Analytical Tools

Semi-Public Space Conflicts and Alliances in Primary Metropolitan Centres: Sylvia Park, Mt Wellington, Auckland

DIGITAL SPACE ANALYSIS

Angelo Bueno, Tanyalak Chalermtip and Manfredo Manfredini

[WORKING DOCUMENT GUS/SP3.1]

Contents

1 ANALYTICAL METHODS

1.1 DATA COLLECTION AND CLEANING

1.1.1 Points of Interest

1.1.1.1 Collection

URLs of Instagram posts from an identified point of interest (POI) are retrieved from the Instagram search page using the Web Scraper tool (Web Scraper is a Google Chrome browser extension built for data extraction from web pages). The page contains a grid of thumbnails which can be selected by the scraper to compile the URLs into a spreadsheet.

1.1.1.2 Cleaning

This extraction process results in duplicates due to the page having an infinite scroll, where newly loaded thumbnails and already extracted thumbnails are visible at the same time. On Excel, duplicate URLs are identified in the spreadsheet by *conditional formatting* then removed by using the *remove duplicates* function to leave only unique URLs.

1.1.2 Posts

1.1.2.1 Collection

This list of URLs is fed into a new scraper which is instructed by manually selecting page elements, forming a rule that can be applied to every post to extract usernames, commenters, link to profiles, mentions, link to the mention profiles, comments, link to images, number of likes, and post time.

(It is worth noting that, as Instagram shows up to the 25 most recent comments when it is first loaded, all posts exceeding this limit require manual access to the expanded list via the "load more comments" button to incrementally expand the comments until all comments are visible.) The resulting spreadsheet is a dataset that includes data of all relevant POIs and comments. The spreadsheet includes 11 series (columns): post URL, usernames, commenters, link to profiles, mentions, link to the mention profiles, comments, link to images, likes number, posted date, and location.

1.1.2.2 Cleaning

Irrelevant and irregular posts are identified and removed. Commercial posts and automated bots are identified through known usernames; spammers are detected by an analysis on excessive posting behaviour.

1.2 DATA PROCESSING

1.2.1 Network Analysis

1.2.1.1 Nodes and Edges

Nodes and *edges/links* are the fundamental elements of a network. They are defined by different attributes that allow a network graph to be plotted. Nodes are actors representing either people (person posting or person commenting) and posts (uploads, comments and likes). Edges/links are interactions between nodes.

While a set of attributes can be contained by an individual node, an edge describes the relationship between two nodes. Due to this difference, each element class must be treated in a separate list. The node list assigns each node an *ID, label,* and *type* (either person or post), while the edge list has a similar arrangement with type (comment, label, posts) but also includes two columns for *source nodes* (people) and *target nodes* (images). Retaining information can find the overlap between the number of people who comment, post, and like. This gives not only the total size of the network measured by the nodes and edges, but further categorises and measures each kind of activity by percentage. Furthermore, the two lists are elaborated with additional attributes to produce graphs which emphasise different aspects such as *centrality*, *community structures*, and *growth*.

1.2.1.2 Centrality

In a bimodal network such as this, *degree centrality* of either photos or images is calculated by counting the number of immediate links pointing toward the node or away from it. These two values would typically refer to the indegree and outdegree centrality. However, within a network which has two node types, these values are the general volume of interactions for each kind of node, rather than the extent to which a node gives or receives comments or likes. Therefore, to find the true indegree and outdegree centralities of people, the same process must be repeated on a graph containing only one node type.

The common solution of using a *bimodal network projection* would be unsuitable as it changes the degree of each nodes and, by extension, the indegree and outdegree centrality as well. Since attributes of each node have been associated early in the scraping process, they can be reconciled at any later steps of the process if needed. This allows all post-nodes belonging to one author to be replaced with a single person-node, so that only one node type is being used and degree and direction is unchanged. After the transformation, the act of counting links in either direction can now identify people who frequently like or comment on other people's posts, and likewise those who attract attention.

Additionally, a single-mode graph also allows the calculation of other kinds of centrality to accurately identify people who can easily reach many parts of the network (closeness centrality), can act as a bridge which connects two groups (betweenness centrality), or are highly connected to other important people. (eigenvector centrality).

These results are visualised in base graphs that a) use colour to distinguish people-nodes and post-nodes, b) show the difference in the centrality values between nodes through size variation (i.e., the higher the value, the larger the node), and c) mark the direction of links (from person to photo).

