Hotel TULIP Web Server Data Analysis

Assignment 2 - SIT742 Modern Data Science

Alex Cummaudo <ca@deakin.edu.au>

Jake Renzella <jake.renzella@deakin.edu.au>

Deakin Software and Technology Innovation Laboratory

School of Information Technology

Deakin University, Australia

May 21, 2017

Executive Summary

This document reports session usage analysis on the Hotel TULIP website by investigating the frequency patterns of how people visited the website in a typical session. Following findings of Assignment 1, we juxtaposed session browsing between the following categories: (1) internal visitors (i.e., guests) of the website; (2) external visitors; (3) visitors from Hong Kong, the United States and Australia, being the top three countries identified who visit the website most; (4) differentiating usage patterns between PCs, Smartphone devices, Tablet Devices; and (5) how bots crawled the website. We developed a visual representation of these identified patterns and highlight areas where people are most likely to navigate between on the Hotel TULIP website to find more information about the hotel in differing fashions. Analysis was performed using the FPGrowth algorithm implementing using Apache Spark on the Databricks platform. Further details on the extraction implementation are attached and an interactive version of this file can be found on Databricks.

Contents

List of Figures

List of Tables

1 Key Findings

A list of key findings in the analysis are as thus:

- Internal visitors in the website tended to most frequently browse the ”About the Hotel” page, then click into rooms and the read about the hotel’s offers, though there is strong correlation between these visitors going from dining page into the offers page,
- external visitors, however, most frequently viewed the ”Above and Beyond” page first, then visited the ”Facilities” page before reading about the hotel’s rooms, though some also went from ”Above and Beyond” directly to the offers and dining in the hotel,
- visitors from Hong Kong usually followed similar patterns to those outlined above but in the inverse order of frequency (visited offers and dining more frequently),
- American visitors usually started by viewing the hotel’s rooms, then the offers made by the hotel, and lastly viewing ”About the Hotel” itself.
- Australian visitors have a further broader visiting pattern, but most frequently viewed the ”Above and Beyond” page first, then read about the hotel’s facilities, and lastly reading about the hotel’s Rooms,
- PC visitors most followed a frequency pattern similar to that of Australian visitors,
- Smartphone visitors had narrow visiting patterns, but started browsing by viewing about the hotel’s offers, then about the facilities of the hotel and lastly about the dining experience at the hotel,
- Tablet visitors most frequently read first about the facilities of the hotel, then about the hotel’s offers and lastly about the dining experiences,
- Bots crawled the website in a completely different fashion, by most frequently crawling between the ”Our City” page, then about the ”Location and Contacts” and lastly the ”About The Hotel” page.

2 Introduction

Browsing patterns on the Hotel TULIP Weblogs were assessed in order to gain insight on how customers navigate through the website within a typical *session*. A user session is defined as a typical visit to the website, and a collation of all the different *informational resources* that were accessed. Information resources refer to web pages that contain primary content about the hotel and its facilities, rather than multimedia and technical-related resources.

In order to assess how *different* customers do so, we extract different information based on the web log data format as prescribed in Appendix B. We contrast those users who make requests:

- internally, such as guests using the internet within the hotel's network,
- externally, such as prospective guests browsing the website for a potential stay in the hotel,
- from users within the top three countries that visit the website (refer to Assignment 1), and
- between users on PCs, Smartphones, Tablets and Bots.

Each of these criteria were analysed against matching sessions that satisfy such criteria. Data mined using the Frequency Pattern was done so using the FPGrowth Algorithm in Apache Spark.

3 Dataset

Hotel TULIP's web server runs Microsoft Internet Information Services (IIS) Server 7.5, and the attributes of this dataset as well as the relevant data dictionary can be found in Appendix B.3.

4 Method

4.1 Assumptions

4.1.1 User Sessions

A *web session* is defined as a single session of a particular user at a given moment of time. It consists of what pages were visited within that period of time, as determined by the requests made to the server in this period.

To extrapolate meaningful data, the definition of *what* a user session was needed to be determined. This is because we need to find patterns made in one particular user session; when a user visits the website in one sitting, what pages do they visit and in what order?

To do this, we assume that a session is made up of the following factors within a series of web requests:

1. The client's IP address must be the same,
2. The client's user agent string must be the same,
3. The timestamp of the session is grouped in the same day, and
4. The timestamp of the session is within the same hour.

We group every request using these four key factors, using a hash as a delimiter, into the field `session_identifier`. Using this identifier, we can group up a series of requests into one particular session. For example, a sample request is made to the server:

- The `c_ip` field is `1.2.3.4`,
- The `user_agent` field is `Safari`,
- The `timestamp` field is `Tue Feb 29 03:18:00 GMT 2017`

Therefore the `session_identifier` would be:

`1.2.3.4#Safari#29-02-2017#03`

Hence, multiple requests made within the 3rd hour of the 29th of February 2017 by the IP address `1.2.3.4` with the user agent `Safari` will be gathered together. Obviously this example is trivial and is more detailed in practice, but as an explanation it helps to be simple.

Advantages of this approach is that there can be multiple requests made from a single IP address that *are not* from the same user agent. For example, internal requests made by users within the hotel would use different computers, and therefore different `user_agent` strings would be made, making the session identifier unique.

Disadvantages of this approach are that there could be two users on an internal network (i.e., same `client_ip`) using the *exact* same user agent within the same exact hour who are technically

not the same person. Additionally, by grouping sessions by the hour, any session that was near the hour may be split into two sessions. These are some limitations of our approach.

4.1.2 Selection of Resources

Not all requests were included in the analysis, as some requests were simply resource requests. We want to focus purely on *informational resources*, and therefore need to filter the number of requests to those that are relevant.

To do this, we consider the following:

1. The request resource is non-multimedia or functional, but informational. That is, requests whose resources are not JavaScript, Cascading Style Sheets, Images and the like, but rather ASPX and ASHX files (ASP.NET and Generic Web Handler pages that contain information about the hotel).

2. The request resource is not under the directories `media`, `layouts`, or `sitecore`, as these directories do not contain informational resources.

3. The request returns a client status that is only 200 (Success), or is not a placeholder page for a status page, e.g., a `404.aspx` page would not be appropriate to consider.

Additionally, we make all request resources consistent for comparison by making them all lower case, which is made visible in the diagrams shown in Section 5.

4.2 Extraction Process

4.2.1 Data Mapping

Following the process made in Assignment 1, all publicly known client IP addresses were extracted from the MaxMind GeoIP2[1] dataset to determine request locations. This therefore allowed us to retrieve sessions from the top three countries, that being: (1) Hong Kong; (2) the USA; and (3) Australia. Determination that these were the top three countries can be found in Assignment 1.

Additionally, user agent strings were parsed to analyse device and browser statistics using the Python `user-agents` library[2] These user agent strings allowed determination of whether or not

[1] See `http://dev.maxmind.com/geoip/geoip2/`.

[2] See `https://pypi.python.org/pypi/user-agents`.

users were using: (1) PCs; (2) Smartphones; (3) Tablets; or (4) if they were actually Bots crawling the website.

