Pub Crawling at Scale: Tapping Untappd to Explore Social Drinking

Martin J. Chorley*
Cardiff University
m.j.chorley@cs.cf.ac.uk

Luca Rossi
Aston University
l.rossi@aston.ac.uk

Gareth Tyson
Queen Mary University of London
g.tyson@qmul.ac.uk

Matthew J. Williams
University of Birmingham
m.j.williams@cs.bham.ac.uk

Abstract

There has been a recent surge of research looking at the reporting of food consumption on social media. The topic of alcohol consumption, however, remains poorly investigated. Social media has the potential to shed light on a topic that, traditionally, is difficult to collect fine-grained information on. One social app stands out in this regard: *Untappd* is an app that allows users to 'check-in' their consumption of beers. It operates in a similar fashion to other location-based applications, but is specifically tailored to the collection of information on beer consumption. In this paper, we explore beer consumption through the lens of social media. We crawled Untappd in real time over a period of 112 days, across 40 cities in the United States and Europe. Using this data, we shed light on the drinking habits of over 369k users. We focus on per-user and per-city characterisation, highlighting key behavioural trends.

1 Introduction

We are witnessing a convergence in our online and offline personas. Activities that were once considered solely the domain of the real-world, have begun to encroach onto online territory. An example of this is the consumption of food and drink. It is now common for users to share details of their meals online, as part of the so-called #foodporn revolution (Mejova et al. 2015). As such, there has been a flurry of recent research looking at consumption habits through the lens of social media (Abbar, Mejova, and Weber 2014).

While there has been a bulk of food-related research using social media data, a topic that has received less attention is that of drinking alcohol. This is partly due to the perception that alcoholic drinks have far less diversity than food. Consequently, past papers have looked at the topic from a rather narrow perspective, *e.g.,* identifying alcohol abuse by identifying tweets containing the word 'hangover' (Culotta 2013). We argue that this, however, misses a significant opportunity. For example, the craft beer community has grown dramatically in recent years. In 2013, craft beer sales saw a year-on-year increase of 79%, with 74 million pints being sold in the UK alone (CGAStrategy 2013). There are significant supply-side expansions too, with 200 new breweries opening each year (Naylor 2014). Thus, studying such a dynamic ecosystem could prove extremely fruitful, offering potential insight into the behaviour of people from around the world.This is because it is well known that many countries form social communities around the drinking of alcohol, which are quite distinct from those formed around eating (Clarke et al. 2000). There are also clear health considerations that such data contribute to. Excessive alcohol consumption is one of the most significant preventable causes of death today. It caused 79,000 deaths and 2.3 million years of potential life lost each year in 2001–2005 in the United States alone (Kanny et al. 2011). Typically, such things can only be studied using labour-intensive (*e.g.,* interviews and surveys) or coarse-grained (*e.g.,* alcohol sales statistics) collection methods. Gathering reliable, up-to-date and fine-grained data on such matters could be extremely helpful for exploring the role and impact of alcohol in society.

With these considerations in mind, one recent social app stands out: *Untappd* centres on the consumption of drinks (primarily beer), allowing users to 'check-in' beers that they are drinking in a given location. These checkins are shared with friends, allowing users to interact. Although the principles underpinning the app may sound unusual, it has proven extremely successful. In 2014 its userbase surpassed 1 million, with 60 million beers being checked-in over just 3 three years[1]. This popularity shows little sign of abating. Hence, Untappd has the potential to shed extremely fine-grained insight into the social drinking habits of a huge population of users. Furthermore, unlike work modelling consumption habits using free-text (*e.g.,* via Twitter), there is no need to perform interpretation or translation of data. Instead, all check-ins are represented using a formal schema.

In this paper, we offer the first characterisation of Untappd, exploring user drinking habits around the world. To this end, we have crawled their publicly available data over a 112 day period, to gather all checkins for users in 40 cities. We begin by characterising the dataset to reveal the app's scale and popularity (§3). Following this, we separate our analysis into three broad themes. First, we shed light on in-

*Authors listed alphabetically

Copyright © 2016, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]Untappd Official Blog, *'Untappd is 1,000,000 strong and growing'*, 2014. http://blog.untappd.com/post/73638076039

dividual user behaviour, highlighting drinking habits, as well as how users interact with Untappd (§4). Second, we explore which beers are most popular, and how these vary across cultural boundaries (§5). Then, third, we inspect the social side of Untappd, exploring Untappd's social graph and homophily (§6). Our key findings can be summarised as follows:

- Untappd is a remarkably large-scale service. Over a 4 month period, we witness in excess of 5 million checkins across 40 monitored cities.
- Most users are responsible in their drinking (75% checkin a maximum of 4 drinks per day), yet we observe a core group of very heavy drinkers. We find that 7% consume in excess of 10 drinks at least once, with 7.5% of users checking in excess of 50 drinks during the period.
- We identify distinct diurnal drinking patterns, revealing differing trends across weekdays, weekends and typical working hours. Cultural trends can also be extracted with different temporal patterns in cities (*e.g.,* drinking peaks in Paris around 9PM compared to 7PM in London).
- We observe clear preferences for particular types of beers across different cities. Through this, it becomes possible to cluster users into geographic regions based on their checkins.
- A nascent Untappd social network exists, with a median friendship group size of 9. Despite its sparsity, we find distinct homophily in the social network, with a tendency for friends to consume similar types of drinks.

We conclude the paper by highlighting key implications from our work (§7 and §8). These include a range a possible apps that could be underpinned by our data and findings, particularly in relation to health monitoring (*e.g.,* enabling interventions through the social network).

2 Related Work

There has been a flurry of recent work into the relationship between food consumption and social media. Abbar *et al.* (Abbar, Mejova, and Weber 2014) investigated the food that people report eating on Twitter. They monitored 210k users to discover what food they reported consuming. They found a strong correlation between the food being consumed and the health of the host country, as well as the attributes of the user (*e.g.,* education). In relation to Untappd, they found that alcohol tended to be mentioned in urban areas, as well as having a weak correlation with obesity. Mejova *et al.* built on these past studies to investigate food consumption patterns in the United States via Instagram and Foursquare (Mejova et al. 2015). They made numerous observations, including that those attending local venues were less likely to be obese. Other work has used Twitter to estimate alcohol sales using keyword matching (*e.g.,* 'drunk' or 'hangover') (Culotta 2013), and text-analysis to extract health information about users (Paul and Dredze 2011). Tamersoy *et al.*, on the other hand, looked at linguistic features on Reddit to characterise long-term abstinence of users in the *StopSmoking* and *StopDrinking* communities (Tamersoy, De Choudhury, and Chau 2015). With respect to these works, Untappd provides far more fine-grained information, allowing individual locations and drinks to be captured, as well as other metadata, *e.g.,* intervals between drinks and ABV.