1.2.1.3 Communities

Definition: Modularity is a measure to effectively gauge the strength of community structures in complex networks containing sparse and dense areas of connections. It identifies modules (networks) composed by clusters of nodes that have relatively high internal connectivity.

Method: Networks are partitioned in communities (modules) whose cohesion is assessed based on the differences between the actual number of connections and the calculated number of expected links. Initially, the entire network is arbitrarily partitioned in a number of modules that appear to be the preeminent ones. Thereafter, each module is given a value between -1 and 1, where positive numbers indicate groups containing more links than expected.

Expected numbers of links are calculated using a *random chance* process. This process uses a *configuration model* which calculates the connectivity of a given network by iteratively transforming its configuration through splitting every edge into two stubs then reconnecting pairs randomly (even allowing self-loops), therefore maintaining the original *node degree distribution*. This allows any random graph to be generated and provides all necessary information for calculating the differences between actual and expected number of links (total number of stubs, total number of rewirings, expected number of edges between two selected nodes, and the actual number of likes between the two nodes.)

Following the *Louvain method* for community detection developed by Vincent D. Blondel et al., this analysis compares changes in modularity affected by moving each individual node into each of its neighbour's communities, until no single move leads to a positive gain in modularity. At this point, maximum modularity has been reached and another iteration of the network is created by merging the members of each community into a single node. The process is repeated iteratively, identifying and merging sets of nodes into fewer and larger groups with each additional pass. This recognises multiple hierarchical communities that are distinct (they do not share nodes with each other) but are able to decompose into smaller groups when viewed at lower resolutions (early iterations). This allows the user to zoom into community structures until a meaningful size and number has been achieved. However, since in large networks an unmanageable number of communities are often detected (even at high resolution), a filter may be applied to hide communities smaller than a specified limit to produce more readable graphs.

1.2.1.4 Network Typology

To give an understanding of the network in its integrity, the elements are interpreted at a large scale by using interactions to associate individuals in networks that can be classified according to specific typological characteristics: *broadcast network, support network, tight crowd, polarised crowd, community cluster*. These types are obtained in three steps: 1) cohesive groups of individuals are identified as single sets representing communities; 2) cohesive groups of communities are identified as larger crowds or macro-groups; and 3) network types are classified using the *Himelboim* method by an assessment of their hierarchy, interconnectivity, and number of isolates.

1.2.1.5 Dynamic Networks

At the community level, the networks dynamics are assessed by counting (and normalising) the number of likes, comments and posts within a given period (e.g. a month). This enables to recognise the general trend across the whole year, showing peaks and dips during specific moments in time. The communities are further categorised into size brackets to show movement of people shifting between communities in a network graph. Nodes are organised from left to right by time, then from top to bottom by size bracket. Line thickness describes the amount of people moving between size brackets. A similar graph is produced to show people introduced each month. This shows how over a given time series (e.g. monthly) individual nodes, representing external members who are not yet introduced, connect to large nodes of the network.

To measure growth within a community, stable groups are used as a point of reference and individuals are added to create sequence (e.g. monthly series create 12-part string per year). Eventually, the largest group which holds the same lineage within a community is identified as its core group. The people not included in this core group are identified as transient members who determine the positive or negative growth of the community over time.

1.2.2 Semantic Analysis

Semantic analysis investigates and traces the main interest of the entire Instagram users collected from a defined location through their comments. This explores textual elements in three scales of the Instagram users defined in network analysis including the full network, communities, and dynamic communities.

1.2.2.1 Identification of Topics of Interests and Keywords

The analysis begins by using a word-frequency analysis tool developed by John Walkenbach to analyse the comments and generate a word-frequency list which can be produced regardless of grammar and word order. However, people's names, meaningless words, foreign languages, punctuation, emoticons, symbols, and numbers are filtered out from the list. The remaining keywords are categorised into 10 different topics of interest including: 1) Animals, 2) Art, Design, and Photography, 3) Beauty, Sports, and Wellness, 4) Events and Entertainment, 5) Fashion and Styles, 6) Food and Drinks, 7) Places and Architecture, 8) Nature, 9) Social and People, and 10) Technology.