4.2.2 IP Address Source Regular Expression

To differentiate between private site visitors and external visitors, a regular expression was used to filter the private and public IP address ranges. The regular expression is shown below:

```
WHERE l.c_ip REGEXP
        '(^127\.)|(^10\.)|
        (^172\.1[6-9]\.)|
        (^172\.2[0-9]\.)|
        (^172\.3[0-1]\.)|
        (^192\.168\.)'
```

Negating this `WHERE` clause of the regular expression will select only public IP addresses.

4.3 Data Mining

Data analysis was gathered using Frequent Pattern mining of user sessions. The Apache Spark implementation (Wendell, 2017) uses FPGrowth, an algorithm described by Han et al. (2000) that calculates item frequencies to identify frequent items in a given set of data. The provided data in our case was the extracted data from our SQL queries which were developed inside the attached Python Notebook.

The Apache Spark implementation requires the data parameter `minSupport`, which is the the threshold for an itemset to be identified as frequent. The default support level chosen was 0.01 (i.e., if a frequent item appears 10 times out of 1000 transactions, then its support is 0.01).

However, this default support value was massaged and relaxed in some instances where few patterns could be identified. These cases are: Hong Kong, Smartphone and Tablet requests, where the support level was relaxed by 50%; and Australian requests, where the support level was relaxed by 75%.

Additionally, we wish to retrieve sessions that contain *three or more* requests in the session. This is to ensure there are multiple hits in the session, and to retrieve more meaningful patterns.

5 Results

In our results, we have visualised the frequency patterns of users via the use of directional network graphs. In these graphs, we are able to visualise the frequency patterns of how people made requests to the website given the assumptions and extraction methods made in Section 4.

Within each graph, a *sequence* is identified as a series of multiple clicks (edges) between pages (nodes). Each sequence is coloured using the same edge colour. This sequence is also identified using a number, which is drawn on the edge label. The frequency of this pattern for the particular sequence identified is given after the forward slash on the label.

For example, a sample directional network graph, a subset of the PC requests, is given in Figure 1. This data is also presented in tabular format as Table 1.

Here we can interpret that the graph has four key sequences, as differentiated by the sequence numbers. Sequence numbers are ordered by decreasing frequency; the higher the sequence number the increased likelihood of the pattern occurring. In this graph, we see that users of PCs are most likely to move between pages in the following order:

1. from the ‘Above and Beyond’ page to the ‘Dining’ page (frequency of 293),
2. from the ‘Facilities’ page to the ‘Dining’ page (Frequency of 288),
3. from the ‘Above and Beyond’ page to the ‘Offers’ page (frequency of 286), and, with equal frequency,
4. from the ‘Facilities’ page to the ‘Offers’ page.

Table 1: Sample Frequency Graph

Sequence	From		To	Frequency
2	above and beyond	→	rooms	293
2	rooms	→	dining	293
3	facilities	→	rooms	288
3	rooms	→	dining	288
4	above and beyond	→	rooms	286
4	rooms	→	offers	286
5	facilities	→	rooms	286
5	rooms	→	offers	286

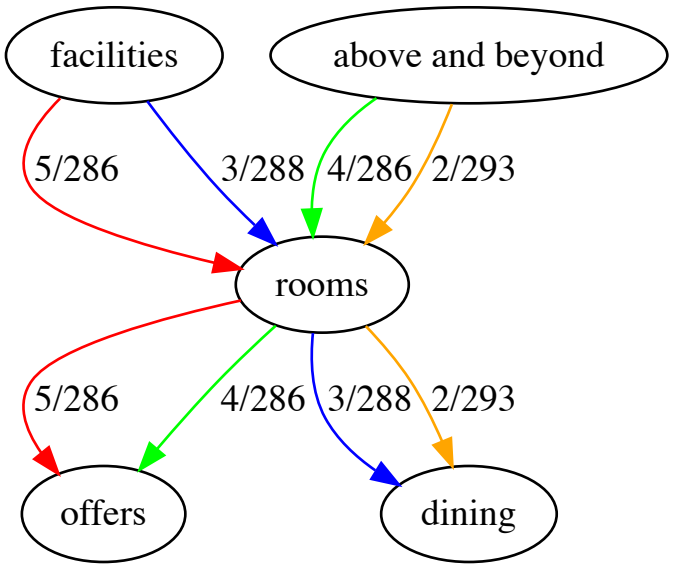

Figure 1: Sample frequency patterns identified in a subset of PC requests. Refer to Table 1 for frequency pattern interactions.

5.1 IP Request Sources

5.1.1 Internal Site Visitors

Refer to Figure 2.

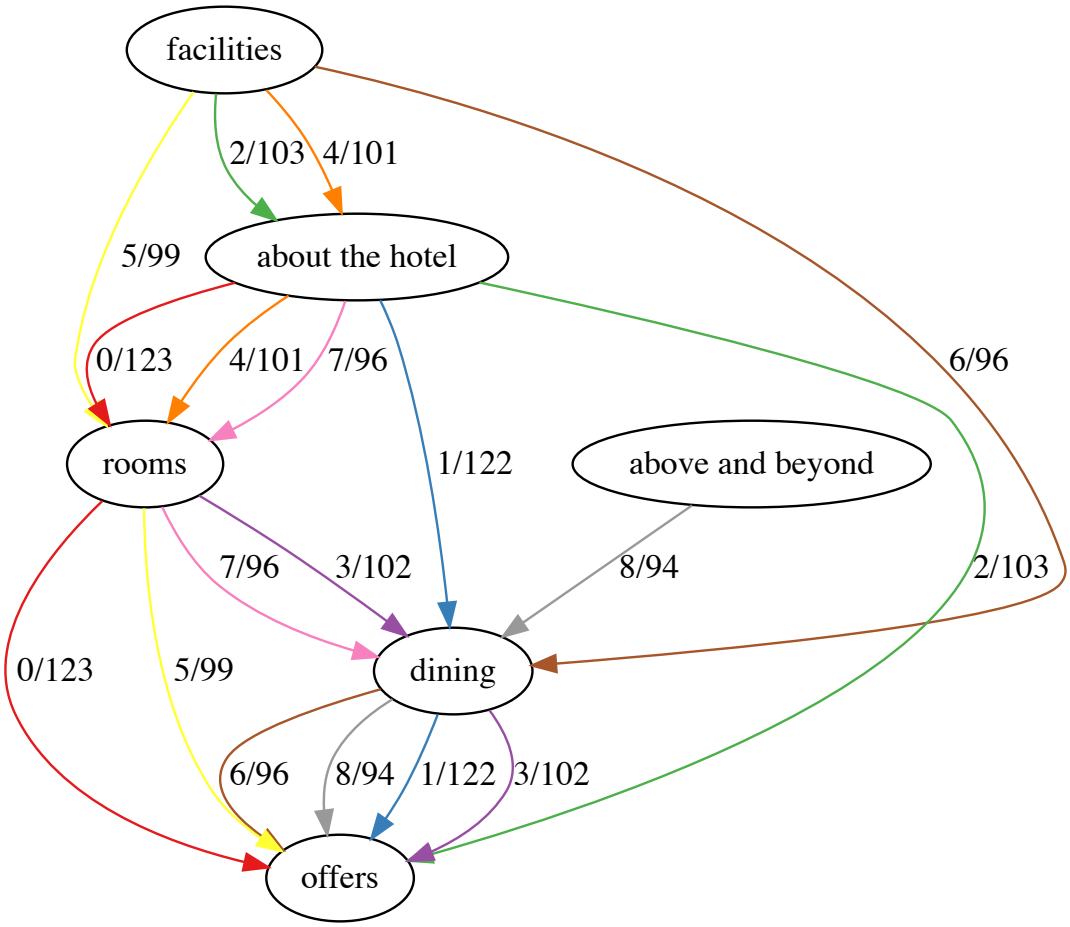

Figure 2: Directional network graph visualising frequency patterns of internal visitors made on an internal IP range. Refer to Table 2 for frequency pattern interactions.