A key contribution of our work is the temporal and spatial analysis of user drinking habits. Silva *et al.* (Silva et al. 2014) explored similar spatial and temporal food/drink preferences via Foursquare. Interestingly, they found that cultural factors were often stronger than spatial (*e.g.,* French food has more in common with Brazil than British food). A common issue with studies such as these is that they use venue choices as a proxy for specific food and drink preferences. This contrasts with Untappd, which provides both venue choices and individual drinks explicitly. We also note that work in this area has typically been limited to food, rather than drink, with particular focus on health implications. We expand this work by studying alcohol consumption, and go beyond health-based implications to explore the social nature of the app. Another important difference between food and drink is the frequency and rate of usage, with often many drinks being consumed per night.

There have also been various studies of more general-purpose location-based social networks (LBSNs). Most notable is Foursquare, which also happens to provide the venue database for Untappd. Cramer *et al.* explored the reasons why people used LBSNs via in-depth interviews (Cramer, Rost, and Holmquist 2011). They found many motivations for using LBSNs. Obvious examples include socially driven desires (*e.g.,* knowing what friends are doing), although more unexpected reasons were discovered as well, such as the endorsement of venues and coordinating meet-ups. It is logical to think that Untappd might hold many synergies here; however, we conjecture that the specific nature of the application likely leads to rather novel motivations as well (*e.g.,* tracking popular beers).

In this study, we build on past work to offer a window into the world of online 'social' drinking. On one hand, we provide the first empirical study of a new social app, which, as of yet, lacks even a rudimentary analysis. On the other hand, we extract and analyse this data to shed light on a number of social patterns, which appear unique to drinking. The rest of the paper explores this topic in depth.

3 Dataset and Characterisation

We have crawled the publicly accessible Untappd API in real time to capture data relating to beer consumption in 40 locations around the USA and Europe between 14th August and 4th December 2015 (112 days). The Untappd API endpoint `/thepub/local` allows retrieval of all public checkins in a given locale. Polling this endpoint at regular intervals allows continuous monitoring of all Untappd checkins within a given geographic area. Note that we only collect data for users who have allowed their checkins to be public. Each checkin contains: the username, the location, the beer, the beer rating given by the user, as well as the number of 'toasts' the checkin received from friends ('toasts' are similar to Facebook 'likes'). It should be noted that only checkins that are associated with a venue are included within this data. 40 cities were selected for monitoring, 34 within the

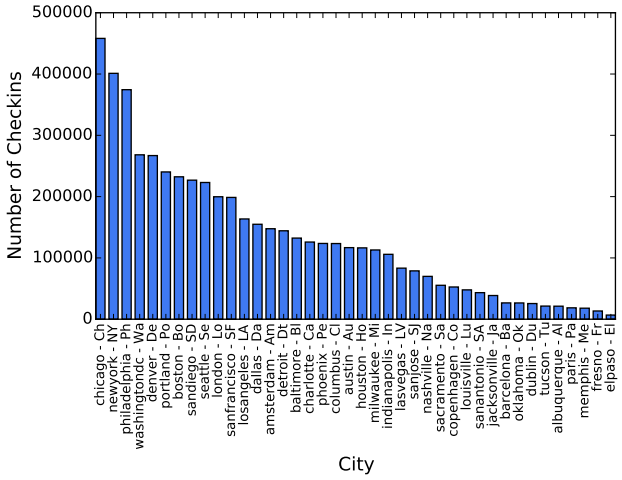


Figure 1: Number of unique checkins per city.

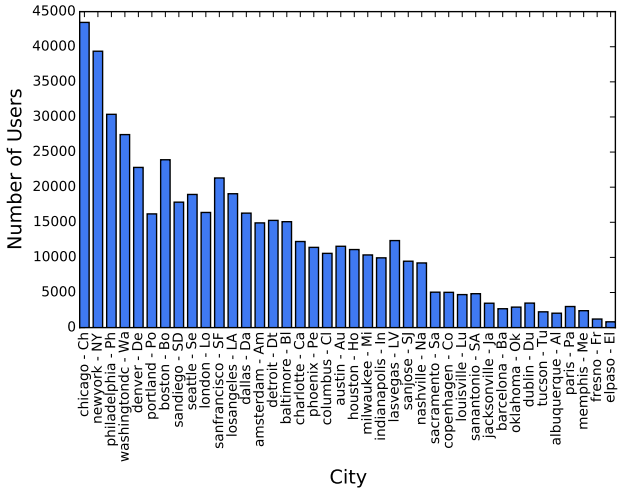


Figure 2: Number of users per city.

USA and 6 within Europe (chosen on the basis of population). In each case, we selected the central point of the city and collected all check-ins within a 25 miles radius.

To augment our checkin data, we also collected the full metadata for every beer encountered during the crawl. This includes the brewery, the type of beer and the ABV. The type of beer is based on a taxonomy produced by Untappd, which itself is an extension of the Beer Advocate styles hierarchy[2]. Since diverging from the Beer Advocate hierarchy, Untappd has introduced additional styles. We therefore manually curated our own taxonomy, mapping from Untappd styles back to Beer Advocate categories. The internationally recognised Cicerone Beer Guide[3] was used to resolve ambiguities. The range of beers covered by Untappd is impressive and missing beers are very rare.

Figure 1 presents the number of checkins collected across all 40 cities. The graphs contain the city names, plus abbreviations, which we will use throughout the paper. We find that cities have a wide range of user populations. Naturally, a key property that might drive this is the city's size. However, we note that Untappd does not particularly adhere to this intuitive assumption. For example, New York (ranked 2nd) has almost 20 million residents compared to under 3 million in Chicago (ranked 1st). More extreme examples are visible too; Denver, with a population of 663k is ranked 5th, well ahead of larger cities such as Los Angeles (3.8m). The European city with the greatest userbase is, by far, London, collecting 199,781 checkins over the measurement period; this still, however, ranks lowly (10th) compared to American cities with much smaller populations.

