Tube Star: Crowd-Sourced Experiences on Public Transport

Neal Lathia
Computer Laboratory
University of Cambridge, UK
neal.lathia@cl.cam.ac.uk

Licia Capra
Department of Computer Science
University College London, UK
l.capra@cs.ucl.ac.uk

ABSTRACT

Public transport information systems have been shown to positively affect passengers' usage of their city's transport infrastructure, by providing information such as the location and schedules of trains and buses. These systems, however, lack qualitative information about passengers' ongoing experiences and may be out of date. In this work, we examine how smartphones can be leveraged to provide crowd-sourced information to transit passengers. We have deployed a prototype for London, England, that merges social media with transport statuses and report on the results of a thematic analysis of the content provided by passengers. We have found that passengers readily share their positive experiences more often than reporting problems, and find evidence that crowd-sourced reporting augments the timeliness of status information, for example, via reports flagging disruptions before the transport authority announces them. We close by discussing the limitations of our prototype and how these findings may inform the design of future transport information services.

Keywords

Public Transport; Crowd-Sourcing; Mobile

Categories and Subject Descriptors

H.5.m. [**Information Interfaces and Presentation (e.g. HCI)**]: Miscellaneous

General Terms

Public Transport; Mobile; Design

1. INTRODUCTION

Public transport already holds a central role in many city dwellers' daily lives. In this context, information systems that facilitate navigating the network play a critical role: they have been shown to strongly influence the uptake and usage of these sustainable means of transport [6]. Historically, these systems have tended to focus on delivering centrally managed, structured information about the *physical* state of the transport system, such as the location of transit services and train inter-arrival times [5]. In doing so, they inherently cannot capture the full breadth of the public transport's service quality (e.g., *qualitative* information about crowdedness), which may influence passengers' travel choices, and rely on being updated centrally, which may affect the timeliness and availability of up-to-date service information.

There are two main challenges that have prevented public transport information systems from providing this kind of data. First, by the mere virtue of being set in cities, public transport is used by a widely heterogeneous population [3], who will have varying preferences with regards to what factors guide their travel choices and their perceptions of the quality of trips; the transport authority simply cannot cater for everyone's needs. For example, a working commuter may care more about timeliness than crowdedness, while others will value seat availability over platform waiting time. Second, transport operators currently have no means of soliciting, collecting, and distributing any forms of information directly from and to their passengers, in order to measure the perceived quality of ongoing journeys. In other words, both *what* information to collect, and *how* to collect it remain unsolved. This setting stands in sharp contrast with the domain of online services and social networks, where users regularly share facets of their real-world experiences via tweets, status updates, check-ins, and photos. This contrast lead us to ask: *could public transport information systems leverage social media and smartphones to crowd-source a solution for the lack of qualitative information in their status updates?*

In this work, we examine the extent that crowd-sourcing can augment the diversity and quality-of-service status information of public transport networks by describing the design and initial evaluation of *Tube Star*, an application for the London Underground. The application merges the official status updates, provided by the transport operator, with functionalities reminiscent of online services: tweet-like reports and 1-5* ratings. We have publicly deployed the application and have received reports from 44 users. A thematic analysis of these reports reveals that:

- Travellers share both positive and negative experiences (the former occurring more than the latter); these reports cover a variety of qualitative information, including crowdedness, seats availability, heat and noise levels, and are often complemented with detailed location information about what station or segment of train line the report refers to. These reports further indicate that individual preferences can be elicited to understand what aspects of the transport system different users care about.
- The spatio-temporal granularity of submitted reports is more fine-grained than official Transport for London status up-

dates: we uncover episodes where passengers reported problems before they were announced by the transport authority, as well as incidences where passengers confirmed ongoing disruptions to their journey.

These findings offer valuable insights into the design of next generation travel information services: rather than passive consumers of travel status updates, travellers can be factored in as valuable sources of real-time qualitative information about their ongoing journey.

2. BACKGROUND AND SETTING

In order to frame our research, we discuss how public transport information systems have been developed and researched, with a particular focus on the growing number of mobile applications being built as a means of interacting with passengers on the go. We then describe the London Underground, including recent structural developments that facilitated the deployment of our application, and elements of its culture, as uncovered by a recent ethnographic study [1], that further motivates the potential of collecting qualitative, crowd-sourced reports from its passengers.

2.1 Public Transport Information and Mobile Phones

Public transport systems across the world are now regularly characterised by information displays, both on platforms, inside transit services, and online. More recently, these systems are being integrated into mapping services, like Google Maps, which allows transit authorities to make their data accessible to passengers via Google's services[1]. The growing adoption of smartphones further allows for many of these systems to be seamlessly accessed on the go. For example, the *OneBusAway* system [6] was built to provide transit riders in the Seattle area with real-time arrival information. The most frequently used interface to the system was the smartphone; overall, the system was shown to induce an increased sense of satisfaction, usage of public transport, and decreased waiting times. All of these systems provide a one-way channel of communication from transport authorities to passengers, and largely focus on *where* services are and *when* they will arrive.

The sensors that are built into modern-day smartphones also offer the opportunity to detect [16], monitor [7], and map or annotate [17] trips and locations in order to aide users' mobility. Services such as Hailo[2] allow for smartphone owners to quickly call for a Taxi with their GPS coordinates, or Waze[3], a community of drivers who volunteer their GPS logs as a means of measuring real-time traffic, fall under this umbrella. In the public transport domain, the *Tiramisu* system [20], for bus transit riders in Pittsburgh, allows passengers to contribute GPS traces of their trips and submit problem reports; passengers could then learn where buses are and, for example, how full they are. However, researchers have also found that "[bus] commuters [from Pittsbugh, USA] had little interest in reporting problems" [20], a question that we similarly address in a different city and transport modality.