5.1.2 External Site Visitors

Refer to Figure 3.

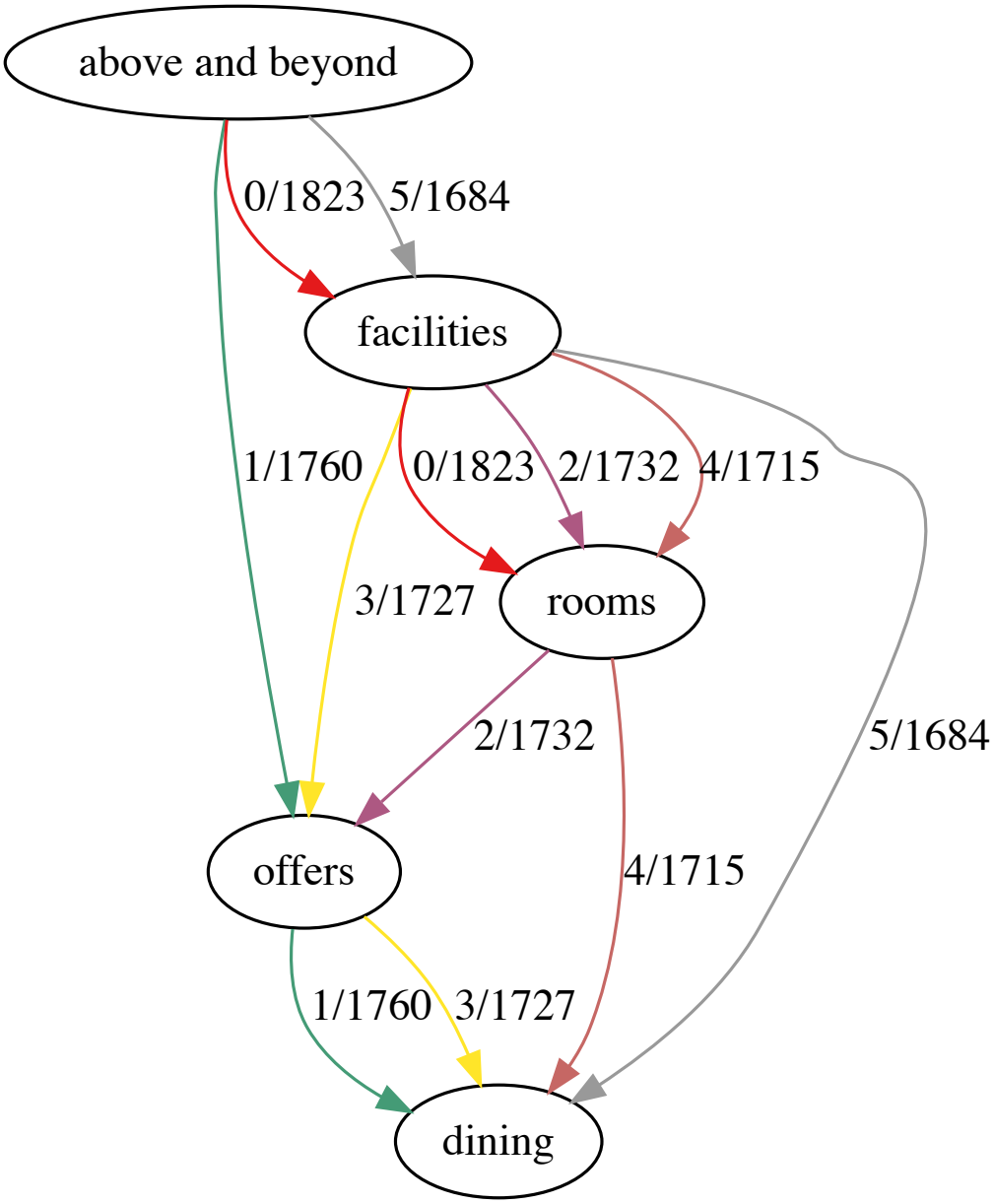

Figure 3: Directional network graph visualising frequency patterns of external visitors made on a non-internal IP range. Refer to Table 3 for frequency pattern interactions.

5.2 Top Three Countries

5.2.1 Hong Kong Visitors

Refer to Figure 4.

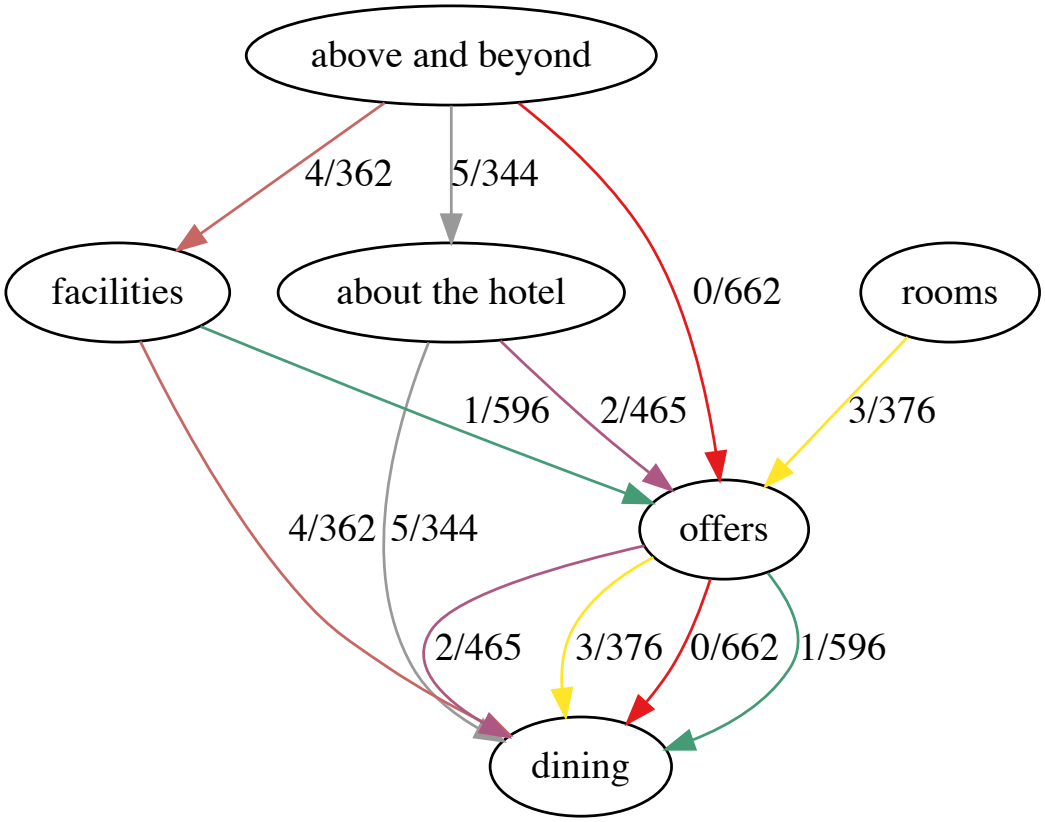

Figure 4: Directional network graph visualising frequency patterns of visitors from Hong Kong. Refer to Table 4 for frequency pattern interactions.