Figure 2 also shows the number of users per city. Broadly speaking, the number of checkins is proportional to the number of users in a city. However, notable exceptions can be seen. For instance, Washington DC and Boston have a disproportionately large number of users making a smaller number of checkins. In contrast, London has a smaller userbase, but with each user performing an above average number of checkins. In total, we witnessed 5,305,543 checkins: 4,834,811 were in America, with 470,732 in Europe. This covered 369,905 users and 139,759 unique beers. It is clear that Untappd has made significant inroads across a large number of cities, although its popularity is far greater in American than in Europe.

Before continuing, it is worth noting the limitations of the dataset. Most notably, there is no guarantee that users *always* checkin the beers that they drink (*e.g.,* a user may stop checking in beers after excessive consumption). Another issue is that we can only collect data when users associate their checkins with a location (within our measurement radius). Thus, we may miss some checkins. Of course, we cannot be sure that users checkin their beers at exactly the time that they consume them either. The data also does not include fluid quantity (*e.g.,* 1 pint vs. 500 ml); consequently, we can *only* measure drink counts, rather than exact amounts of alcohol consumed. Finally, as with all studies of this type, we note that our sample relies on a population of individuals who own a smartphone and are also eager to adopt mobile apps such as Untappd. Consequently, we are careful to scope our study as a exploration of Untappd users, rather than the population at large.

4 Characterising Untappd Users

We begin by characterising the usage of Untappd. Here, we focus on how individual users interact with the application, particularly in relation to the frequency and range of beers consumed.

4.1 Usage Frequency

First, we inspect the frequency of usage by Untappd users. Figure 3 shows the complementary cumulative distribution function (CCDF) of the number of checkins per user. We use the `powerlaw` package (Alstott, Bullmore, and Plenz 2014) to determine the best fit with respect to four candidate

[2] `http://www.beeradvocate.com/beer/style/`

[3] `https://cicerone.org/files/Certified_Cicerone_Syllabus_v2.pdf`

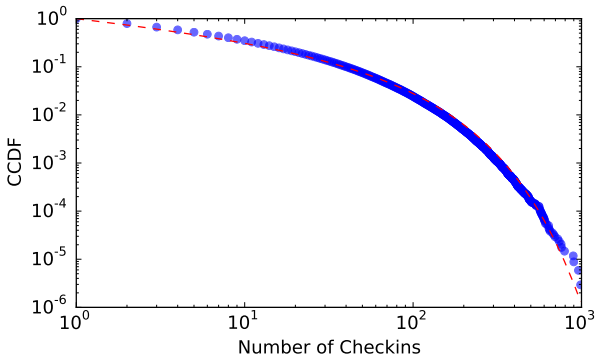


Figure 3: Users checkins distribution. The dashed red line shows the truncated power-law fit $f(x) = x^{-\alpha}e^{-\lambda x}$, with $\alpha \approx 1.29$ and $\lambda \approx 0.01$.

distributions, i.e., exponential, power law, truncated power law, and lognormal. In particular, we compare the truncated power law with alternative hypotheses via a likelihood ratio test, finding that in all cases the best fitting distribution is a truncated power law ($p < 0.01$). Note that a checkin indicates a drink has been consumed. A large number of users are quite occasional, with 65.2% having under 10 checkins. However, there are a small subset of extremely dedicated users. For example, we see that 2.44% of users have over 100 checkins in a 112 day period. Most users spread their drinks across a long period, however, we also observe many examples of intensive drinking. To check this, we compute the maximum number of drinks each user consumes on a daily basis. Most users drink quite responsibly, with 75.44% checking in a maximum of 4 drinks per day. However, we also find a notable subset of heavy users; 6.68% of users consume in excess of 10 drinks a day at least once during the measurement period.

We can also see how this usage frequency varies across cities. Figure 4 presents a box plot for the number of checkins per user on a daily basis. The mean number of daily per-user checkins is typically between 0.05 and 0.15. There is no clear rank that might relate to the numbers of users in a city. There are, however, some notable outliers worth mentioning. Interestingly, these tend to be either smaller American cities or European cities. The city with the highest per-user checkin rate is Portland (mean of 0.13 checkins per user per day), although London, Amsterdam, Copenhagen and Barcelona all report similar per-user per-day checkin rates to the larger US cities (between 0.08 and 0.1 checkins per user per day). This shows that, although America has a large userbase, Europeans are more active. This may be explained by the higher alcohol consumption reported in Europe (Peter Anderson and Galea 2012). Another possibility is that users in Europe have had to be more proactive to discover Untappd and, therefore, have a propensity to be more-committed to its usage.

That said, users in smaller American cities are also disproportionately active in using Untappd. The two highest ranked cities by absolute number of checkins are Chicago

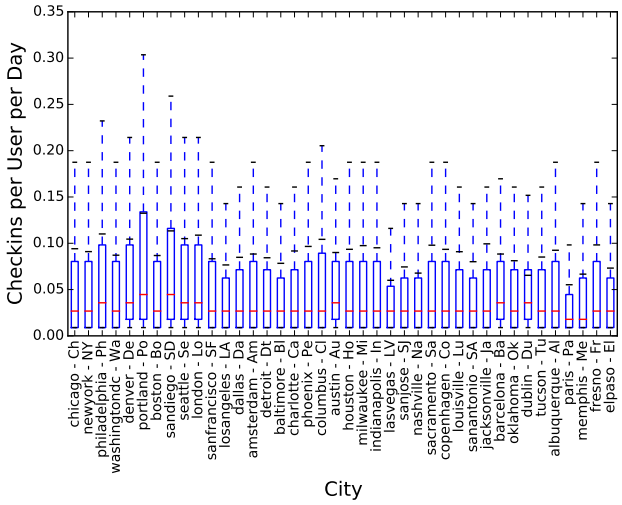


Figure 4: Average number of unique checkins per user per city on a daily basis (ordered by number of checkins across whole measurement period).

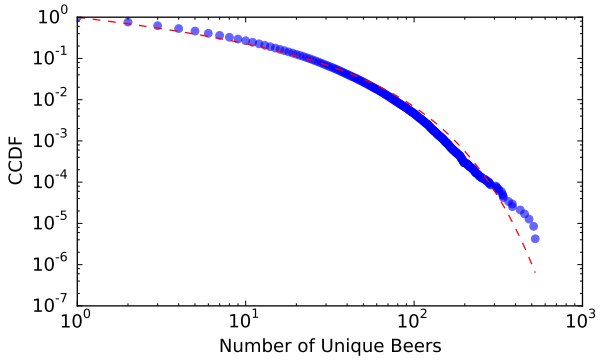


Figure 5: Unique beers distribution. The dashed red line shows the truncated power-law fit, with $\alpha \approx 1.29$ and $\lambda \approx 0.01$.

and New York. On a per-user basis, their ranks drop to 14th and 19th, respectively. Instead, small cities such as Denver, Portland and San Diego report far higher rates. In Portland, for example, 0.78% of users checkin a beer once every two days on average. This is in contrast to Chicago, for instance, where the equivalent percentage is 0.39%. Again, this is likely because users in small cities have had to be more proactive in discovering Untappd, whilst those in large cities have probably been exposed to it (*e.g.,* via friends, in bars) and installed it without necessarily being particularly interested in its long-term usage.