2.2 The London Underground

Structure. London's public transport system is an interconnected multi-modal network of bus, train, boat, tram, and taxi services, operated by Transport for London (TfL). These services include three rail systems: the London Underground (commonly known as 'the tube'), the London Overground, and the Docklands Light Railway, an automated train network in the east of the city. These train networks are formed of 13 distinct lines and approximately 260 stations that have been geographically clustered into 9 concentric fare zones. The transport system itself is also one of the oldest in the world and has a daily ridership in excess of 3 million passengers: planned disruptions (e.g., for infrastructure upgrade work), rush-hour or large-scale event crowding, and unforeseen disruptions tend to occur rather frequently.

Information Services. Transport for London operates a number of information services to support travel planning and navigation. Most notably, these include the Journey Planner[4], which supports multi-modal routing across the city, and the Live Travel News[5], which provides an API to status updates for all train services and any station disruptions. More recently, TfL has begun operating a host of Facebook and Twitter accounts[6] that broadcast service status updates about each train line. We note that these accounts do not seem to be automated, but rather are manually operated by TfL employees.

Mobile Connectivity. Despite its name, less than half of the London Underground is actually underground, where passengers' mobile devices have no 3G coverage. During the summer of 2012, the problem of lack of mobile connectivity was further reduced: Transport for London and Virgin Media began deploying Wi-Fi at selected stations throughout the network, with an initial focus on those stations in the centre of the city that are underground. This rollout provides Internet connectivity which was initially freely accessible. Note that there continues to be no connectivity inside of underground tunnels.

A complimentary aspect of the London Underground is the perception, usage, and cultural habits of its passengers. Bassoli *et al.* [1] and Brewer *et al.* [3] have examined this in depth by performing an ethnographic study of a sample of tube passengers in 2006; in this section, we report on the aspects they uncovered that are relevant to our research.

Discouraging Social Interaction. The London Underground has a distinct set of social norms; amongst them is a trend of *civil inattention*, i.e., ignoring and not openly interacting with fellow passengers in an effort to not seem rude. Technology and traditional printed material has a mediating role in this environment: the usage of mobile phones, music players, books and newspapers allows passengers to create a social shield that prevents them from interacting with their neighbours. In this sense, mobile applications are a useful gateway towards maintaining social isolation within this often overcrowded setting.

Encouraging Social Interaction. Civil inattention is sharply contrasted with the interest and curiosity that passengers have about those around them, and subtle forms of inter-passenger communication patterns that regularly occur. In fact, technology and printed media also served the seemingly contradictory role of *enabling* social interaction: both implicitly, without requiring spoken words (for example, passengers leaving newspapers for each other), to

[1] http://maps.google.com/help/maps/transit/partners/

[2] https://hailocab.com

[3] http://www.waze.com

[4] http://journeyplanner.tfl.gov.uk

[5] http://www.tfl.gov.uk/tfl/livetravelnews/realtime/tube/default.html

[6] http://www.tfl.gov.uk/socialmedia

explicit conversation starters. Mobile applications that capture and reflect the ongoing experience of co-located passengers therefore also have the potential of building a further means of communication that continues to fall within the overriding social norms.

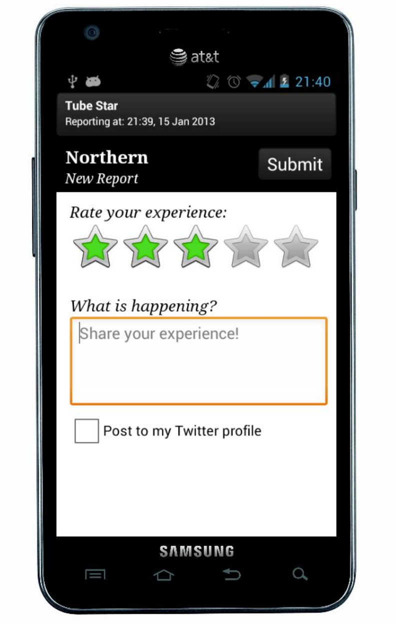
Given the above, which focused directly on our target deployment area, we believe that there is ample opportunity to explore the design of mobile applications for the London Underground which intersect the norms of social isolation with peoples' natural curiosity about one another, while addressing the open need to augment transport information systems. In the following section, we describe the design of *Tube Star*, an application that gives passengers a topic-centric platform to share their experiences about their mobility.

3. TUBE STAR APPLICATION DESIGN

We built a native Android application whose main functionalities are to allow passengers to (1) access information about the transport system and (2) share reports relating to their experience while using London's transport system. The mobile application dialogues with our server, which returns only those reports that had been created within the preceding two hours, and with TfL's public APIs for information from the transport authority. The client-side application stores both the most recent crowd-sourced reports as well as any reports that the user has submitted but have not been uploaded (for example, due to lack of connectivity). The application also runs a background service that seeks to upload all pending reports when the phone detects that it has regained connectivity. Lastly, the client uses the interface to the device's coarse-grained location sensor in order to filter any reports that are attempted to be submitted from outside a pre-defined radius from the centre of London. In the following, we offer a more detailed description of the application's functionalities.

Reports (Figure 1). We implemented a reporting interface that contained three main components: a 1-5* star rating, paired with a *tweet*-style text entry that is limited to 140 characters, and a colour-varying title bar that localises the report. Transport for London uses a unique colour code for each rail line that it operates, which is consistent across all interfaces that relate to the tube. For example, the Northern Line is colour-coded black: all maps, journey planners, and notice-board status updates show this line in the same colour. Our mobile interfaces use the same colour assignment. This interface also allows users to opt in to cross-posting their report to their Twitter profile; if they chose to, they could also automatically add the Twitter handle for the official TfL line that they were reporting about, as a potential means of notifying the transport authority of their report.

The rating is not tied to any particular aspect of the system (such as delays or crowds), so to avoid raising attention to pre-defined (often 'problem-related') aspects of the system. Similarly, the text entry does not ask for particular information but is left open to any content. The only guidelines that directed users as they entered reports were (a) to *rate your experience*, and (b) to answer the question "*what is happening?*" which included a hint text, "*share your experience*." We purposefully left these interfaces open to users' interpretation, in order to be able to see how their reports would be shaped by the context surrounding their current tube trip and, moreover, to analyse the extent that this context transpires in their reports.