Interest level and significant topics of interest are determined by the number of keywords in each category. In addition, Māori keywords are especially detected and categorised to observe the interest topics and level of interest expressed through local language.

1.2.2.2 Comments Labelling for Interest-Based Networks

Keywords are used as the attributes to assign the interest labels (Art, Design, and Photography; Beauty, Sports, and Wellness; Events and Entertainment; Fashion and Styles; Food and Drinks; Places and Architecture; Nature; Animals; Social and People; and Technology) to each comment. A list of comments is labelled using an Excel formula to detect the keywords and assign the interest labels to which those

keywords belong. For example, if a comment contains “burger,” which belongs to the Food and Drinks topic, a Food and Drinks label is assigned to the comment. Each comment can be assigned more than one label as comments can be associated with more than one topic.

The interests of commenters identified during the comments labelling process allow the creation of a new nodes type which represents interest topics. A network is formed showing the connection between people and various interest topics. Moreover, a correlation matrix is calculated to find highly correlated pairs of interests, revealing topics which tend to be mentioned together in the same comments.

1.2.2.3 Tracing Timeline of Interests and Events

As the comment’s dates are unknown (only the dates for the posts tagged are available), the keywords are separated into 12 months following the dates of the posts. The number of keywords in each topic represents the interest level of those topics in each month. Furthermore, significant dates, public holidays, and events held at the site in 2017 are retrieved to observe their effects on the changes of interests’ level throughout the year.

1.2.2.4 Semantic Analysis for Static Communities

Moreover, semantic analysis analyses the comments from each community defined in network analysis to explore the differences of interests in each community, identify communities’ characteristics, and track the changes of communities’ interests in relation to events held at Sylvia Park Shopping Centre.

1.2.2.5 Interests of Each Community Size Bracket

Comments are separated into four lists for different-sized communities (large, medium, small, and isolates). This enables an analysis of interests within each community size bracket based on keywords detected in the comments list. Nevertheless, the number of categorised keywords in each community size is measured in percentages to reveal the distribution of interest topics within different community size brackets.

1.2.2.6 Interests of Large-Sized Communities

Communities identified within the large size bracket are further explored to detect their interest patterns and their dominant interest topics. External and internal and influences on the communities’ interests are explored by investigating potential causes (e.g. significant dates and events or influential (central) person’s interests).

1.2.2.7 Activities and Interests

Number of activities including posts, likes, and comments of the large communities are compared alongside the level of communities’ interests based on the number of keywords identified in each community. This reveals the overall pattern of activity and interests of large-sized communities, allowing an understanding of the general trend of the patterns as well as identification of distinguishing patterns. Furthermore, languages used in the comments of each community are identified to recognise the causes of difference in the patterns of activities and interests between communities with English comments and communities with foreign languages.

1.2.2.8 Distribution of Interests and Dominant Interest

The number of keywords in each interest topic determines the interest variety of each community. The categorised keywords in the large-sized communities are portrayed in diagrams that reveal the distribution of interests and allow the detection of communities’ hierarchy of interest.

1.2.2.9 Timeline of Interests in Each Community and Influences from Events

Comments within each large-sized community are identified and separated into 12 months based on the dates of the original posts, which the comments are made about. Alongside, to detect significant changes which may be due to major events, a timeline of public holidays, school holidays, and events held at Sylvia Park Shopping Centre is compiled and used to explore the correlations between these events and 1) the amount of comments made for the posts in each community, and 2) the patterns of interest.

1.2.2.10 Influences of Central Person

The interests of the central person, identified in the network analysis, of each large-sized community are explored by observing the keywords used in their comments. This enables the detection of the central person's interests and dominant interest. In addition, this reveals whether the central person's interests align with their entire community's interests, showing the strength of influence central persons may have on their communities. Semantic Analysis for Dynamic Communities

The keywords are separated into 12 months for each large-sized dynamic community identified in the network analysis. This allows an investigation on the interests' pattern to recognise the interest which initiate the community's establishment.