4.2 User Drinking Preferences

Next, we analyse the range of beers that individuals checkin. Figure 5 shows the CCDF of the number of unique beers per user. We find that it follows a truncated power-law distribution: A small fraction of users checkin a high number of unique beers, with the remainder far less active. A user, on

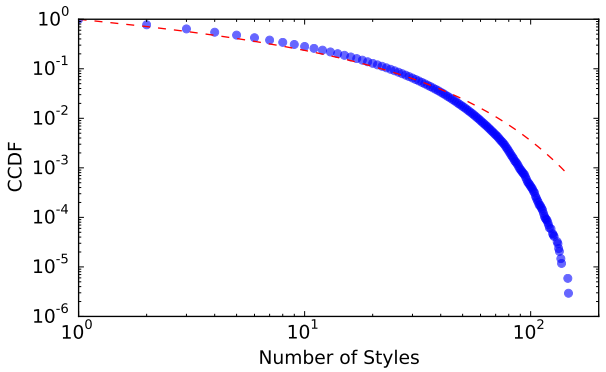


Figure 6: Beer styles distribution. The dashed red line shows the truncated power-law fit, with $\alpha \approx 1.23$ and $\lambda \approx 0.03$.

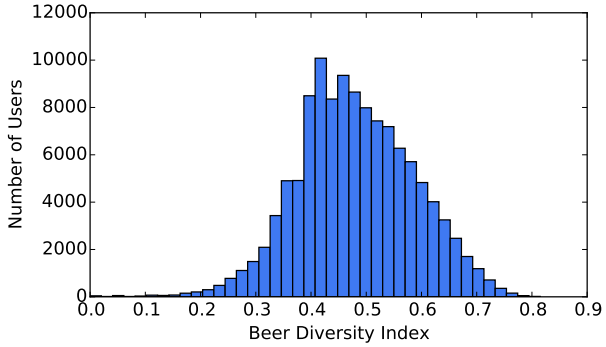


Figure 7: Shannon Entropy of the per user checkins frequency for all users with at least 10 checkins.

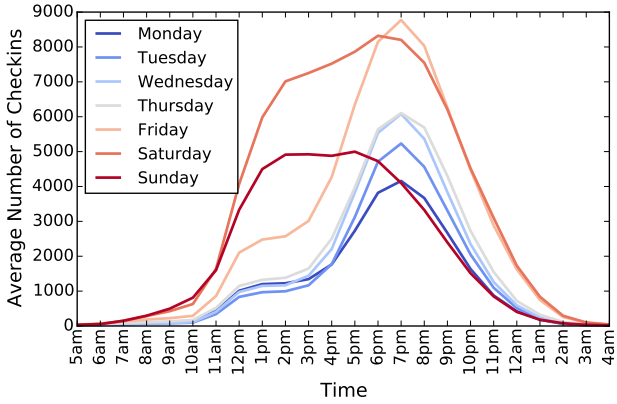


Figure 8: Average number of checkins per hour.

average, drinks 14.34 unique beers, with 54% of the users trying at most 5 different beers; 6.5% trying more than 50; and 1.9% trying more than 100 unique beers.

In order to get a better insight into the users drinking preferences, we characterise the affinity that individuals have to certain *types* of beer, *e.g.,* lagers, British ales, American ales, Belgian beers *etc.* Figure 6 presents a CCDF of the number of types of beers that users checkin. This is based on the beer taxonomy presented in §3. It can be seen that the bulk of users only drink a small range of different beer types. On the one hand, this is because many users simply do not checkin many times. However, on the other hand, we find that some users are extremely exploratory. The average number of beer types a users consumes is just 8.9, yet we find that some users (12.06%) consume over 20 different styles in our measurement period, and 6.1% consume more than 30 different styles. For reference, the total number of styles in the taxonomy is 222. If we restrict this analysis to those users who have made at least 25 checkins during the monitoring period (*i.e.,* those users above a 'casual' level of usage) we find the average beer types consumed rises to 30.13. Of these more active users, 98% have drunk at least 10 different styles of beer, 71% have drunk at least 20 different styles, and 37% have drunk at least 30 different styles of beer during the monitoring period. This illustrates the diversity of styles sampled by the active userbase of Untappd.

Finally, to address the varying number of checkins per user, we compute the Shannon Entropy of the per user checkin frequency. Higher entropy means a more uniform distribution across beer types, *i.e.,* a user who explores many different types. A low score indicates a user who limits themselves to a small range of types. Figure 7 presents the distribution of scores across all users that have at least 10 checkins. It can be seen that a normal distribution is followed. Whereas a few users are highly experimental, the majority show a strong propensity towards 4 – 8 preferred types. In summary, it is clear that most users *do* have clear affinities to beer types, with only a small minority of explorers willing to experiment widely.

4.3 Temporal Drinking Patterns

A particularly novel aspect of Untappd is the ability to checkin individual drinks 'live'. This gives us insight into the drinking patterns of different cities in terms of the *time* that people drink. Figure 8 presents the average number of checkins per hour across our entire dataset (normalised by timezone). It can be seen that clear patterns are followed. Most checkins occur between 4PM and 10PM, peaking around 7–8PM. Checkins during the working day are surprisingly frequent, starting most noticeably at 12PM (18.28% occur between 12–6PM).