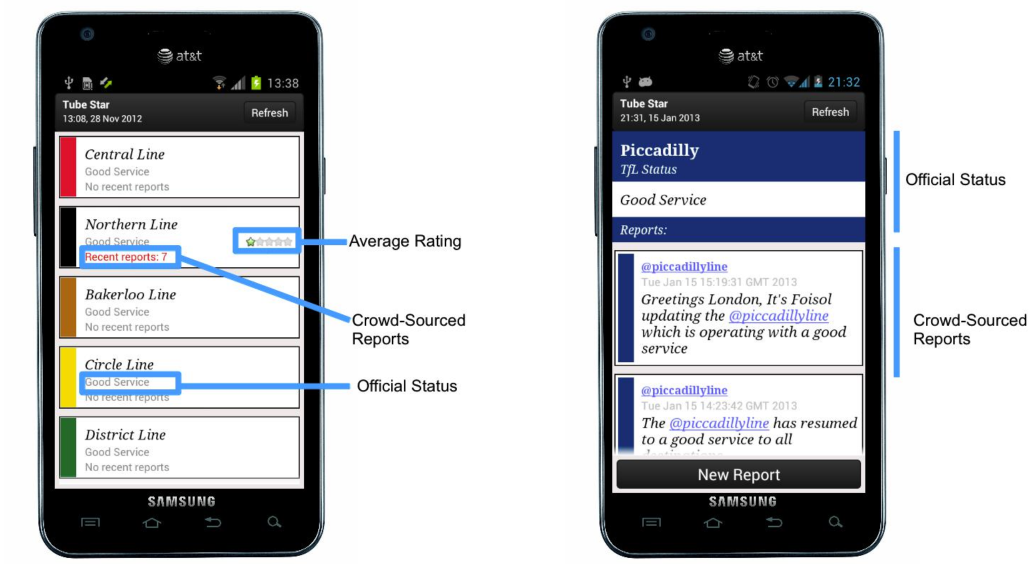
Tube Star
Reporting at: 21:39, 15 Jan 2013

Figure 1: Tube Star's interfaces for submitting a new report. The topmost part localises the report by being coloured and titled with the line (or station on a line) that is the subject of the report. Users are then asked for a 1-5* rating and a tweet-style report that asks them to *share their experience*. Finally, users also have the option to post their report's text to their Twitter profile.

Identities. Users can log in via their Twitter account or anonymously. An anonymously logged in user is limited to only providing star-rating reports (no text). A Twitter-authenticated user can input full reports (ratings and text), or rating-only reports; they also have a separate, optional check box to post their text reports to their Twitter profile. The default value for this option is to *not* post to Twitter; we further note that it was rarely used. We added a strong emphasis on the identity of users to limit potential spamming, although this is not our primary research goal: any reports that were submitted by Twitter-authenticated users were then linked to their Twitter profile.

Feedback (Figure 2). We built a set of interfaces that concurrently displayed both the official and crowd-sourced status data. The two sources of information are shown side-by-side; one interface shows the aggregated statistics for each line, while the second shows the detailed reports from each user underneath the TfL official status. We do not otherwise merge the two sources of information: any disparities between the two streams will be presented to the user.

Cold-Start Content. A typical problem of ungoverned user-generated content applications conditioned our design: initial lack of data (i.e., the cold-start problem [20]). To address this, we adopted a solution reminiscent of filtering agents, that has been implemented in online recommender systems [18]. The main idea behind *filtering agents* is to have a set of fully automated users who add content to the system in order to foster engagement from others. In our case, rather than attempting to algorithmically submit consistent, understandable reports, we opted to implement a set of crawlers that would pull the content of TfL's public Twitter posts and sub-



Tube Star
13:08, 28 Nov 2012
Refresh

Central Line
Good Service
No recent reports

Northern Line
Good Service
Recent reports: 7

Bakerloo Line
Good Service
No recent reports

Circle Line
Good Service
No recent reports

District Line
Good Service
No recent reports

(a) All Tube Lines

Tube Star
21:31, 15 Jan 2013
Refresh

Piccadilly
TfL Status

Good Service

Reports:

@piccadillyline
Tue Jan 15 15:19:31 GMT 2013
Greetings London, It's Foisol updating the @piccadillyline which is operating with a good service

@piccadillyline
Tue Jan 15 14:23:42 GMT 2013
The @piccadillyline has resumed to a good service to all

New Report

(b) The Piccadilly Line

Figure 2: Tube Star's interfaces showing both crowd-sourced and official public transport status information. On the left, we show the list of tube lines alongside their official colour, and include a summary official status, the number of recent passenger reports, and the average rating for the line in the last two hours. Clicking on an entry of this list will bring up an interface like that on the right, which shows the Piccadilly Line's current TfL (official) status, and a list of recent crowd-sourced reports. The latter also allows users to submit a new report for this line.

mit them, as reports, to the application.

4. APP DEPLOYMENT

The broad research question that we sought to address, while designing *Tube Star*, is the extent that transit passengers would readily share their experiences as they are navigating the London Underground. In order to investigate this, as well as analyse the content of submitted reports, we wanted to avoid recruiting participants who would be "prescribed" to use the app and instructed to submit reports. We therefore opted to publicly release the application, and advertise it via mailing lists, social media and a press release; we note that majority of the app's downloads occurred in the days following coverage from a local blog that focuses on London's public transport[7]. The biggest drawback of deploying the application this way is that it obscures our knowledge of who is actually using it [12]. For example, we know nothing of our users' demographics, and thus cannot analyse this. Moreover, we cannot track if a user has installed the app more than once: in the following, we use "user" to denote a unique identifier held by the system.

We have a number of means of tracking how many users the application has. First, from the Google Play developer console, we can track the number of *active* device installs, i.e., the number of people who have installed, and not then subsequently uninstalled, the application. Figure 3 shows the number of active installs relative to when we first put the app on the market; the surge in installs happened precisely when the blog post about our app was published. At the time of writing, the application has 156 active installs. The latest version of the application also records log-in requests, which is the first action that any user will need to do after downloading the app. From this, we see that 193 users have installed this version of the app, and 24 (12.43%) logged in with Twitter: the remaining users will be limited to reporting with ratings only.