2 DATA ANALYSIS

2.1 NETWORK ANALYSIS

2.1.1 Relationship Between Elements in a Multi-Modal Network

2.1.1.1 Typology

Although the value of a single like, comment, or post is difficult to interpret on its own, an entire system of repeated and reciprocal interactions between thousands of people provides valuable information which differs with each kind of communication. As likes are the most ubiquitous form of interaction, it ties many people together and gives a better understanding of community structure. In addition, the keywords found in comments can be analysed to identify the major subjects that are discussed within and between communities.

a. Proportion and Size

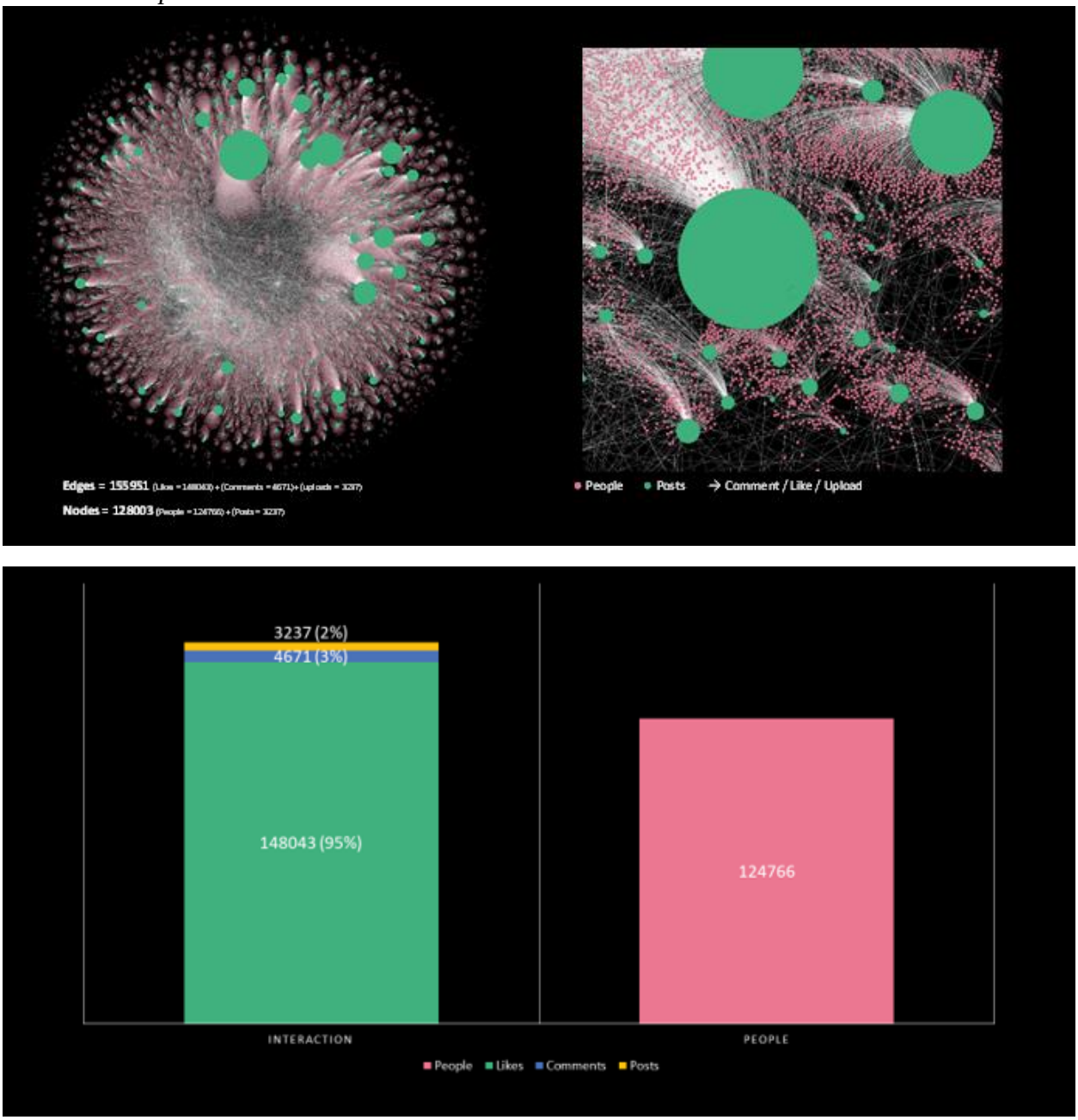


Figure 1. Network size (top) and umber of interactions vs people (bottom).