It can also be seen that the rate at which people drink increases during the week: People checkin noticeably more on Thursday than on Monday (average of 12.7% checkins vs. 8.34%). Of course, Friday and Saturday evening witness the highest rates. It is particularly interesting to see the difference between these two key days. Whereas Saturday has far more checkins than Friday, these are spread across a much longer period of time (with a large number of users drinking between 12PM and 10PM on Saturday). Users also start to drink around this time on Sunday, although to a lesser extent. In contrast, on Friday a large number of drinks are consumed, but these are consumed in a far shorter time period. In fact, there is a higher peak on Friday than on Satur-

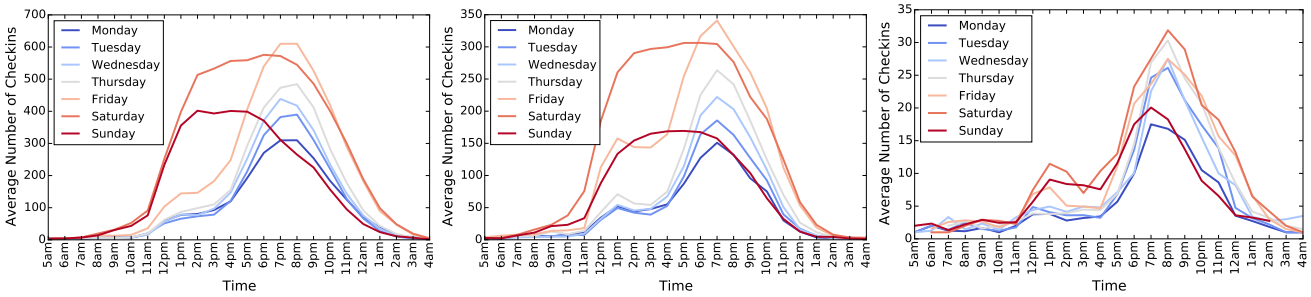


Figure 9: Average number of checkins per hour in (*i*) New York, (*ii*) London and (*iii*) Paris.

day (average of 20.57% vs. 19.23% of all checkins per hour at the peak). Overall, people checkin 55% more drinks on the weekend than during the week. People also start to drink earlier on Friday than other weekdays, suggesting that people often finish work early for 'after work' drinks. Saturday and Sunday have the greatest rate of daytime drinking. On Sunday a tenth of all checkins occur before 6PM.

Next, we inspect how these patterns vary across cities. Broadly speaking, all cities display similar trends across the week and weekend. Subtle cultural differences can, however, be extracted. We select three cities for comparison, shown in Figure 9: New York, London and Paris. First, it can be seen that the drinking culture in London is more binge-oriented than New York or Paris; when comparing Friday and Saturday night to other cities, drinkers in London consume more and over a longer period of time when compared against their weekday activities. For example, on a Monday, London collects an average of 996 checkins compared to 2,668 on a Friday, a 267% increase. This can be compared to just a 206% increase in New York. This propensity is particularly evident on Saturday, where users in London drink heavily during the day, peaking at 3PM. It can also be seen that Parisian users exhibits very different tendencies to both of these cities. Their daytime drinking is far less than in New York or London. Instead, they centre their drinking in at late evening, peaking around 9PM for on all days.

5 Finding the Beers that Matter

Next, we investigate which beers are important within Untappd. Unlike the previous section, we now focus on how beer popularity varies across geographic regions.

5.1 Where is beer heaven?

We start by inspecting which cities have the greatest range of beers available. Figure 10 presents the number of unique beers observed in each city. For each city, we normalise by the number of checkins there. For example, a value of of 0.1 indicates that 90% of checkins were for beers that had already been checked in there. The plot is ordered by the absolute number of checkins per location as in the previous section. It can be seen that the trends are very different, with an inverse relationship between the fraction of unique beers and the user population in the city. This is intuitive, as a larger population would indicate a greater chance of two users checking in the same beers. For example, in Chicago

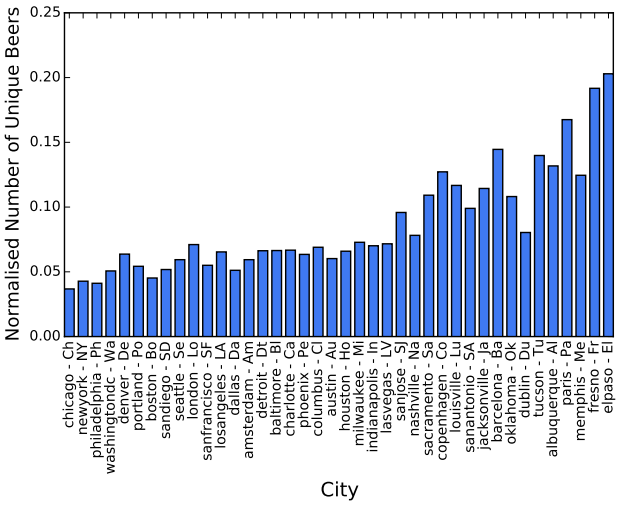
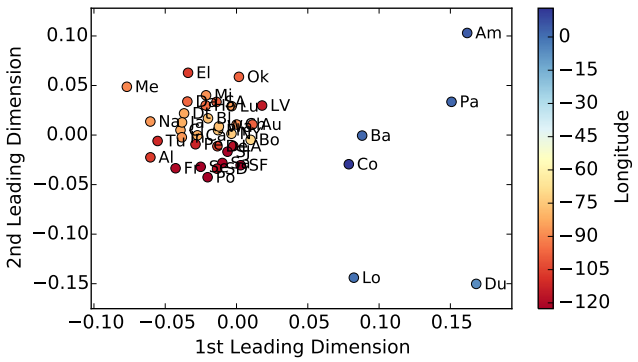


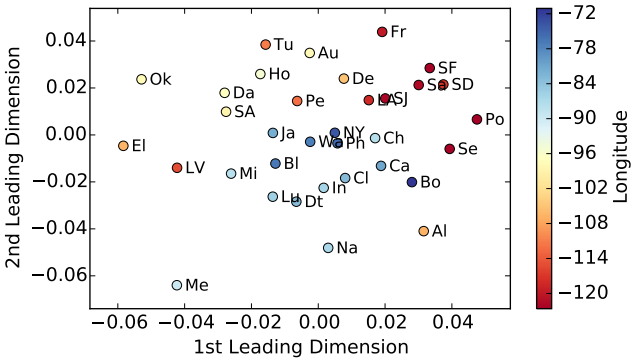
Figure 10: Number of unique beers per city normalised by the number of checkins in the city (ordered by number of checkins across whole measurement period).

and New York only around 4% of checkins are for a new beer, compared to over 20% for El Paso. In absolute terms, however, the opposite is the case, with the number of unique beers per city closely following the trend of the number of users there.