5. ANALYSIS METHODOLOGY

There are two forms of analytic methodologies that we adopted: this section outlines these approaches, the precise questions that each of them can answer, and how they relate to the broader context of our application's evaluation.

5.1 Quantitative Analysis

The first step we took into analysing the application's data was quantitative, and focused on the 1 to 5-star ratings that users submitted; recall that users who opted to not log into the app via their Twitter account were limited to only providing this form of feedback. Collecting this set of ordinal values allowed us to formulate two main questions:

[7] `http://london-underground.blogspot.co.uk/2012/07/tube-star-app-connecting-london.html`

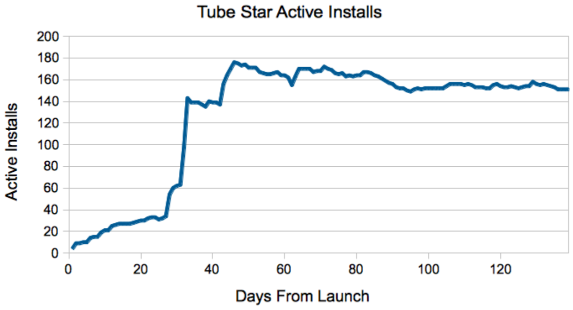


Figure 3: The number of *active installs* of the Tube Star application, relative to when it was first put on the Google Play market. Majority of the installs occurred when a blog that is exclusively about the London Underground published a post about the app.

- **Q1: Are ratings positive or negative?** Are users only reporting about disruptions, delays or, more broadly, only negative experiences? To investigate this, we examined the distribution of ratings received.
- **Q2: Are ratings consistent with the official status?** Do passengers report having negative experiences when a negative official status is being broadcast by the transport authority? Examining this aspect entailed a three-step procedure: first, we grouped TfL's status updates to uncover generic categories of information, then linked passengers' reports to the temporally closest status and, finally, compared the ratings that users gave across each group.

Both of these approaches solely focus on the ratings in user reports; we next describe how we analyse the snippets of text that users submitted.

5.2 Themes and Content of Reports

To analyse the submitted reports' content, we performed a *thematic analysis* of the textual content of each report. Thematic analysis has historically been used as a methodology to perform a systematic analysis of qualitative data, including free text survey responses and semi-structured interview transcripts [2, 9]. Much like Grounded Theory [8], this is a procedure where insights into topics and themes within unstructured text is created by iteratively visiting and encoding (or labelling) the data until an acceptable level of understanding has been reached.

The methodology behind these approaches to qualitative research are still open to debate, and precise methodological descriptions tend to be lacking in the literature [8]. In general, the intent is to draw inferences by following a well-defined procedure for identifying patterns in the data. In practice, there continues to be a tension between those advocating for purely inductive approaches versus those who allow the research to be driven by researchers' prior theory or understanding of the domain [19]. In order to clarify how we conducted this research, we describe, as above, the set of questions that motivated this particular kind of analysis.

In the case of *Tube Star*, our qualitative analysis was driven by a number of practical constraints. First, unlike semi-structured interviews, we do not have any means to seek further details if information is vague or seemingly incomplete: we must take the system's reports as they are. Furthermore, all of the data that is available for analysis was submitted by users who freely decided to do so, without reward: our insight is limited to reports from self-selecting participants. Given these constraints, we defined the following questions:

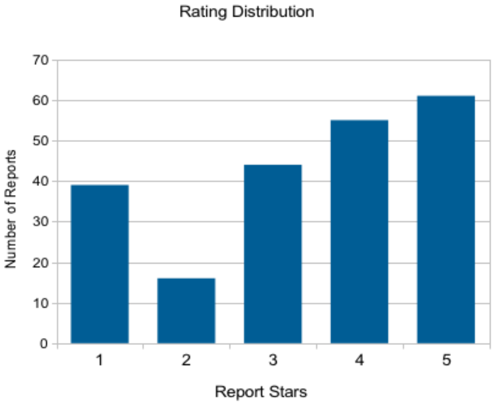


Figure 4: Tube Star's distribution of user-submitted report ratings: the spread of these ratings indicates a propensity to report both when positive and negative events are occurring.

- **Q3: What do users report about?** As we iterated over the data, we assigned a number of topic-labels to each report, in order to quantify the proportion of reports that discuss each topic. Our methodology for assigning topic-labels was not restricted to assigning a single topic, but rather gave a binary score to each report for *all topics* that we uncovered. This way, we were also able to investigate the extent that users reported on co-occurring topics.
- **Q4: What do ratings expose about topics?** We analysed whether there are themes that are, on aggregate, viewed negatively or positively. We examined the extent that topics correlate with positive or negative ratings by comparing how the encoded reports relate to the ratings that were associated to them.
- **Q5: Does crowd-sourcing add new information?** Finally, we investigate the extent to which user-generated reports are (mis)aligned, both in time and space, with official TfL status updates.

The following section reports on the results of analysing the data according to these six questions.

6. ANALYSIS RESULTS

In this section, we describe the results of analysing the reports that were submitted to the *Tube Star* application. We precede this analysis by briefly describing the aggregate statistics of the data, in terms of reports submitted by users and by our filtering agents.

Between June 28th 2012 and March 4th 2013, the app received 215 (rating-only and rating with text) reports by 44 users; this is an average of 4.89 ± 6.24 per user. We also note that the app's content was highly dynamic; the tube line accounts that our filter bots collected data from averaged $1,365.92 \pm 226.05$ reports each.

6.1 Are Ratings Positive or Negative?

The overall average rating for the reports was 3.39 ± 1.43 stars. Figure 4 plots the distribution of these ratings. The application interface sets the default value to 3 stars, which accounts for just over 20% of the ratings received. Most notably, over 50% of the ratings that were submitted are *4 or 5 stars*. This distinct skew towards the positive end of the scale indicates that the application's users are proactively submitting more positive reports than otherwise.