5.2.2 USA Visitors

Refer to Figure 5.

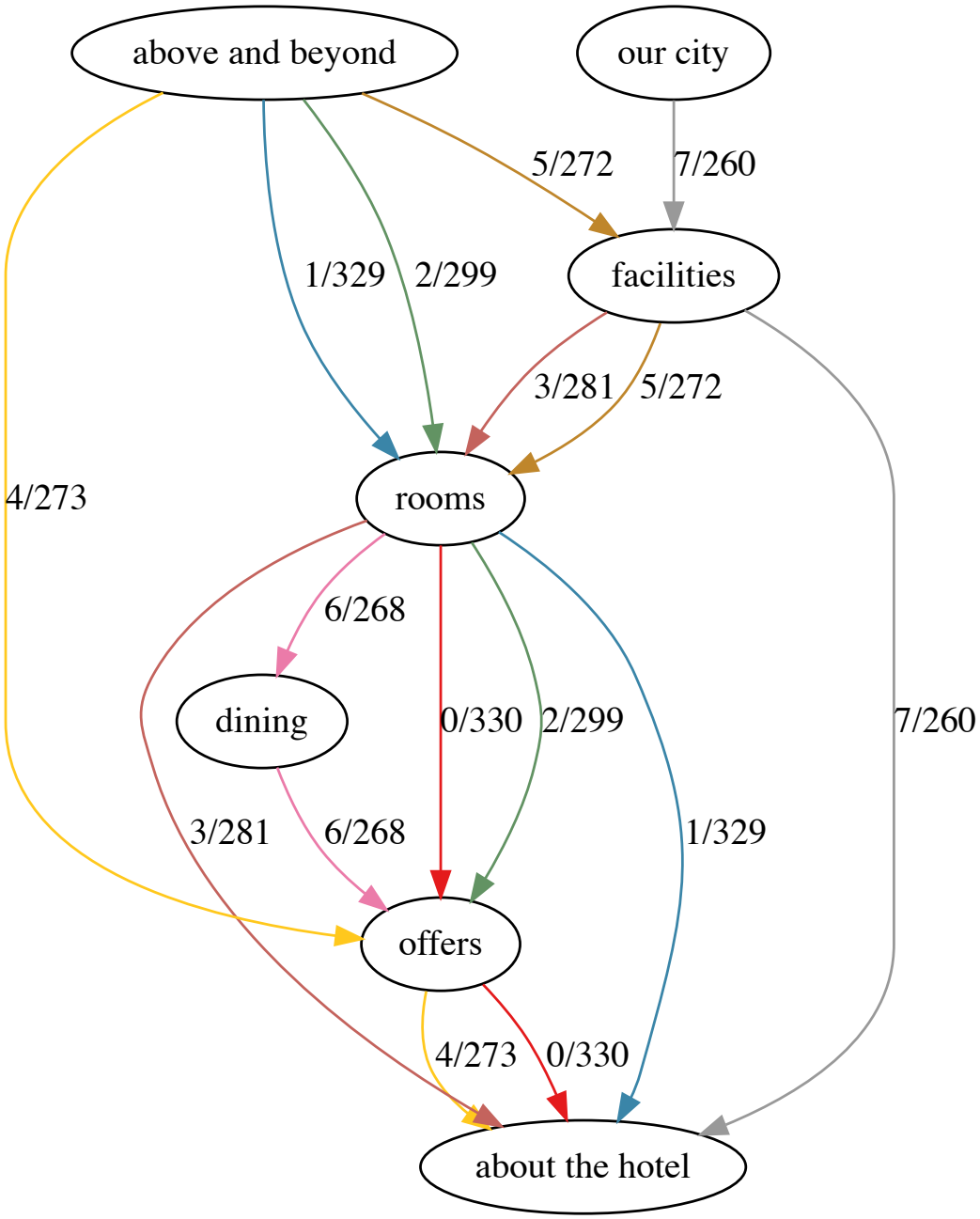

Figure 5: Directional network graph visualising frequency patterns of visitors from the United States. Refer to Table 5 for frequency pattern interactions.

5.2.3 Australian Visitors

Refer to Figure 6.