We can also look at the styles of beers available in each area. We separate all checkins into their respective cities, and compute a beer signature for each one of them. A signature is a vector, in which each element contains the probability of a given beer style being consumed in the city. We then compute the distances between each city's signature using Jensen-Shannon divergence (Lin 1991) and project their locations onto a two-dimensional space using Multidimensional Scaling (Cox and Cox 2000). Figure 11(a) presents the results for all cities, whilst Figure 11(b) presents the results for just American cities. Clear clusters can be seen, with individual locales exhibiting strong preferences towards certain types of beers. Most noticeable in Figure 11(a) is the divergence between American and European cities. This is largely led by the tendency that American users have towards American lagers and ales. Although many users in Europe also drink American beers (27%), they show greater



(a) All cities



(b) USA only

Figure 11: Similarity of beer signatures across cities. The signatures are computed using the probability that each beer style is consumed in a city.

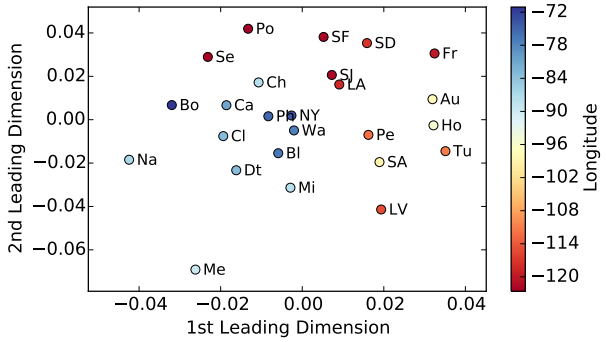


Figure 12: Similarity of beer signatures across American cities. To mitigate the impact of beer availability, the data is subsetted to only leave checkins for beers that are available in all cities.

diversity, consuming various English, Belgian and German beers more regularly. Thus, it is interesting to see that Untappd effectively captures these cultural boundaries.

We can also observe divergence within America itself, shown in Figure 11(b). This divergence emerges between East and West coast cities. A large part of this is, clearly, the availability of beers in the area. To explore this, we compute the set of unique beers consumed in each American city and then calculate their intersection. We find that 64% of beer styles are available in all American cities. This suggests that availability alone is not the only factor in the differing preferences seen. To understand the role of availability, we subset the American cities to leave those with in excess of 50k checkins. We then subset the checkins to leave only those beers that are available in all the cities. Through this, we mitigate the impact that availability has on our earlier analysis. Figure 12 shows the results. It can be seen that the clusters still remain, suggesting that different cities do have inherent cultural preferences towards particular beer types. The reasons for this could be diverse, and includes not only availability, but also differing costs to buy local beer versus imported beer. However, it is evident form this analysis that regional factors (beyond availability) do heavily influence the beers consumed.

5.2 Ale or lager?

We can inspect the different preferences across cities. Figure 13 presents a heat map showing the location quotient (Isserman 1977) (LQ) of checkins in each city for various style categories. The location quotient for a particular city and category indicates the number of checkins to that category in the city, relative to the global proportion for that beer style. Red indicates a strong preference (more than the global average) for a particular genre of beer; blue indicates the opposite. Various cultural preferences can be observed. For instance, unsurprisingly, it can be seen that London exhibits a strong propensity towards 'English Ales' (probability of an 'English Ale' is 4.0 times the global likelihood), whilst these remain broadly unpopular across other cities. This, of course, is largely a product of availability, which will be less in non-English areas. However, it is worth noting that even some American cities (*e.g.,* Phoenix and Dallas) show a liking for English ale. Other cities show less obvious trends, spreading their consumption across various types of beers. American ales do not stand out as particularly popular in the American cities. This is because *all* American cities tend towards American ales, leading none to an above average dominance. The European beer with the greatest export popularity is Germany, outperforming both English and Belgian ales in America.

5.3 Beer ratings

A particularly helpful feature of Untappd is the ability to rate beer from 0 — 5. This can be used to inform your friends of good/bad beers, but also to keep a personal diary of beers that have been enjoyed. Here, we briefly explore the distribution of these ratings across the entire dataset, shown in Figure 15. On the whole, users seem relatively positive about the beers they drink, with a mode average rating of

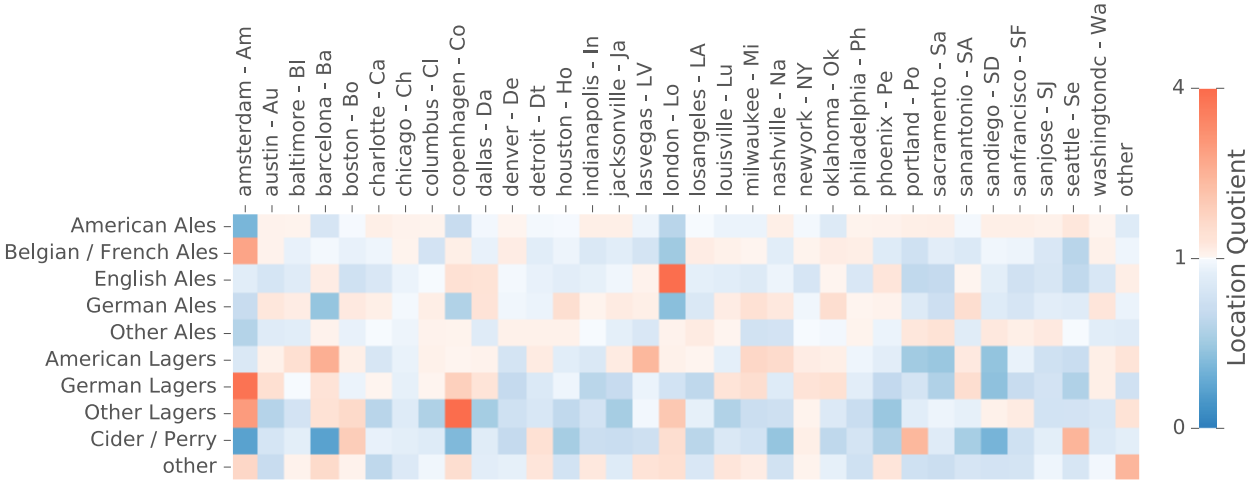


Figure 13: Beer styles location quotient. The location quotient for a city and category indicates the proportion of checkins to that category in that city with respect to the global proportion for that beer style. Red indicates a preference towards a style, whereas blue indicates a preference against. Styles with fewer 60,000 checkins and cities with fewer than 26,000 checkins have been aggregated into 'other' categories.

4. Very few beers score below 2 with the overwhelming majority 64% falling between 3 and 4. A particularly curious attribute of our findings is the difference between European and American ratings. Whereas the average rating in American cities is ≈ 3.75, this is instead ≈ 3.50 in Europe. Although this may not seem significant, the consistency of these scores are remarkable with American users nearly always marking one ordinal point above Europeans. The exact reason for this is unclear, however, we posit that differences in culture lead to American users being more generous.