6.2 Are Ratings Consistent with the Official Status?

The generic statuses provided by the TfL API cover three topics: service *suspensions* ("Suspended" and "Part Suspended"), *closures* ("Planned Closure" and "Part Closure"), and *delays* ("Severe Delays" and "Minor Delays"). In the absence of any of these, a "Good Service" is reported. Note that more than one problem may cooccur, e.g., a part suspension *and* severe delays. In the following, we show how these service levels relate to the ratings that users submitted alongside their reports.

To analyse how the ratings that the users gave to their surroundings matched with the official status, we aligned users' reports to the temporally closest official Tweet broadcast, prior to each report, about the line that the user was reporting about. To account for potential service level changes, we only compare ratings to the official tweets that appeared up to two hours before the rating, which reduces the dataset to 79 ratings that arrived between 21 seconds and 144 minutes after the official status update (average: 44.38 ± 33.52 minutes).

We found that there were two largest groups: 61 ratings were submitted while there was a *Good Service* reported, while 7 reports were submitted on lines that were currently *Part Suspended*; the final 3 were during times of *Major Delays*. Figure 5 shows the normalised proportions of how these ratings were distributed for each kind of status. In general, there seems to be an agreement between the official status and subjective experience of passengers: the largest proportion (55.73%) of ratings during times of *Good Service* are 4 or 5 stars, and 60% of ratings given to lines that were *Part Suspended* were 1 or 2 stars.

However, these reports also highlight the possibility of *contradictory* experiences arising in urban settings [1]. In particular, 26.2% of the ratings submitted during times of *Good Service* were below 3 stars, and there is a 5-star rating while the service is *Part Suspended*. This has two implications: on one hand, users may perceive rather different experiences, even when travelling in the same contextual conditions; the ability to harvest users' preferences in context (e.g., in the form of ratings) is thus an essential step towards building personalised users' profiles and content filters (e.g., only showing reports from users of similar preferences). On the other hand, there may be a temporal misalignment (e.g., delay) between disruptions occurring, and official TfL feeds being updated. Crowd-sourcing may also offer a solution to gathering more realtime reports than what centralised solutions can afford, as we shell later discuss in relation to Q6.

6.3 What do Users Report About?

To uncover the topics that reports dealt with, we iterated over the reports' text, labelling each report with a set of one or more words that captured its topic. Our first iteration resulted in a preliminary set of categories: information about the *speed* of the service and any *crowds*. Given that reports could contain both positive or negative comments about these topics, we further decomposed each theme into two groups (which we refer to as dual-theme topics). For example, reports about crowds covered both instances where the service was very full or particularly empty.

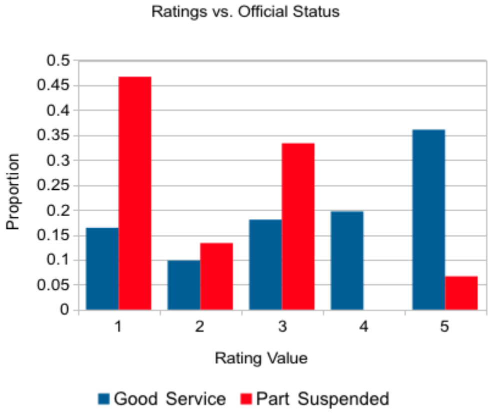


Figure 5: Tube Star's normalised distribution of user-submitted report ratings, when TfL reported a *Good Service* and *Part Suspended* services: while the overall distributions seem to agree with the official status, the potential for contradictory experiences arises.

Aside from these dual themes, there were a substantial number of reports dealing with *location* information (45.45%), *disruptions* that were encountered (18.94%), and *events* that were occurring (7.58%). A large proportion of reports detailed where the user was (for example, specifying which two particular stations a train they were on was, or which branch/direction of a line they were travelling on); users localised their reports, even though the application itself did not ask them to manually encode their precise location. In case of events and disruptions, additional information was provided: for example, one report was concerned with the break out of a fire at a particular station, and another one was formulated as a request for *"someone"* to clean up vomit on a platform. The top part of Table 1 shows these themes on the left-hand column, and the proportion of reports that were labelled with such category on the right-hand column. Note that very few reports deal with seating (a topic solicited in [20]); on the contrary, Londoners seem to care a lot about crowdedness, a topic that the transport authority's information systems do not convey; reports discussing a service as *quiet* or *busy* were also considered to be referring to crowding levels.

Finally, there were also a number of topics that we identified in smaller proportions of the reports. These included: the *temperature* (summer heat and cold from the air conditioning) and comments about the correctness of *local information* offered by stations' announcements or notice boards, and users whose reports engaged with the TfL status. This latter group included users who both reposted the official status (e.g., a report being *"good service"* when the official status was exactly the same) or commented on it: *"is it really good service? surprise, surprise"*.

Our analysis also labelled reports as *positive*, which contained words like *good, great* and *love*, or *negative*, which had words like *awful* or unhappy emoticons. Overall, we found a greater tendency to report positive rather than negative experiences: 25.76% of the re-

Theme	Proportion of Reports
Dual Themes	
Fast	11.36%
Slow	15.91%
Empty	14.39%
Crowded	21.97%
Non-Dual Topics	
Location	45.45%
Disruption	18.94%
Event	7.58%
Seating	3.03%
Temperature	6.82%
Local Information	7.58%
Sentiment and Self-Reference	
Positive	25.76%
Negative	15.91%
Self-Reference	16.67%

Table 1: Thematic analysis of the textual component of users' reports. This table shows the topics of the largest proportions of the reports. Proportions do not sum to 100% since a report could be about more than one topic.

ports were clearly positive, while only 15.91% of the reports were labelled negative. This finding suggests that Londoners are willing to engage with information services not just to notify fellow passengers of disruptions, but also (and perhaps primarily) as a way to share experiences. Moving through the city is thus not seen as a problem to be solved, but as a moment to implicitly interact, as advocated by [1]. We also counted the proportion of reports that contained explicit references to the person producing them: 22.1% of the data fell under this category, again suggesting that travellers care about sharing *their own* experience, rather than objective status reports (which official travel feeds already provide).