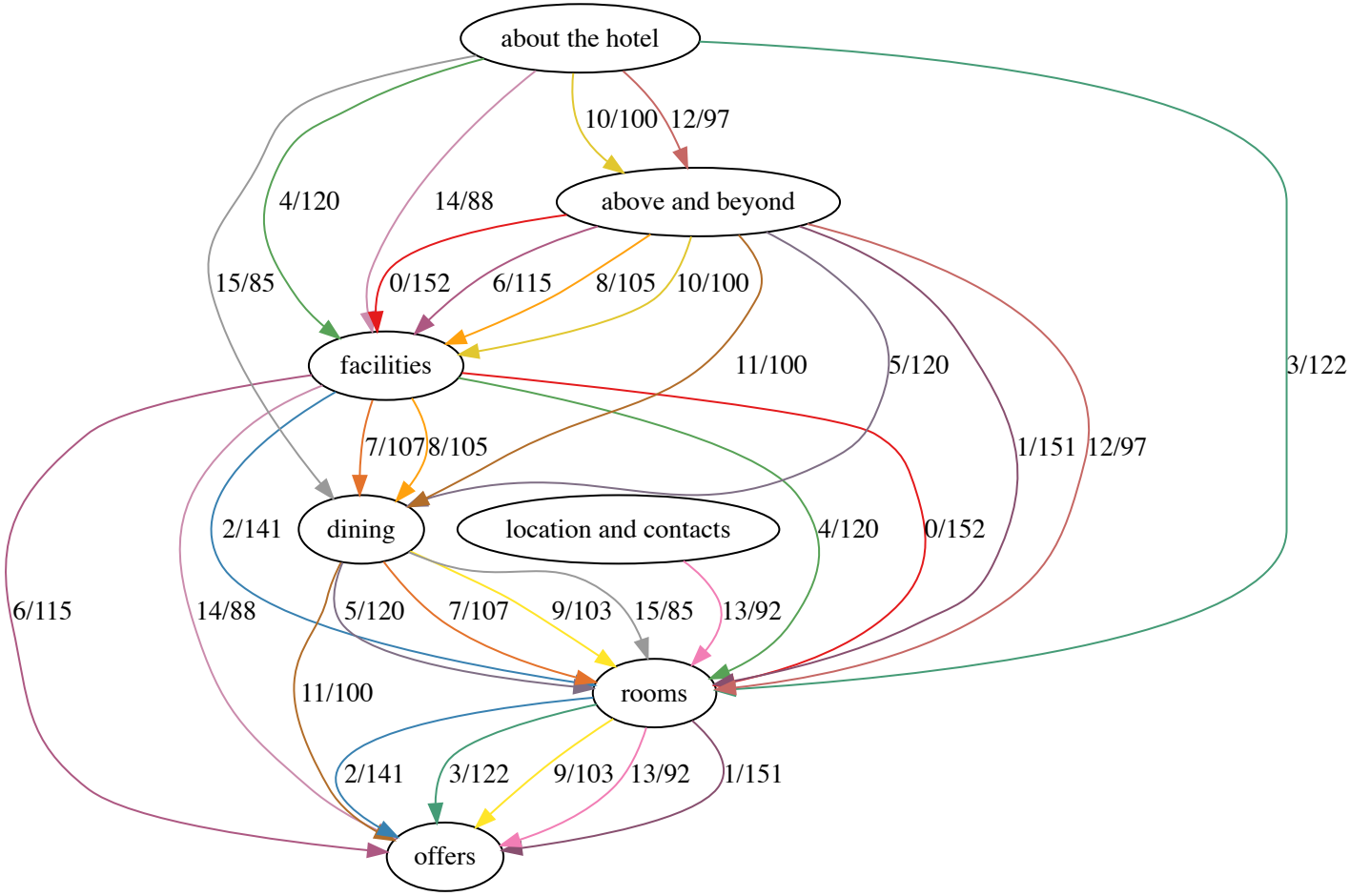

Figure 6: Directional network graph visualising frequency patterns of visitors from Australia. Refer to Table 6 for frequency pattern interactions.

5.3 Platform Categories

5.3.1 PC Visitors

Refer to Figure 7.

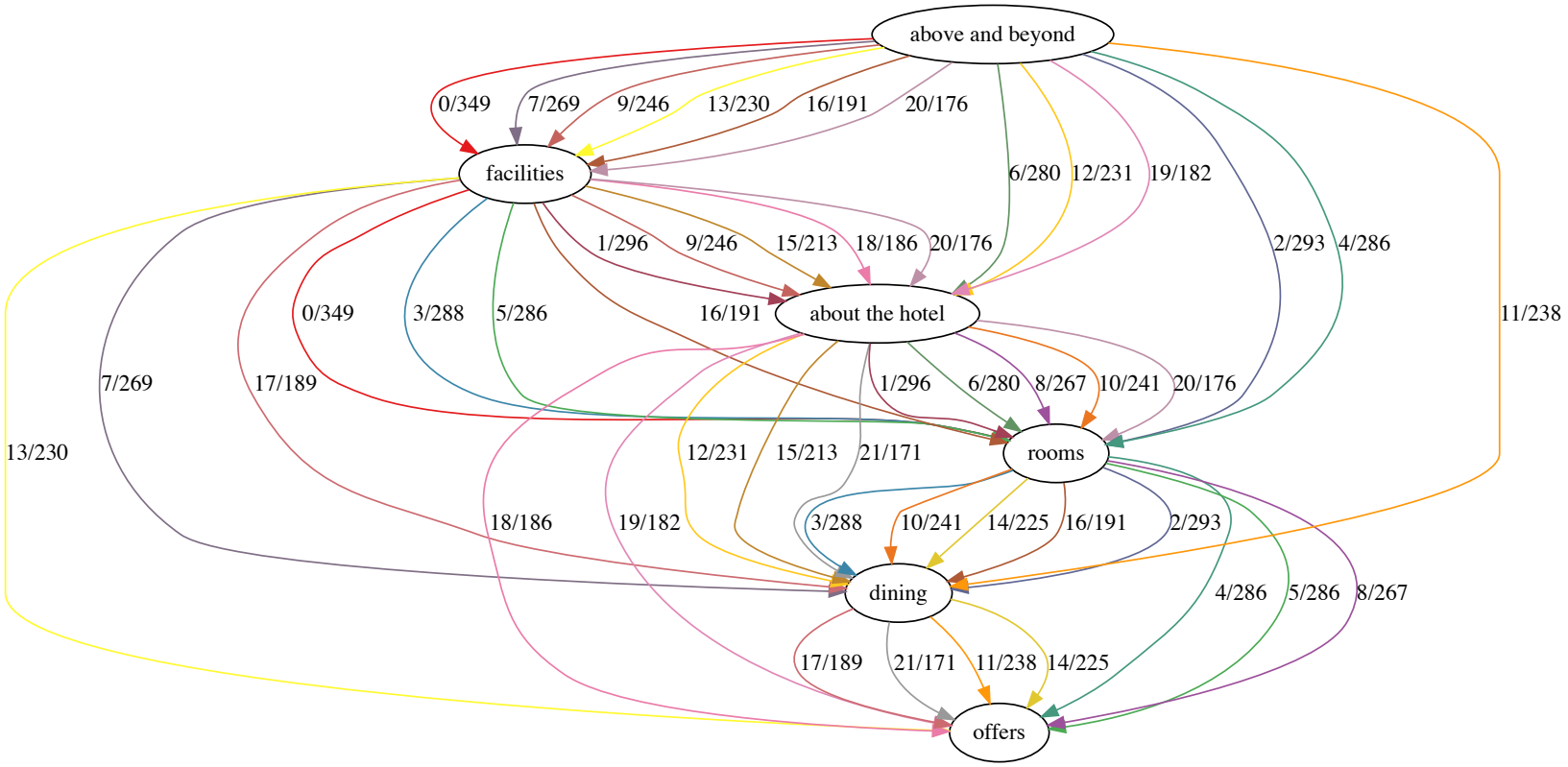

Figure 7: Directional network graph visualising frequency patterns of requests made by PCs. Refer to Table 7 for frequency pattern interactions.

5.3.2 Smartphone Visitors

Refer to Figure 8.