New relationships can be found between the different forms of interaction by dissecting the network into its individual parts. This points out the number of people using multiple types of communication, the preferred type of communication, and the proportion of people using each kind of interaction. The network comprises 155,951 total edges and 128,003 total nodes. The edges are further categorised by 148,043 likes (95 %), 4,671 comments (3 %), and 3,237 uploads (2%). Furthermore, nodes are made up of 124,766 people (97 %), and 3,237 posts (3 %). A total of 95 % of people interact solely through likes, compared to 1 % of people who only use comments; 2 % of the network, or 2,343 people, interacts through both comments and likes. Only 40 people, or 0.03 %, post, comments and likes.

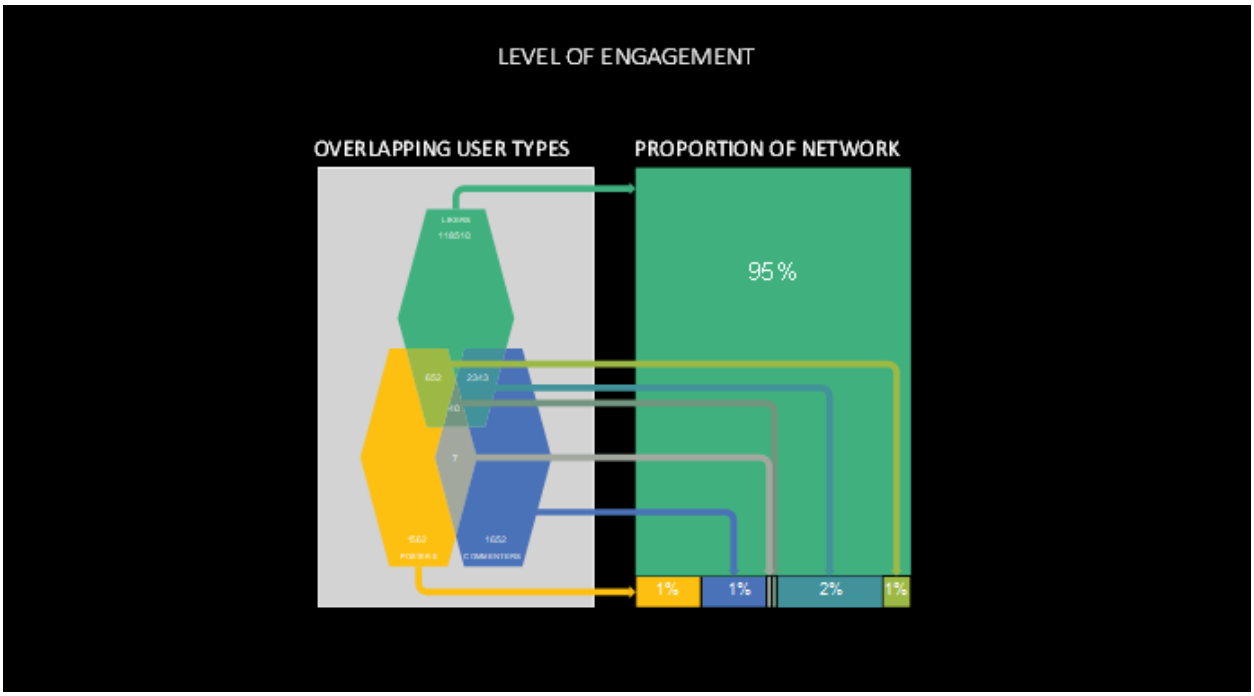


Figure 2. Level of engagement.

2.1.1.2 Key Features and Classification Process

Various metrics are rated, from high to low, to identify a set of key features for classifying distinct network typologies. The method has been developed by Itai Himelboim et al. for use in many research fields by identifying recurring structures that appeared in different social networks ranging from social movements, viral topics, and international conferences. In contrast to monetary approaches to the analysis of public space, Himelboim's method has great potential as it revolves around people's interaction. This method enables to understand the aggregate performance of the network by relating all elements and focusing on three key metrics.

First, all members of the same community are merged into one node and their influence is measured with centrality distribution. Most hierarchical systems have a great difference between the centrality of the most dominant and second-most dominant communities. Communities such as these resemble a hub-and-spoke which points inward *(broadcast network)* or extends outward *(support network).*

Second, the interconnectivity between a subset of nodes is assessed using modularity and density. This identifies nodes which share mutual contacts, then creates a larger crowd that can be identified by the same colour. This generates a varying number and size of crowds which are either unified (tight crowds) or divided (polarised crowds)

Third, the extent of exclusion from crowd uses the number of isolates and their size to investigate the lack of connectivity in specific parts of the network resulting in fragmented (*brand clusters*) or clustered structures (*community clusters*).