Finally, we find that there is no correlation between the average score and the number of checkins of a beer. We measure Kendall's tau coefficient and we find that it is approximately equal to 0.0192, with p-value $< 10^{-8}$. This is surprising, as one would expect popular beers to also receive high ratings. Instead, it seems that users tend to experiment with new beers often rather than repeatedly consuming the same highly ranked beers. This gives powerful insight into the usage patterns of Untappd.

6 Untapping the Social Network

Untappd is equipped with a social network, allowing users to 'follow' each others' activities. Untappd's friend mechanism is symmetric; *i.e.,* both users involved in the friendship must accept it. A logical question is to what extent do users within a social network influence each other. To reconstruct the Untappd social graph, we generated a list of all users we observed with at least one check in before 30 September 2015. We then retrieved each user's publicly accessible friends list from the API. This crawl constitutes 254,868 users across all monitored cities. We then combined these 254,868 friend lists (*i.e.,* ego-centric networks) to build a single undirected network.

We first inspect the degree distribution of the social graph, shown in Figure 16. It is has been shown that real networks, including those emerging from human societies, display a power-law degree distribution with exponent α between 2 and 3 (Ahn et al. 2007; Kwak et al. 2010). Indeed, we find that the degree distribution of the Untappd social network follows a power-law with $\alpha \approx 1.85$. This reveals a sparse social graph, lacking the large social groups seen in other social networks. 52.74% of users have under 10 friends. The median friendship size is just 9 (mean 21.64), indicating that social interactions on Untappd are localised to a relatively small group of people. Further, we find a network density of just 0.0000069693, further highlighting the sparsity of the Untappd social graph.

Another common observation of the structure of social networks is *homophily* — the principle that individuals tend to be similar to their friends (McPherson, Smith-Lovin, and Cook 2001). Next, we briefly study homophily in the Untappd network, asking the simple question: to what extent are individuals' beer drinking preferences similar to their friends? To explore homophily, we use the approach of Aiello et al. (Aiello et al. 2012), whereby we measure the average similarity between friends and compare this empirical measurement to appropriate random null models. A null model preserves the same social structure (ensuring the same degree distribution and community structure), but randomises other features in the network.

We first construct and compare beer style feature vectors of two users using cosine similarity (Salton 1989). A feature vector represents the number of checkins by a user to each beer style in the taxonomy described in §3. User similarity is therefore based on the propensity to drink the same types of beers. Before analysis, we filter users to remove those that have fewer than six checkins or have visited fewer than six different venues (leaving 103,403 users and 456,426 friendship links).

Figure 14 depicts the distribution of friend similarities. The intra-city null model shuffles users within the same city,

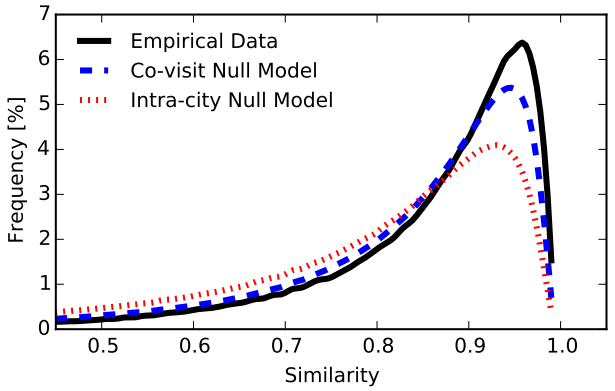


Figure 14: Homophily of Untappd users. Distribution of cosine similarities between friends according to the empirical data and two null models.

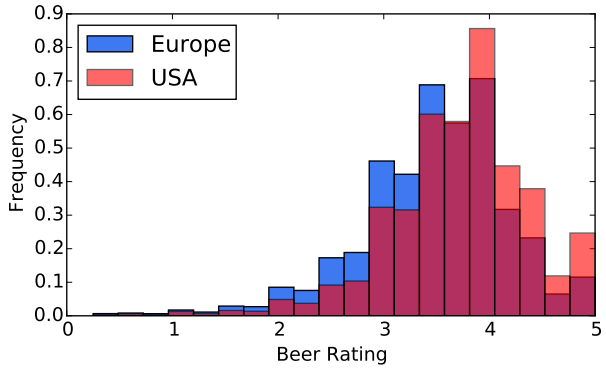


Figure 15: Beer ratings distribution in European and American cities.

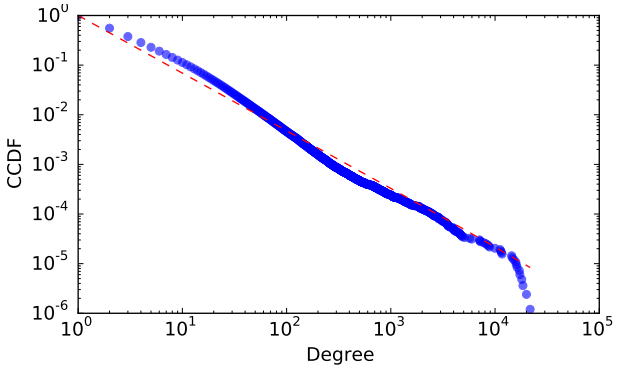


Figure 16: Degree distribution, social network.

while the co-visit null model randomly re-assigns a profile with a user who has visited the same venue (to avoid issues with beer availability). In both null models, the structure of the friendship network is not changed. The null models only act to randomly re-assign user profiles. It can be seen that the empirical similarities tend to be higher than the null models, confirming the presence of homophily in friends' drinking preferences. Histograms of null models are approximated by numerical experiments. Applying Kolmogorov-Smirnov tests, we find that both null models statistically differ from the empirical data ($p < 0.01$).

The average similarity for the empirical data, co-visit null model, and intra-city null model are 0.853, 0.829, and 0.782, respectively. Unsurprisingly, the co-visit null model tends to result in higher similarities than the intra-city null model, indicating that commonality in the venues visited by friends constrains the range of beers they are exposed to, pushing their profiles to become more similar. However, the statistical difference between the empirical data and the co-visit model indicates that, even after accounting for these correlations, homophily is present.