Based on these topic labels, we also computed a co-occurrence matrix, in order to identify how the relations between these categories emerged from the reports. Figure 6 visualises the result as a graph, where nodes are topics and the weights beside each link indicate the number of times we observed co-occurring themes.

6.4 How are Different Topics Rated?

Having labelled the reports into a set of well-defined topics, we now examine how topics relate to the ratings that users input alongside their reports. To do so, we computed the average and standard deviation of the set of ratings within each topically grouped set of reports. Note, as above, that reports can have more than one topic, and their rating will therefore appear in more than one group.

Figure 7 shows the result of this analysis, which agrees with common intuitions about public transport service quality (i.e., crowds and delays are bad; empty and fast services are good). Most notably, we found that those reports that contain one form of sentiment produce the most extreme average ratings: positive sentiment is accompanied with an average rating of 4.44 ± 0.77, while negative sentiment sits at the other extreme of the scale (1.61 ± 1.09). In the middle of this scale (and the topic with the highest rating standard deviation) are those reports that include self-references: reporting about a personal experience does not necessarily imply commenting on a particular type of event quality. This agreement between topics and ratings may indicate that the rating itself contains sufficient information to signal the difference between broadly 'good' and 'bad' service levels, although it is such a simple form of feedback.

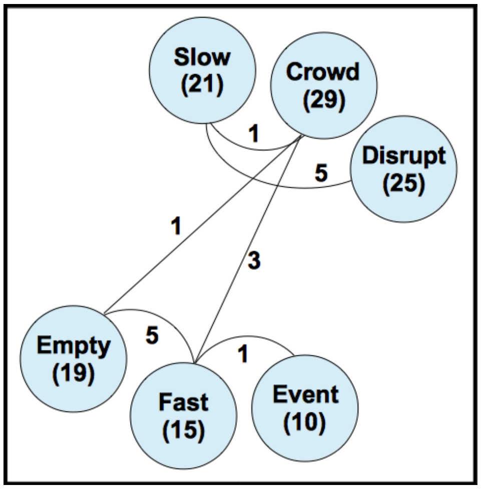


Figure 6: Graph showing the overlaps in identified topics between passenger reports. The numbers in each node indicate the number of reports identified with the given theme; links between themes are weighted with the number of reports where the two themes co-occur. Note that the apparently contradictory link between *empty* and *crowded* was in fact a report about a crowded train arriving at an empty station.

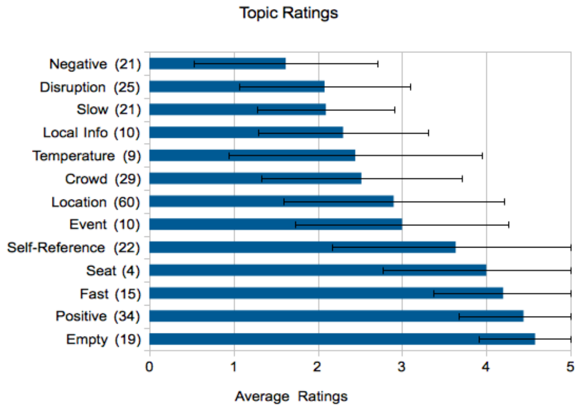


Figure 7: Relation between the topics that reports contained and the average/standard deviation of ratings given for those reports. In general, we find a strong agreement between the ratings and the topics covered: for example, reports containing negative sentiment have the lowest average ratings.

6.5 Does Crowd-Sourcing Add New Information?

We close our analysis by considering the extent that collecting report data from a crowd of passengers, who are surrounded by potential disruptions, compares to the centralised information from a *spatial* and *temporal* perspective.

Geographically Dispersed Reporting. As above, the official TfL statuses come from a limited set of categories (e.g., *Good Service*, *Part Closure*), which invariably represent a tube line as a single entity with potential localised problems. User reports, instead, can freely report about locations in an unstructured way, and have the ability to therefore provide finer grained information across the city. We returned to the report data to investigate the extent that this emerges, by selecting all reports that mention more than one location within the report text: in the following, we highlight three examples.

A user submitted a report when the Northern Line was officially experiencing *Severe Delays*. This line splits into two separate branches in central London, that are referred to as the *Bank* and *Charring Cross* branch (reflecting the names of two stations that are on each branch).

> *TfL staff says via Bank only, display says via Charing Cross only. Anyway, train overcrowded.*

There are two important observations exemplified by this report. First, the user-generated report offers much more detailed spatial information than what the online TfL feeds offer: rather than just stating *Severe Delays* for the whole line, it adds details about the way the functioning of the line is affected across its two branches. This additional information is of practical value to fellow passengers that need reaching stations on one specific branch (for example, they may use this information to divert their travels towards other routes – in London, it is often the case that multiple routes are available to reach the same destination). Second, the report highlights a contradiction that the passenger experiences, where the announcement made by TfL staff at the station does not match the information on the platform's notice board. Note that, regardless of the contradiction, this additional information from TfL staff is only available to passengers who are at the station at the moment of the disruption; crowd-sourcing reports via mobile applications can thus offer a practical means to promptly disseminate travel information across the city, leveraging travellers themselves as distributed information sources.

Similarly, two reports were separately submitted to the Piccadilly Line during times of *Good Service*:

> *At Kings Cross, they are still announcing that the line is part suspended near Uxbridge.*
>
> *Passenger alarm on train at Caledonian Road (just announced at Kings Cross).*

Both reports offer detailed spatial information about their ongoing travel experience. While relying on TfL staff to disseminate these reports may not be economically sustainable (in that there are not enough human resources on TfL to do so), we can build travel information services that exploit crowd-sourcing to harvest such information in near real-time in a highly distributed and efficient way, directly from the many travellers that at any point in time use the public transport network.