Classifying the entire system using this method provides the opportunity to draw on existing knowledge on all network types. This reveals how drastically different social conditions naturally arrange all network elements into specific configurations.

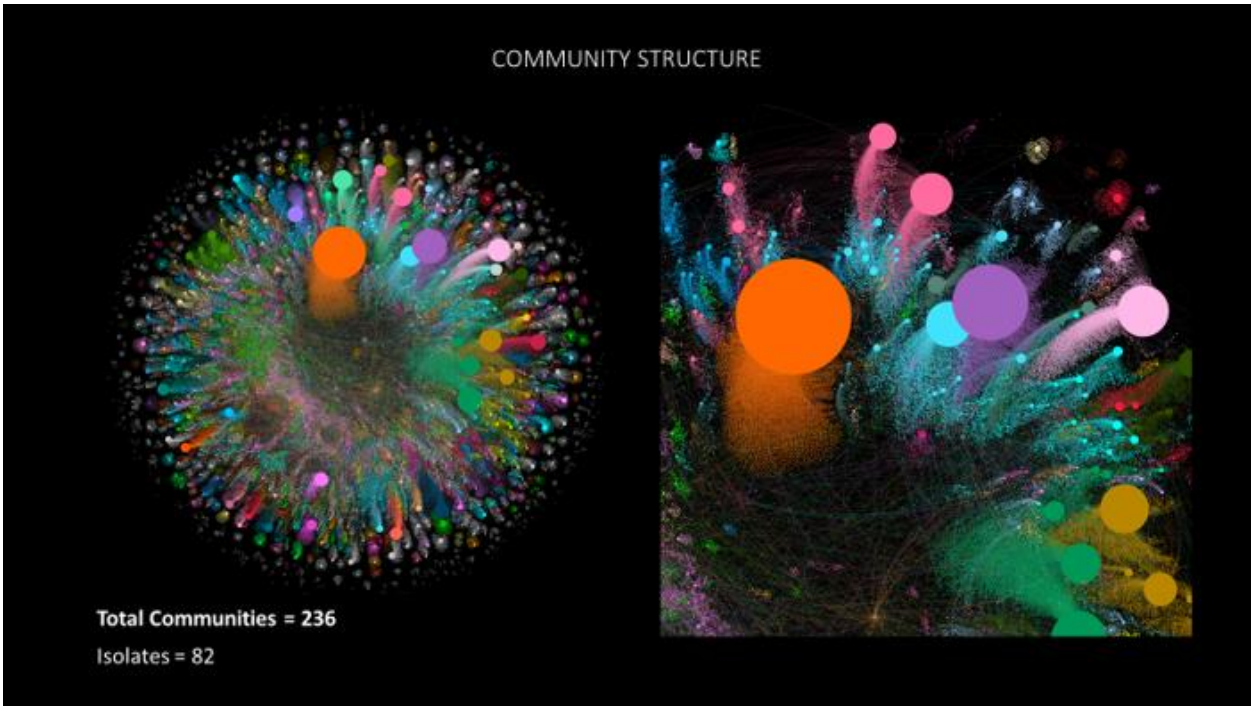


Figure 3. Community structure.

