultures. Ultimately it is market share that matters, and within any given industry we feel that our rankings should plausibly foreshadow most brand loyalty market share shifts.*

15. There is an additional data field that is available in the "live" metrics suite that is not available in the historic data files:

– An event risk percentile – warning of any potential revenue damaging event-related publicity (usually created by some sort of PR nightmare). These nightmares could be created either by especially visible or spectacular product or service failures, safety recalls, corporate fiscal implosions, M&A activities and/or executive suite turnover/turmoil, scandals, or malfeasance/criminal behavior. This metric captures very short-term (7 calendar days) multi-sigma deviations in citation rates, and is reported as a peer-relative percentile ranking. Four sigma spikes will typically place a corporation at or above the 98th percentile;

16. From time to time clients have asked us to provide them with the "raw" underlying data. As mentioned above, raw citation rates are generally meaningless without the context of the highly variable sample sizes – which can fluctuate daily in some locales by an order of magnitude due to both bandwidth and latency issues. This problem requires comprehensive normalization based on our internal bandwidth and latency statistics (and several additional proprietary technologies) before any analytics can be performed. The "citation" data we provide has been normalized for several reasons:

– Daily "raw" citation counts are materially impacted by a number of factors that we have to normalize for -- including the highly variable day-to-day effective bandwidth experienced by our virtual server farm, latency issues in a number of locales, our ongoing server instance scaling, and exponential growth in both web content and user connectivity. Our proprietary normalization methodologies are one of our core technologies and they are based in part on carefully instrumenting our collection processes and on proprietary "standard candle" approaches.

Our "standard candles" are non-brand related words that should have on-line citation rates (i.e., luminosities) which are largely invariant over time. Significant variations from the historical norms for the citation rates of these standard candles (especially if widely observed within the *group when intra-group share ratios are preserved) can provide quantitative normalization factors to compensate for the daily bandwidth and latency fluctuations that we experience, both globally and regionally. To assist with compensating for inadvertent sampling biases, the "standard candles" are deliberately chosen to cover the same wide range of luminosities as we experience among the brands that we cover.*

– The source data is roughly 30 zettabytes of web content, and nobody has the resources or bandwidth to "scrape" or "crawl" the entire content in any reasonable time frame. In fact, vast contemporary cloud based server farms can scan only about a billionth of the available content on a daily basis. That billionth, in turn, needs to be representative of the whole and provide standard sample sets. We use several sophisticated sampling strategies to construct normalized aggregate citation rates.

– For any given equity we generally require 3 sigma signal-to-noise ratios to generate meaningful citation rate metrics. For less luminous equities these signal-to-noise ratios can be obtained only by using moving multi-day sampling apertures. Therefore for many equities the published daily citation rates are derived from statistically rigorous meta-analysis.

– Consumer citations, like most elective consumer actions, follow a weekly activity cycle – which can vary globally based on local customs. For that reason all of our reported citation metrics have been normalized for that cycle – and trailing "years", "quarters" or "months" within our metrics are in fact sliding time windows that are multiples of 7 days to capture full weekly cycles in each case.

17. Our citation data (and the metrics derived from it) have been normalized in the following steps:

– The nightly citation counts are normalized for bandwidth and latency fluctuations based on our internal bandwidth and latency statistics (and independently verified using several additional proprietary technologies, e.g. "standard candles").

– Citation counts for low luminosity equities (i.e., signal to noise ratios < 3 sigma) are separately derived using meta-sampling techniques involving sliding time windows that are multiples of 7 days each.

– Derived metrics (such as YOY change and trailing quarter share growth) are based on moving time apertures that are multiples of 7 days (e.g. 35 day "months", 91 day "quarters" and 364 day "years").

18. The primary operational data files available to clients are the current day's data, which contain only the most recent day's data and are available to clients on a daily basis in a .CSV format (although on special request .XML or .SQL dump formats are available). Clients may *"pull" them from our web site, or we can "push" them to a client's FTP site should the client so request.*

19. For some components of our metric suite we do not have data extending back to 2006. In certain cases (e.g., the correlation based data) it is necessary to have a sufficient span of historic numbers before the metric can be derived. For other metrics the raw information was simply not available (or being extracted from our sources) during earlier market periods. For these reasons some portions of our "live" metric suite are either not populated in the earliest eras of our history files or are missing entirely from the files.