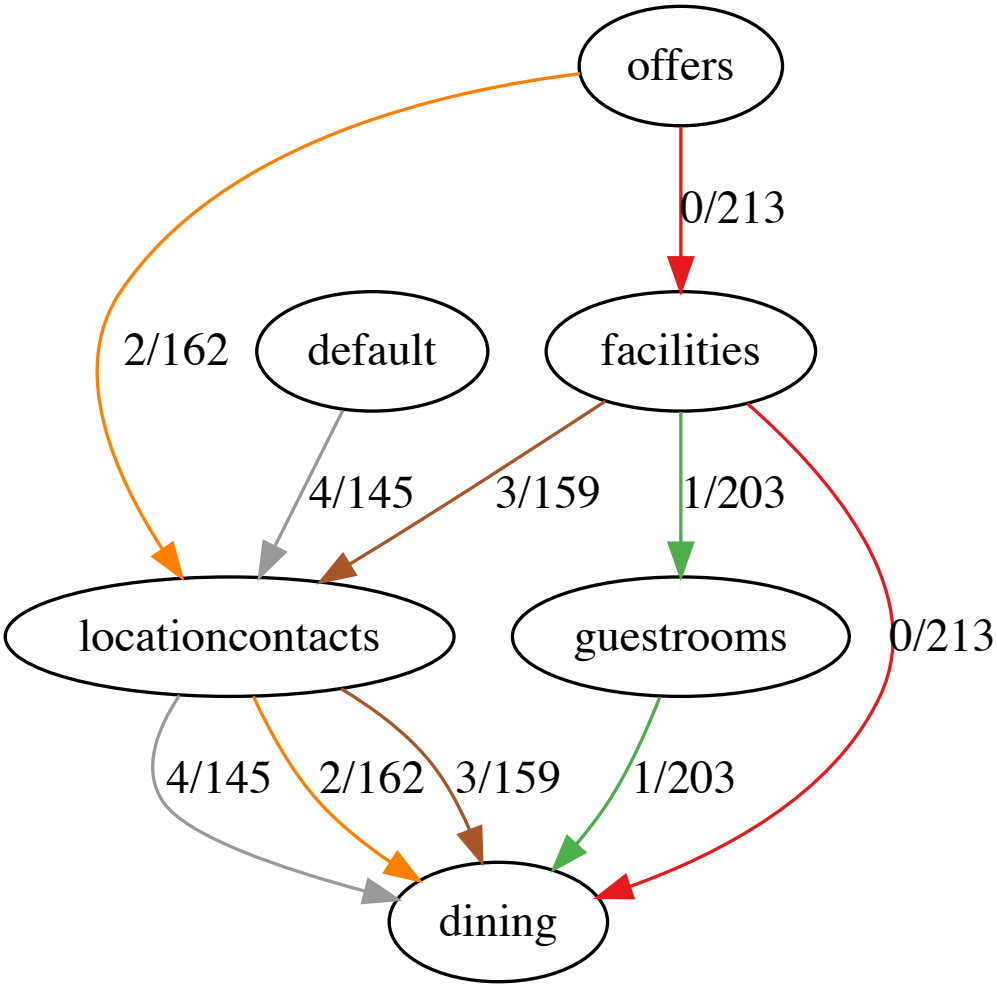

Figure 8: Directional network graph visualising frequency patterns of requests made on smartphones. Refer to Table 9 for frequency pattern interactions.

5.3.3 Tablet Visitors

Refer to Figure 9.

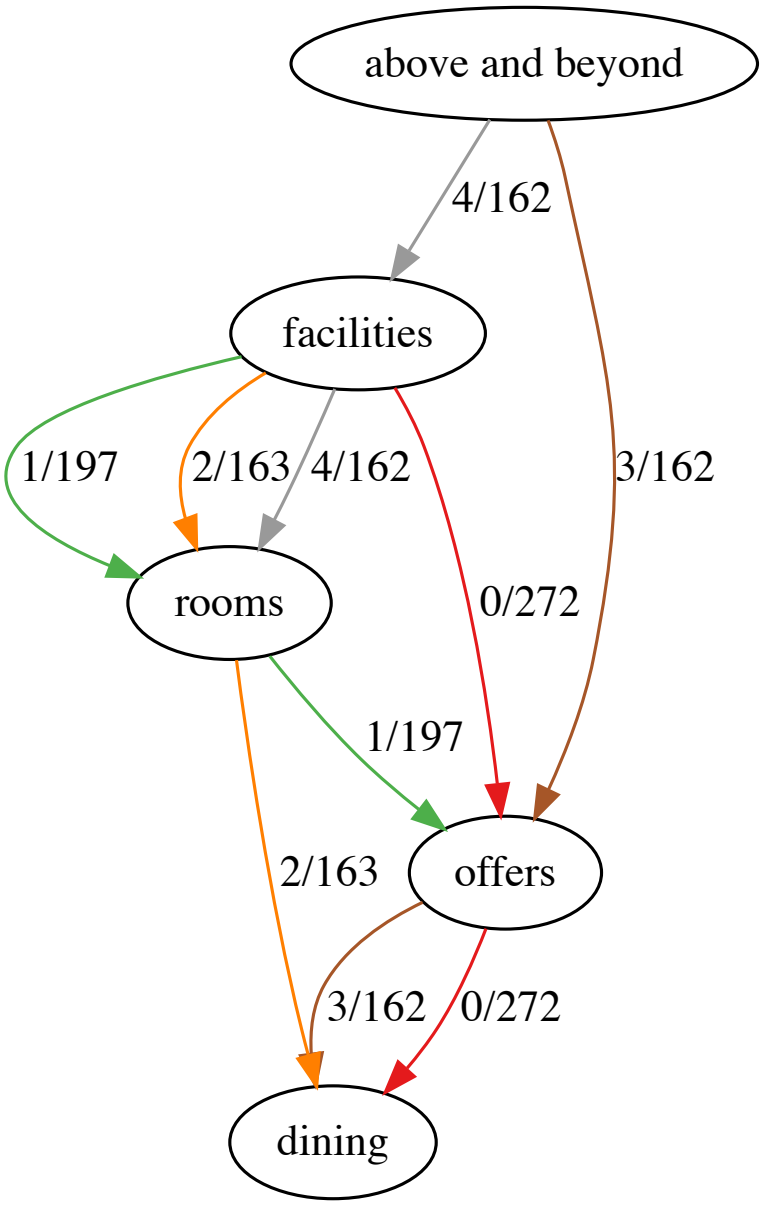

Figure 9: Directional network graph visualising frequency patterns of requests made on tablets. Refer to Table 8 for frequency pattern interactions.

5.3.4 Bots Visitors

Refer to Figure 10.