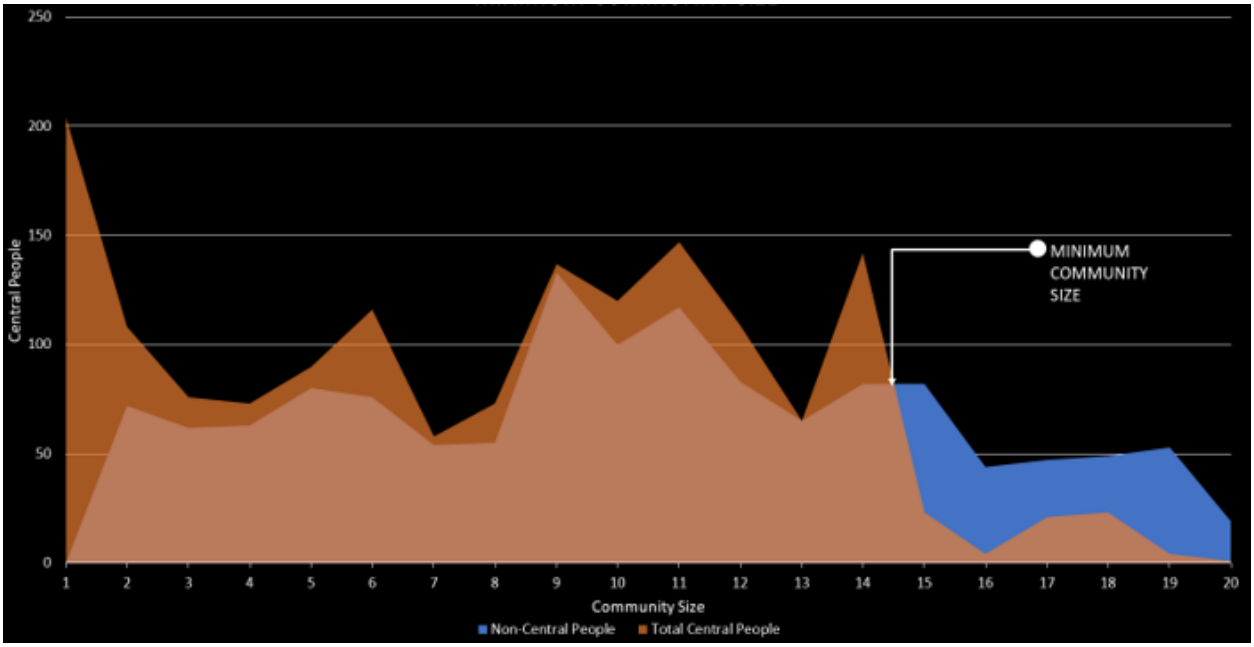


Figure 4. Minimum community size.

2.1.1.3 Finding Isolates and Minimum Community Size

Criteria are needed for finding a minimum community size, as modularity detected communities ranging in size from one member to 3,781 members. Rather than setting an arbitrary threshold, we look for community with no distinguishable central figures. Communities that had fewer than 15 members were found to have more central figures than non-central figures. Due to the lack of size and low level of interaction, a few likes are enough to become the central figure in each group. As multiple people tied for

the highest centrality in these small groups, they have no disguisable leaders but only those that under-engage. However, in groups larger than 15 people there is enough difference between members to identify central figures and analyse the distribution of interactions.