Timeliness of Reporting. Finally, we consider the extent that the reports preempt changes to the official TfL status. The main challenge of this analysis is that we cannot say, with certainty, that a report and status change are about the same event(s). For example, one of the reports about overcrowding on the Victoria Line was followed, five hours later, by an official update to *Minor Delays*. These two statuses are difficult to link, although a potential causal relation exists (overcrowding leading to delays). We therefore limited the amount of allowable time between reports/status updates, and sought instances of reports that were submitted within a 90-minute time window (both before and after) an official status change. We also manually inspected each report to ascertain whether the content of the reports was of a similar nature to the updated official status; the resulting set of four reports (see Table 2) co-occur with an official status update about the same topic.

Report (Summary)	Status	Difference
Passenger Alarm 16:14, July 10	*Part Suspended* 16:35, July 10	15 mins
Station Closed (fire) 07:11, July 31	*Part Suspended (fire)* 07:15, July 31	4 mins
Next train in 40 mins 08:34, July 31	*Severe Delays* 08:25, July 31	-9 mins
Station Closed 20:42, July 12	*Part Suspended* 20:05, July 12	-35 mins

Table 2: Do passenger reports anticipate similar changes to the official transport status? We found instances reflecting both (a) reports that preempt the official status and (b) reports that reflect the official status.

The examples we found include both reports that preempt an official change (positive time difference), such as a user reporting about a fire before a line is officially part suspended due to the same fire, and come after the change (negative time difference). The former set of reports reveal the potential of using crowd-sourcing as a means to dramatically reduce the time elapsed between the moment when a problem arises, and when information services notify travellers of such problem: for example, because the reporting passenger was on the train when the alarm was pulled, s/he was able to report the event 15 minutes before the official travel feed reported any problem (first row in Table 2). The latter set of reports is useful to reinforce the validity of centralised updates: for example, a passenger seeing a timestamp of 35 minutes ago as the latest TfL feed update might doubt its temporal accuracy; a just-issued traveller's report confirming the state of the system may add confidence in the official message (third row in Table 2).

7. DISCUSSION AND LIMITATIONS

The analysis above provides a number of key insights into how *Tube Star* is being used by London Underground's passengers. However, there are a number of constraints that we had to accommodate for that limited the extent we were able to evaluate the application. This section describes these limitations and how they may be addressed.

Design. As with similar systems [20], we were unable to anticipate the extent that users would contribute to the application's content until it was deployed. More specifically, we re-designed and re-built the application three times around different interaction schemas, which included asking users to check-in to their trip (in exchange for points that counted the number of miles they had travelled) and submitting well-defined reports that were structured around a hierarchy of questions, prior to the design that we ultimately deployed. Both of these designs were abandoned after testing them with a small group of University staff and students, as they were deemed overly constraining and inflexible.

Contact with Participants and Understanding Why. We continue to lack insights into how those users who adopted the application use it, or why they are incentivised to submit reports: in the future, this gap could be filled via *in situ* evaluations with the app's users and by collecting data about how regularly the application is used. The current version of the application also does not have explicit incentives for participation (such as virtual rewards or self-tracking functionalities), which is likely to have affected users' perception of the value of the application. Moreover, this emphasises that the content of our reports is inherently determined by a group of self-selecting passengers. Although this was our target audience, we note that this may not be representative of the wide range of views that passengers of the London Underground may hold. We have also taken a preliminary step into understanding why participants contributed by asking the Twitter-authenticated ones to complete a short survey, which has to date received 8 free-text responses that each contained *one or more* reasons for submitting reports. A full analysis is beyond the scope of this paper; however, these responses give an early indication that motivations ranged from helping others (5), passing time (1), adding information that is more accurate than official feeds (2), and hoping others will reciprocate (2).

Evaluation. Overall, deploying this application has enabled the thematic analysis of what passengers freely and willingly contribute to the system. This evaluation is thus relatively one-sided: it focuses on those who contribute to the application, rather than those who use the app passively without contributing. We note that the latter group is, as reflected in the analysis above, the majority of the application's audience. The next stage of our evaluation entails turning to the other side, and examining the value that passengers perceive from reading others' reports, by means of users' studies and surveys.

Open Research Questions. Finally, we enumerate an initial set of pending questions that we did not address in this research:

- **Generalisability**. Would the observed patterns hold if there were a larger experimental population? What if the application were deployed in a different city's public transport system? While the former question may be addressed via design and marketing attempts to expand the app's set of active users, the results already agree with larger-scale studies of the heterogeneity of public transport journeys measured via smart-cards [11]: each individual will differ in terms of travel time and service-level information preferences. The latter question, instead, may require revisiting and repeating the ethnographic studies [1] that uncovered the cultural facets of London's public transport.
- **Privacy**. To what extent are passengers inhibited from contributing because of privacy concerns? Our prototype did not focus on privacy, but allowed for varying levels of contribution: anonymously-authenticated users could contribute ratings that were not linked to their identity; reports were tied to tube lines rather than precise locations, which inherently have a large spatial granularity.
- **Malicious Usage**. Our analysis uncovered situations where the reported service level was different from the official information. We assumed that this arose due to differences between the lived experience of the user and the official service quality information, rather than suspecting participants of willingly submitting incorrect information. Previous research [10] indicates that, to be effective, malicious attacks would need to submit more reports than honest users: in our current context, where the broad amount of information received is low, we cannot address this further.

8. BEYOND PUBLIC TRANSPORT

Beyond the work described in the *Background and Setting* section, recent research also points to the growing role that social media has when mediating interactions between urban residents and their surroundings.

While smartphones provide a mobile window into public transport information systems, social media is also increasingly being accessed via mobile devices and, more broadly, its content is increasingly being associated with geographic locations and urban mobility. Prominent examples include Foursquare's check-ins and geotagged microblogging on Twitter. This wave of applications capture peoples' mobility without seeking to aide a journey. Location-based applications focus on social signalling and urban discovery [4]: this family of applications view mobility in the city "as a source of experience" [1] and an opportunity to create interactive experiences, rather than as a problem to solve.