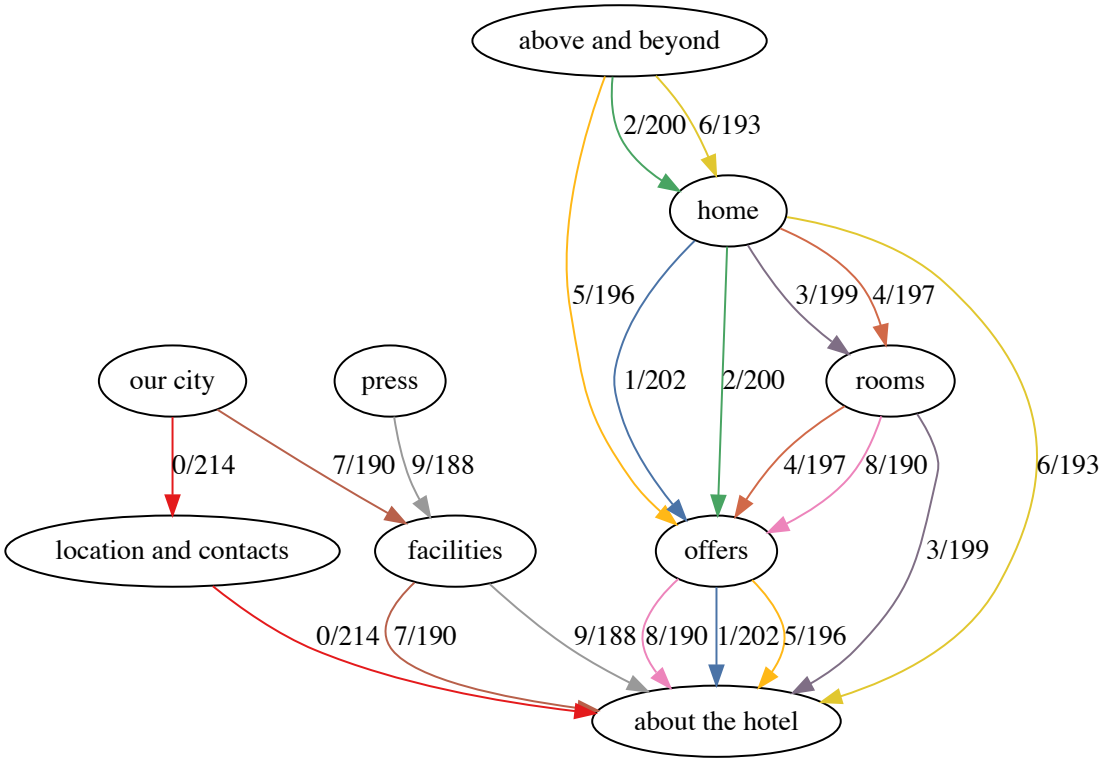

Figure 10: Directional network graph visualising frequency patterns of requests made by bots. Refer to Table 10 for frequency pattern interactions.

References

Fielding, R., J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee (1999). RFC 2616, Hypertext Transfer Protocol – HTTP/1.1. Retrieved 11 May 2017, <http://www.rfc.net/rfc2616.html>.

Hallam-Baker, P. and B. Behlendorf (1998). Extended Log File Format. Retrieved 11 May 2017, <https://www.w3.org/TR/WD-logfile-960221.html>.

Han, J., J. Pei, and Y. Yin (2000, May). Mining frequent patterns without candidate generation. *SIGMOD Rec. 29*(2), 1–12.

Microsoft Corporation (2003). W3C Extended Log File Format (IIS 6.0). Retrieved 11 May 2017, <https://www.microsoft.com/technet/prodtechnol/WindowsServer2003/Library/IIS/676400bc-8969-4aa7-851a-9319490a9bbb.mspx?mfr=true>.

Microsoft Corporation (2017). Description of Microsoft Internet Information Services (IIS) 5.0 and 6.0 status codes. Retrieved 11 May 2017, <https://support.microsoft.com/en-au/help/318380/description-of-microsoft-internet-information-services-iis-5.0-and-6.0-status-codes>.

Wendell, P. (2017). Frequent Pattern Mining - RDD-based API. Retrieved 11 May 2017, <https://people.apache.org/~pwendell/spark-nightly/spark-master-docs/latest/mllib-frequent-pattern-mining.html>.

Wolf, M. and C. Wicksteed. Date and Time Formats. https://www.w3.org/TR/NOTE-datetime.

A Additional Tables

Below are tables of frequency pattern results for each section identified in Section 5.

Table 2: Internal Request Frequency Patterns

Sequence	From		To	Frequency
0	about the hotel	→	rooms	123
0	rooms	→	offers	123
1	about the hotel	→	dining	122
1	dining	→	offers	122
2	facilities	→	about the hotel	103
2	about the hotel	→	offers	103
3	rooms	→	dining	102
3	dining	→	offers	102
4	facilities	→	about the hotel	101
4	about the hotel	→	rooms	101
5	facilities	→	rooms	99
5	rooms	→	offers	99
6	facilities	→	dining	96
6	dining	→	offers	96
7	about the hotel	→	rooms	96
7	rooms	→	dining	96
8	above and beyond	→	dining	94
8	dining	→	offers	94

Table 3: External Request Frequency Patterns

Sequence	From		To	Frequency
0	above and beyond	→	facilities	1823

Continued on next page...

Table 3 (*continued from Page 32*): External Request Frequency Patterns

Sequence	From		To	Frequency
0	facilities	→	rooms	1823
1	above and beyond	→	offers	1760
1	offers	→	dining	1760
2	facilities	→	rooms	1732
2	rooms	→	offers	1732
3	facilities	→	offers	1727
3	offers	→	dining	1727
4	facilities	→	rooms	1715
4	rooms	→	dining	1715
5	above and beyond	→	facilities	1684
5	facilities	→	dining	1684

Table 4: Hong Kong Frequency Patterns

Sequence	From		To	Frequency
0	above and beyond	→	offers	662
0	offers	→	dining	662
1	facilities	→	offers	596
1	offers	→	dining	596
2	about the hotel	→	offers	465
2	offers	→	dining	465
3	rooms	→	offers	376
3	offers	→	dining	376
4	above and beyond	→	facilities	362
4	facilities	→	dining	362

Continued on next page...

Table 4 (*continued from Page 33*): Hong Kong Frequency Patterns

Sequence	From		To	Frequency
5	above and beyond	→	about the hotel	344
5	about the hotel	→	dining	344

Table 5: USA Request Frequency Patterns

Sequence	From		To	Frequency
0	rooms	→	offers	330
0	offers	→	about the hotel	330
1	above and beyond	→	rooms	329
1	rooms	→	about the hotel	329
2	above and beyond	→	rooms	299
2	rooms	→	offers	299
3	facilities	→	rooms	281
3	rooms	→	about the hotel	281
4	above and beyond	→	offers	273
4	offers	→	about the hotel	273
5	above and beyond	→	facilities	272
5	facilities	→	rooms	272
6	rooms	→	dining	268
6	dining	→	offers	268
7	our city	→	facilities	260
7	facilities	→	about the hotel	260

Table 6: Australian Request Frequency Patterns

Sequence	From		To	Frequency
0	above and beyond	→	facilities	152
0	facilities	→	rooms	152
1	above and beyond	→	rooms	151
1	rooms	→	offers	151
2	facilities	→	rooms	141
2	rooms	→	offers	141
3	about the hotel	→	rooms	122
3	rooms	→	offers	122
4	about the hotel	→	facilities	120
4	facilities	→	rooms	120
5	above and beyond	→	dining	120
5	dining	→	rooms	120
6	above and beyond	→	facilities	115
6	facilities	→	offers	115
7	facilities	→	dining	107
7	dining	→	rooms	107
8	above and beyond	→	facilities	105
8	facilities	→	dining	105
9	dining	→	rooms	103
9	rooms	→	offers	103
10	about the hotel	→	above and beyond	100
10	above and beyond	→	facilities	100
11	above and beyond	→	dining	100
11	dining	→	offers	100
12	about the hotel	→	above and beyond	97
12	above and beyond	→	rooms	97

Continued on next page...