20. There are two different versions of our historic data that are available: "proforma" and "as published":

(a). Using a provided link, clients may perform daily downloads of our proforma "BrandLoyalties Historic Data" file, which contains proforma data for every market day back to January 2008. This particular historic data file is intended as a real-time decision making aid for clients seeking to actively understand the historical performance of brand names that have caught their current focus -- especially during historic time frames with similar (or soon anticipated) economic circumstances. For that reason, those "historic" files are presented proforma, with the rankings of equities involved in merger/acquisitions or spin-offs presented as if the current equities had been historically represented by their current brand name mix.

These "proforma history" files are also proforma by virtue of utilizing the current brand name mix for each of the current equities in our universe – capturing newly covered equities and new product citation rates retrospectively and excluding delisted equities and discontinued products or brand names. Thus from time to time the mix of data sources, equities and brand names covered will vary for any given proforma date range -- causing the relative rankings and citation rates to change from an earlier proforma file to the latest proforma file for the same equity and time period. Note that ad hoc client requests for historic brand name or pre-IPO research are also prepared using the proforma approach. These files generally carry file names that are some variation of "BrandLoyalties_Full_History."

This version of the history files is also proforma in the sense that it can now contain newer types of data that were not available (or being published) at the earlier dates covered by the files. An example of that is the metric describing the correlation between citation share growth rates and corporate revenue. That data field was added to our "live" metric suite in 2014, but it can also be easily recast retroactively for earlier quarters. Analysts and managers interested in utilizing that metric should construct models that are back tested using this more inclusive series of data fields.

Another difference between our "proforma" and our "as published" history files is that our early "as published" equity universe was materially smaller (~500 US equities) than our current universe (~2500 US equities, and an additional ~2500 Asian and European equities). That initial *set of 500 equities was selected from among the largest capitalization corporations publicly traded in the US. As a consequence, the early "as published" (2012 through 2016) is disproportionately "large cap" compared to our current universe and the "proforma" data. For that reason the "proforma" data better represents the equity choices available currently, and it is arguably a more appropriate back-testing set than the much smaller and large cap biased "as published" data set.*

Yet another data science issue behind the need for proforma data sets is that we continually encounter previously unknown -- but date-able -- citations that our lexicographic data collection had not tallied before. In order to provide the most accurate historical data possible, our current citation counts for any historic given day are the cumulative counts for on-line citations made by consumers on that day, regardless of when the pages containing those citations were crawled by us for the first time. Since 2008 we have expanded both the bandwidth and server instances that we utilize within the AWS ecosystem by several orders of magnitude. Since on any given night we cannot possibly crawl the entire 30 zettabytes of web content, we do our best to crawl a statistically robust -- but largely non repetitive -- sample of the web each night. That sampling methodology -- coupled with the massive bandwidth and instance expansions -- inevitably results, each and every night, in the acquisition of back date-able citations that we had not previously tallied.

Although our citation and share counts are constantly changing for any historical date (and are revised accordingly in our proforma data sets), in general newly discovered citations from 2008 are far rarer than newly discovered citations from last week. In principle the citation counts for any given date will asymptotically approach a statistically reliable value.

As a necessary consequence of the expanding coverage, the evolving brand name mix being used for any given equity and the tallying of previously unknown historical citations, identically named data fields (e.g., "Percentile Ranking") will generally differ between the "proforma" and "as published" for the same date and ticker. This is due to both the coverage driven share and ranking dilutions and the continually evolving brand names matrix represented in the "proforma" metrics (even without corporate actions).

We have also recently begun to provide another version of the "proforma" file that is designed for back-testing in a ML / AI environment. All of the fields in that version of the file have been fully scaled/regularized and the file is ready to be loaded into a neural network platform (e.g., Google's Tensor Flow) for training and modeling purposes, and there are metrics for each and every calendar day.

In our ML / AI data sets the "tickers" shown are not point-in-time. The "tickers" are used merely to identify the current corporate owner of a certain mix of brands (e.g,, the "ticker" AMZN:US identifies a corporation that has a current mix of brands that includes Whole Foods). For that reason our "tickers" do not track the complex point-in-time history of trading symbols, but rather *are simply used to identify brand name mixes that can then be associated to luminosity values during neural network training.*

The key point is that for the historical dates in the data set, the brand mixes are invariant (i.e., all rows using the "ticker" AMZN:US to identify Amazon, Inc. represents the historical brand citation luminosity for the brand mix currently owned by that corporation -- including Whole Foods). Our expertise is in the historical luminosity of brand names, not the historical use of trading symbols. Typically our clients use their own corporate actions databases to determine the historical trading symbols used at any specific point in time for the corporations that we currently include in our coverage.