Research into the digital footprints collected by these services demonstrates the extent that social media captures urban experiences. For example, geo-located tweets provide insights into the subjects and emotions expressed by neighbourhoods of varying social deprivation [15] and Foursquare check-ins uncover a city's spatio-temporal trends [13]. However, while these services readily capture urban residents' daily experiences, the channeling of this information centres around social networks of friends and followers; in doing so, they often are removed from the physical context which they describe and become seemingly unstructured (motivating systems to automate filtering their content [14]).

9. CONCLUSION

In this work, we have presented the design of *Tube Star*, an application that uses crowd-sourcing to harvest qualitative and real-time experiences of navigating within the public transport network in London, thus augmenting the information that current travel information services provide. More precisely, the application leverages the same techniques (ratings and tweet-style text) that social media sites use, but channels these into interfaces that reflect the structure of the transport system, rather than the users' social networks. We analysed the diversity of information that user reports contain via both a quantitative and thematic analysis of the reports we received. In particular, we found that majority of the reports submitted were rated positively and uncovered topics (e.g., crowdedness and heat) that the official transport information systems do not tackle; they provide a complimentary stream of information that passengers may use. We also found early indications of the potential for this system to provide reports about events before they emerge in the official status. The combination of *qualitative* and *real-time* information that can be gathered directly from travellers can be leveraged to build the next generation of personalised travel experience applications advocated in [1]. To achieve this goal, techniques to engage travellers into sharing experiences, as seamlessly and continuously as they may produce status updates on their Facebook and Twitter accounts, are required.

10. ACKNOWLEDGMENTS

The authors would like to thank Adam Barwell, Alexandros Zenonos, and Daniel Martinez-Oeckel who worked on an earlier prototype of Tube Star, and Annie Mole for blogging about the app on `http://london-underground.blogspot.co.uk`.

11. REFERENCES

[1] Bassoli, A., Brewer, J., Dourish, P., Martin, K., and Mainwaring, S. Underground Aesthetics: Rethinking Urban Computing. *IEEE Pervasive Computing* (July 2007).

[2] Braun, V., and Clarke, V. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology 3*, 2 (2006), 77–101.

[3] Brewer, J., Mainwaring, S., and Dourish, P. Aesthetic Journeys. In *ACM DIS* (Cape Town, South Africa, 2008).

[4] Cramer, H., Rost, M., and Holmquist, L. Performing a Check-in: Emerging Practices, Norms, and 'Conflicts' in Location-Sharing Using Foursquare. In *MobileHCI* (Stockholm, Sweden, 2011).

[5] Dziekan, K., and Kottenhoff, K. Dynamic At-Stop Real-Time Information Displays for Public Transport: Effects on Customers. *Elsevier Transportation Research Part A 41* (July 2007).

[6] Ferris, B., Watkins, K., and Borning, A. OneBusAway: Results from Providing Real-Time Arrival Information for Public Transit. In *ACM CHI* (Atlanta, USA, 2010).

[7] Froehlich, J., Dillahunt, T., Klasnja, P., Mankoff, J., Consolvo, S., and Harrison, B. UbiGreen: Investigating a Mobile Tool for Tracking and Supporting Green Transportation Habits. In *ACM CHI* (Boston, USA, 2009).

[8] Furniss, D., Blandford, A., and Curzon, P. Confessions from a Grounded Theory PhD: Experiences and Lessons Learnt. In *ACM CHI* (Vancouver, Canada, 2011).

[9] Inglesant, P., and Sasse, A. The True Cost of Unusable Password Policies: Password Use in the Wild. In *ACM CHI* (Atlanta, USA, 2010).

[10] Lathia, N., Hailes, S., and Capra, L. Temporal Defenses for Robust Recommendations. In *ECML/PKDD Workshop on Privacy and Security Issues in Data Mining and Machine Learning* (Barcelona, Spain, September 2010).

[11] Lathia, N., Smith, C., Froehlich, J., and Capra, L. Individuals among commuters: Building Personalised Transport Information Services from Fare Collection Systems. *Pervasive and Mobile Computing* (2012).

[12] Morrison, A., McMillan, D., Reeves, S., Sherwood, S., and Chalmers, M. A Hybrid Mass Participation Approach to Mobile Software Trials. In *ACM CHI* (Austin, USA, 2012).

[13] Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. An Empirical Study of Geographic User Activity Patterns in Foursquare. In *AAAI ICWSM* (Barcelona, Spain, 2011).

[14] Phelan, O., McCarthy, K., and Smyth, B. Using Twitter to Recommend Real-Time Topical News. In *ACM RecSys* (New York, USA, 2009).

[15] Quercia, D., Seaghdha, D., and Crowcroft, J. Talk of the City: Our Tweets, Our Community Happiness. In *AAAI ICWSM* (Dublin, Ireland, 2012).

[16] Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., and Srivastava, M. Using Mobile Phones to Determine Transportation Modes. *ACM Transactions on Sensor Networks 6*, 13 (February 2010).

[17] Reddy, S., Shilton, K., Denisov, G., Cenizal, C., Estrin, D., and Srivastava, M. Biketastic: Sensing and Mapping for Better Biking. In *ACM CHI* (Atlanta, USA, 2010).

[18] Sarwar, B. M., Konstan, J., Borchers, A., Herlocker, J., Miller, B., and Riedl, J. Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System. In *ACM CSCW* (Seattle, USA, 1998).

[19] Thomas, G., and James, D. Reinventing grounded theory: some questions about theory, ground and discovery. *British Educational Research Journal 32*, 6 (2006), 767–795.

[20] Zimmerman, J., Tomasic, A., Garrod, C., Yoo, D., Hiruncharoenvate, C., Aziz, R., Thiruvengadam, N., Huang, Y., and Steinfeld, A. Field Trial of Tiramisu: Crowd-Sourcing Bus Arrival Times to Spur Co-Design. In *ACM CHI* (Vancouver, Canada, 2011).