7 Implications

There are a number of implications from our work. We believe our analysis could offer a powerful insight for sociologists, dieticians, and psychologists, specialising in alcohol consumption. Here, we briefly discuss two key applications that could be built. First, our work could form the underpinning of future (automated) documentation of drinking activities. Various organisations frequently perform large-scale studies into drinking habits (*e.g.,* the World Health Organisation (Peter Anderson and Galea 2012)). These are usually aimed at assisting governmental policy construction. However, they are slow to be performed and often coarse-grained in terms of the data collected. Untappd offers the potential to near-automate this process, with huge bodies of data readily available across the globe. Of course, a key challenge is normalising and interpreting such data so that it can be generalised across an entire population. Combing Untappd with more traditional methodologies (*e.g.,* surveys) is a promising line of work here. Second, a number of further apps could be built atop of Untappd. These could have a variety of foci, however, intuitive possibilities include recommendation apps, targeted advertisement, social connectivity services (*e.g.,* facilitating meet-ups) and heath-related services. For instance, health features could be integrated into the app itself, with warnings provided to users when exceeding certain amounts. This might be particularly helpful for users who consume drinks containing a wide range of alcohol strengths.

8 Conclusion and Future Work

In this paper, we have explored the consumption of beer through the lens of social media. Using nearly 4 months of data, collected from the Untappd social app, we have characterised the activities of users from both Europe and America. We have found clear trends on both a per-user and per-region basis. In both cases, we find a strong affinity for certain types of beers. This is particularly prominent on a per-region basis, with different cities having very evident beer signatures. The clarity of these signatures was surprising, with the ability to automatically cluster users into geographic areas. Culture and availability are clearly paramount here, although it is difficult to ascertain exactly which plays the greater

role. Similar findings apply to temporal drinking patterns as well. These factors confirm that beer consumption *can* offer a fine-grained insight into the cultural properties of a region; whereas past work has also identified these ingrained traditions, we are the first to do this on a large-scale. These traditions relate closely to Untappd's growing social network. It was particularly interesting to see the presence of homophily in regards to friends' preferred beer styles. Again, this raises interesting questions regarding causality. We posit that Untappd friends may often share experiences (*e.g.,* drinking the same drinks together, or recommending drinks to each other). However, within our current dataset this is impossible to determine. With this in mind, we are keen to emphasise that our insights do not necessarily generalise to whole populations but, instead, shed light on the activities of Untappd users. Despite this, we find that the scale and real-time nature of Untappd data is unrivalled in this particular research domain.

There are a number of interesting areas of future work we plan to focus on. Most notably, we wish to expand our work on the Untappd social network. For example, it would be interesting to see how often Untappd friends drink together and how they influence each other. We posit that drinks might 'spread' through the social network via both online and offline recommendations (*i.e.,* two people drinking together in a pub). Quantifying this could be extremely powerful for recommendation engines and targeted advertisement. Another key line of future work will be to further investigate how cultural aspects impact drinking. So far, we have explored primarily American users, however, we intend to increase this to include users on all continents. We also plan to develop some of the apps discussed in §7. Finally, we conclude by saying that inspecting alcohol consumption through social media is a topic that has yet to receive sufficient attention. We therefore hope that Untappd could provide a catalyst for researchers to explore this topic more deeply.

Acknowledgements

We thank Lorrie Bonser and Thomas Galloway for valuable insights into the beer industry. We also acknowledge Dr. Jason Khan for his fruitful beer-driven discussions.

References

Abbar, S.; Mejova, Y.; and Weber, I. 2014. You tweet what you eat: Studying food consumption through twitter. *arXiv preprint arXiv:1412.4361*.

Ahn, Y.-Y.; Han, S.; Kwak, H.; Moon, S.; and Jeong, H. 2007. Analysis of topological characteristics of huge online social networking services. In *Proceedings of the 16th international conference on World Wide Web*, 835–844. ACM.

Aiello, L. M.; Barrat, A.; Schifanella, R.; Cattuto, C.; Markines, B.; and Menczer, F. 2012. Friendship Prediction and Homophily in Social Media. *ACM Trans. Web* 6(2):9:1–9:33.

Alstott, J.; Bullmore, E.; and Plenz, D. 2014. powerlaw: a python package for analysis of heavy-tailed distributions. *PloS one* 9(1):e85777.

CGAStrategy. 2013. Craft beer sales jump 79%. `http://www.peach-report.com/Trends/2039066/craft_beer_sales_jump_79.html`.

Clarke, I.; Kell, I.; Schmidt, R.; and Vignali, C. 2000. Thinking the thoughts they do: Symbolism and meaning in the consumer experience of the british pub. *British Food Journal* 102(9):692–710.

Cox, T., and Cox, M. 2000. *Multidimensional scaling*. CRC Press.

Cramer, H.; Rost, M.; and Holmquist, L. E. 2011. Performing a check-in: emerging practices, norms and'conflicts' in location-sharing using foursquare. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, 57–66. ACM.

Culotta, A. 2013. Lightweight methods to estimate influenza rates and alcohol sales volume from twitter messages. *Language resources and evaluation* 47(1):217–238.

Isserman, A. M. 1977. The location quotient approach to estimating regional economic impacts. *Journal of the American Institute of Planners* 43(1):33–41.

Kanny, D.; Liu, Y.; Brewer, R. D.; for Disease Control, C.; (CDC), P.; et al. 2011. Binge drinkingunited states, 2009. *MMWR Surveill Summ* 60(Suppl):101–4.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, 591–600. ACM.

Lin, J. 1991. Divergence measures based on the shannon entropy. *Information Theory, IEEE Transactions on* 37(1):145–151.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 27:415–444.

Mejova, Y.; Haddadi, H.; Noulas, A.; and Weber, I. 2015. # foodporn: Obesity patterns in culinary interactions. In *Proceedings of the 5th International Conference on Digital Health 2015*, 51–58. ACM.

Naylor, T. 2014. The craft beer revolution: how hops got hip. *The Guardian*.

Paul, M. J., and Dredze, M. 2011. You are what you tweet: Analyzing twitter for public health. In *ICWSM*, 265–272.

Peter Anderson, L. M., and Galea, G. 2012. Alcohol in the european union. consumption, harm and policy approaches. World Health Organisation.

Salton, G. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Silva, T. H.; de Melo, P. O.; Almeida, J.; Musolesi, M.; and Loureiro, A. 2014. You are what you eat (and drink): Identifying cultural boundaries by analyzing food & drink habits in foursquare. *arXiv preprint arXiv:1404.1009*.

Tamersoy, A.; De Choudhury, M.; and Chau, D. H. 2015. Characterizing smoking and drinking abstinence from social media. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, 139–148. ACM.