Generally, BrandLoyalties clients use our cleaner "Proforma" back data for initial model building or neural network training. They then confirm that model or network using the "out-of-sample" data provided by our "As Published" back data to test for any possible "survivor bias" in their results.

The possibility of survivor bias is more of an academic issue than a real-world portfolio issue when using data sets based on consumer brand loyalty. Because of our data collection methodology, our core equity coverage consists of highly liquid consumer companies with an actively engaged customer base. Such companies are rarely delisted because of failed business operations, and instead are usually "non-survivors" only because of M&A or other corporate actions -- which, in fact, generally result in portfolio gains. For this reason survivor bias within our back data is only likely to be found in portfolio models that focus on less liquid and smaller capitalization equities.

(b). On request we can provide an "as published" data file with every market day back to when we first went "live" with our first institutional clients in January 2012, containing the equity rankings for each date from January 2012 to date as they were initially published (i.e., as of the morning of each historic date). In that version of the historic data the rankings of equities subsequently involved in de-listings, merger/acquisitions or spin-offs are presented as they were on each of the covered days, with each equity represented by their then-appropriate brand name mix. This file may be used for back testing, but clients should be aware that the date range is more limited and the underlying YOY data is intrinsically unreliable, since the mix of brand names for any given equity will be subject to change at any time – often rendering the YOY and slope data an "apples to oranges" comparison. Additionally, any YOY comparison of the non-proforma files will be distorted from time to time as we augment and expand our coverage – since the older "as published" data is neither re-scaled nor normalized to reflect the newer coverage. And, as mentioned above, the "as published" data (by definition) does not include metrics added after the date of original publication of our data. These files generally carry file names that are some variation of "BrandLoyalties_As_Published_Full_History."

As mentioned above, because of the large cap bias in our early universe composition, the "as published" set does not accurately represent our current capitalization mix for the years 2012 through 2016. For this reason back-testing with it is arguably not representative of the results that could be expected moving forward with our current mid-to-large cap universe.

We also caution that real-world portfolio managers, however quantitative their methods may be, are generally not blindly agnostic to "survivability" issues. For this reason real-world managers should significantly outperform portfolios based purely on this data.

21. It is interesting to note that even our "full universe" shows a significant alpha relative to the S&P Consumer Discretionary Index. Our robust screens require a high level of web luminosity – meaning that even our "full universe" is highly selective. These include many forward looking retailers that embrace the web and social media for both marketing and distribution (including a number with highly luminous transaction portals).

22. We have constructed a number of "Proof-of-Concept" Indexes that use our metrics and a very simple rule book. These indexes (and their recent performance data) are shown in a table near the top of our BrandLoyalties.com home page under the heading "Examples of BrandLoyalties.com Metrics in 'Smart Beta' Indices (Proforma Performance)". Links to corresponding Fact Sheets and a Data Usage Example paper are also provided.

Most of the indexes share a simple and basic set of investment rules:

– The constituent equities are included in the BrandLoyalties published list of covered equities (and by virtue of that have a mean daily on-line brand name citation rate that is greater than three times the standard deviation of their daily citation rates).

– The constituents generally have mid and large market capitalizations (>= $2 billion).

– The constituents have a materially positive trailing year BrandLoyalties citation share growth ranking to trailing price correlation.

– The corporations included in the index at any reconstitution are in the best 20% of BrandLoyalties ranked corporations.

– Each index is completely rebalanced and reconstituted quarterly to equal valuations.

The differences between the various indexes shown on the BrandLoyalties.com home page result primarily from the application of sector and industry selection criteria to the list of BrandLoyalties covered equities.

Upon request we can provide the simple rules used to calculate the indexes, quarter-by-quarter allocation information, and day-by-day index valuations. Additional information can also be found in our "Data Usage Example" paper. We would encourage quantitative investors to use those indexes as initial benchmarks for their own proprietary strategies for utilizing our metrics.