Table 6 (*continued from Page 35*): Australian Request Frequency Patterns

Sequence	From		To	Frequency
13	location and contacts	→	rooms	92
13	rooms	→	offers	92
14	about the hotel	→	facilities	88
14	facilities	→	offers	88
15	about the hotel	→	dining	85
15	dining	→	rooms	85

Table 7: PC Request Frequency Patterns

Sequence	From		To	Frequency
0	above and beyond	→	facilities	349
0	facilities	→	rooms	349
1	facilities	→	about the hotel	296
1	about the hotel	→	rooms	296
2	above and beyond	→	rooms	293
2	rooms	→	dining	293
3	facilities	→	rooms	288
3	rooms	→	dining	288
4	above and beyond	→	rooms	286
4	rooms	→	offers	286
5	facilities	→	rooms	286
5	rooms	→	offers	286
6	above and beyond	→	about the hotel	280
6	about the hotel	→	rooms	280
7	above and beyond	→	facilities	269

Continued on next page...

Table 7 (*continued from Page 36*): PC Request Frequency Patterns

Sequence	From		To	Frequency
7	facilities	→	dining	269
8	about the hotel	→	rooms	267
8	rooms	→	offers	267
9	above and beyond	→	facilities	246
9	facilities	→	about the hotel	246
10	about the hotel	→	rooms	241
10	rooms	→	dining	241
11	above and beyond	→	dining	238
11	dining	→	offers	238
12	above and beyond	→	about the hotel	231
12	about the hotel	→	dining	231
13	above and beyond	→	facilities	230
13	facilities	→	offers	230
14	rooms	→	dining	225
14	dining	→	offers	225
15	facilities	→	about the hotel	213
15	about the hotel	→	dining	213
16	above and beyond	→	facilities	191
16	facilities	→	rooms	191
16	rooms	→	dining	191
17	facilities	→	dining	189
17	dining	→	offers	189
18	facilities	→	about the hotel	186
18	about the hotel	→	offers	186
19	above and beyond	→	about the hotel	182

Continued on next page...

Table 7 (*continued from Page 36*): PC Request Frequency Patterns

Sequence	From		To	Frequency
19	about the hotel	→	offers	182
20	above and beyond	→	facilities	176
20	facilities	→	about the hotel	176
20	about the hotel	→	rooms	176
21	about the hotel	→	dining	171
21	dining	→	offers	171

Table 8: Tablet Request Frequency Patterns

Sequence	From		To	Frequency
0	facilities	→	offers	272
0	offers	→	dining	272
1	facilities	→	rooms	197
1	rooms	→	offers	197
2	facilities	→	rooms	163
2	rooms	→	dining	163
3	above and beyond	→	offers	162
3	offers	→	dining	162
4	above and beyond	→	facilities	162
4	facilities	→	rooms	162

Table 9: Smartphone Request Frequency Patterns

Sequence	From		To	Frequency
0	offers	→	facilities	213
0	facilities	→	dining	213
1	facilities	→	guestrooms	203
1	guestrooms	→	dining	203
2	offers	→	locationcontacts	162
2	locationcontacts	→	dining	162
3	facilities	→	locationcontacts	159
3	locationcontacts	→	dining	159
4	default	→	locationcontacts	145
4	locationcontacts	→	dining	145

Table 10: Bots Request Frequency Patterns

Sequence	From		To	Frequency
0	our city	→	location and contacts	214
0	location and contacts	→	about the hotel	214
1	home	→	offers	202
1	offers	→	about the hotel	202
2	above and beyond	→	home	200
2	home	→	offers	200
3	home	→	rooms	199
3	rooms	→	about the hotel	199
4	home	→	rooms	197
4	rooms	→	offers	197
5	above and beyond	→	offers	196
5	offers	→	about the hotel	196

Continued on next page...

Table 10 (*continued from Page 39*): Bots Request Frequency Patterns

Sequence	From		To	Frequency
6	above and beyond	→	home	193
6	home	→	about the hotel	193
7	our city	→	facilities	190
7	facilities	→	about the hotel	190
8	rooms	→	offers	190
8	offers	→	about the hotel	190
9	press	→	facilities	188
9	facilities	→	about the hotel	188

B Data Dictionary

B.1 Dataset Description

Table 11: Dataset Description

Key	Entry
Name	Hotel TULIP Web Log Dataset
Size	17.06 GB (954.7 MB compressed)
Release Date	30/4/17
Attributes	14
No. Records	73,368,256
Provider	Dr Beer Guts, CIO, Hotel TULIP (Information Technology Division)[3]
Privacy	Confidential[4]

B.2 Contact Information

Table 12: Dataset Contact Information

Key	Entry
Prepared by	Team-SIT742
Point of Contact	Alex Cummaudo <ca@deakin.edu.au>
	Jake Renzella <jake.renzella@deakin.edu.au>
Team Members	Alex Cummaudo <ca@deakin.edu.au>
	Jake Renzella <jake.renzella@deakin.edu.au>

[3]Download URL: https://d2l.deakin.edu.au/d2l/le/content/520519/topics/files/download/3482057/DirectFileTopicDownload,https://d2l.deakin.edu.au/d2l/le/content/520519/viewContent/3482057/View?ou=520519.

[4]Exclusively available for educational purposes only for the Deakin University unit SIT742. Redistribution is prohibited.

B.3 Data Dictionary

Table 13: Data Dictionary

Attribute Name	Data Type	Data Subtype	Description	Examples	Notes
date	MC	DATE - Date	Date: Date when request occurred	2014-08-01	UTC time zone; ISO 8601 date format (YYYY-MM-DD)
time	MC	DATE - Time	Time: Time when request occurred	09:51:23	UTC time zone; ISO 8601 24-hr time format (hh:mm:ss)
s-ip	CN	ADDR - Address - IP Address	Server IP Address: IP of the server generating the log responding to the request	10.130.0.12	N/A
cs-uri-stem	CN	URL - Uniform Resource Identifier (URI)	URI Stem: Stem portion of the full URI made in the client to server request	/sitecore	N/A
cs-uri-query	CN	URL - Uniform Resource Identifier (URI)	URI Query: Query portion of the full URI made in the client to server request	cmd=GetTreeview	May be empty; Query string is specifically matched as URI. See Hallam-Baker & Behlendorf (1998).

Continued on next page...

Table 13 (*continued from Page 42*): Data Dictionary

Attribute Name	Data Type	Data Subtype	Description	Examples	Notes
s-port	CN	ADDR - Address - Port Number	Server Port: Port for the server which request is made	80	N/A
cs-username	CN	STR - Free String	User Name: Name of authenticated user accessing the server	-	Anonymous users are represented with a hyphen.
c-ip	CN	ADDR - Address - IP Address	Client IP Address: IP of the client making the request	10.120.7.23	N/A
sc-substatus	CN	ID - Identification - Microsoft IIS Named HTTP Response Sub status Code	Protocol Sub status: The sub status error code	0	Refer to Microsoft (2017) for specific codes. A zero indicates no sub status error code.
sc-win32-status	CN	ID - Identification - Microsoft IIS Named Windows Code	Win32 Status: The Windows status code	0	Generally not applicable to non-Windows devices; defaults to zero.

Continued on next page...

Table 13 (*continued from Page 42*): Data Dictionary

Attribute Name	Data Type	Data Subtype	Description	Examples	Notes
time-taken	MC	DATE - Time - From zero	Duration: The time taken for the request to complete.	39	Measured in milliseconds.

C Extrapolation Results

Attached on the following pages are the results from Databricks. You may also interact with this online on Databricks.