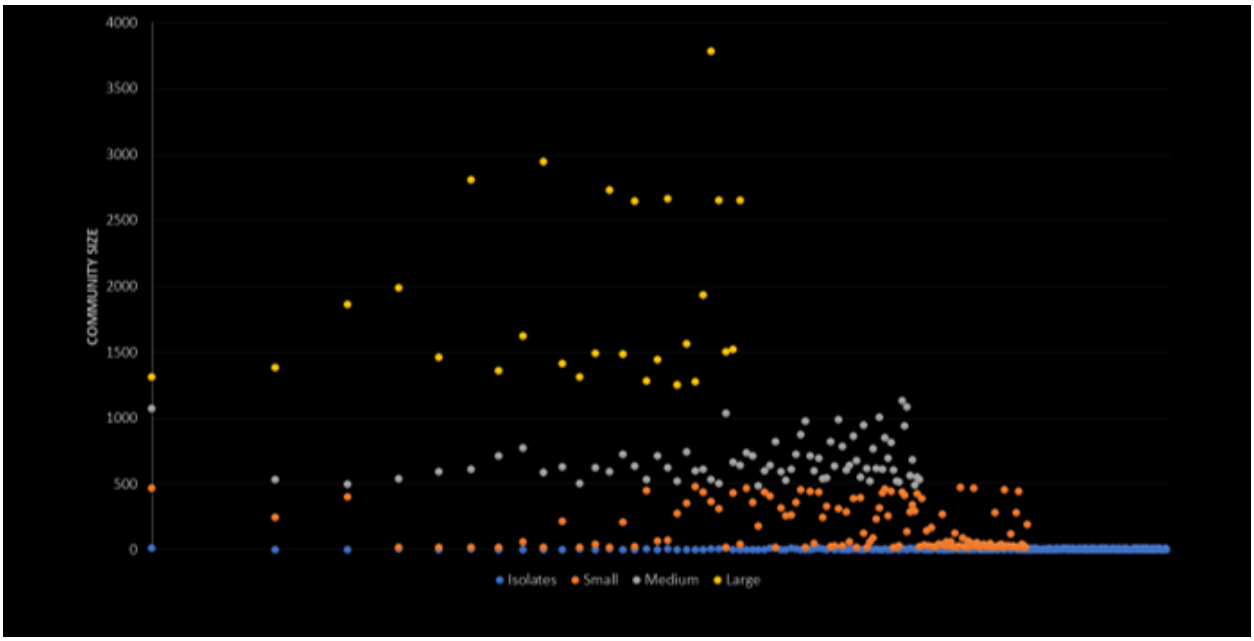


Figure 5. Arranging communities into size brackets.

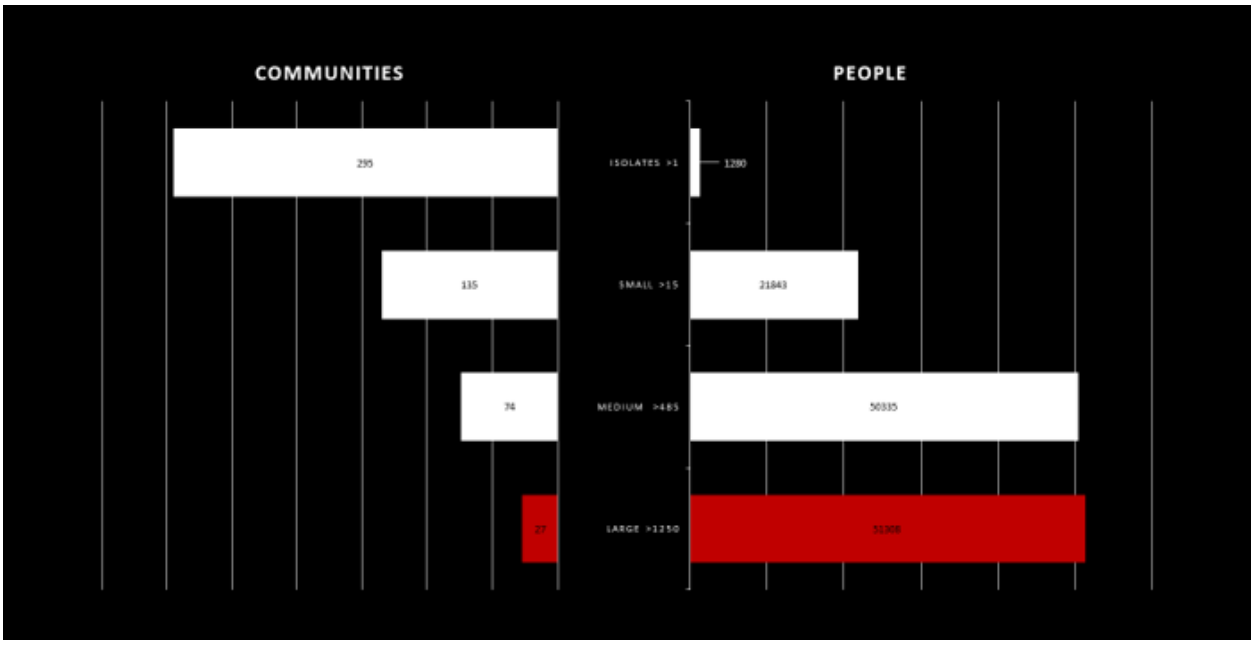


Figure 6. A few large communities have more impact than hundreds of isolates.

2.1.1.4 Defining Size Brackets

Unlike minimum community size, a maximum community size cannot be found as there is continual growth. Additionally, the distribution of community sizes cannot be understood by using only two

reference points (min and max). However, by counting the number of communities at each size, we can detect multiple sizes where communities naturally stabilise using kNN. In addition to the previously established threshold for isolates, three size brackets are found: isolates > 0; small > 15; medium >485; large >1,250. If the number of people within a community is known, it can be arranged into a size bracket. It was discovered that, while there are 295 isolate groups, their impact on the network is negligible. In contrast, there are only 27 large communities, yet they make up 41 % of the entire network. Therefore, as size brackets become larger, there are fewer communities, but they have a greater impact on the entire network.

2.1.1.5 Network Typology (Hybrid)

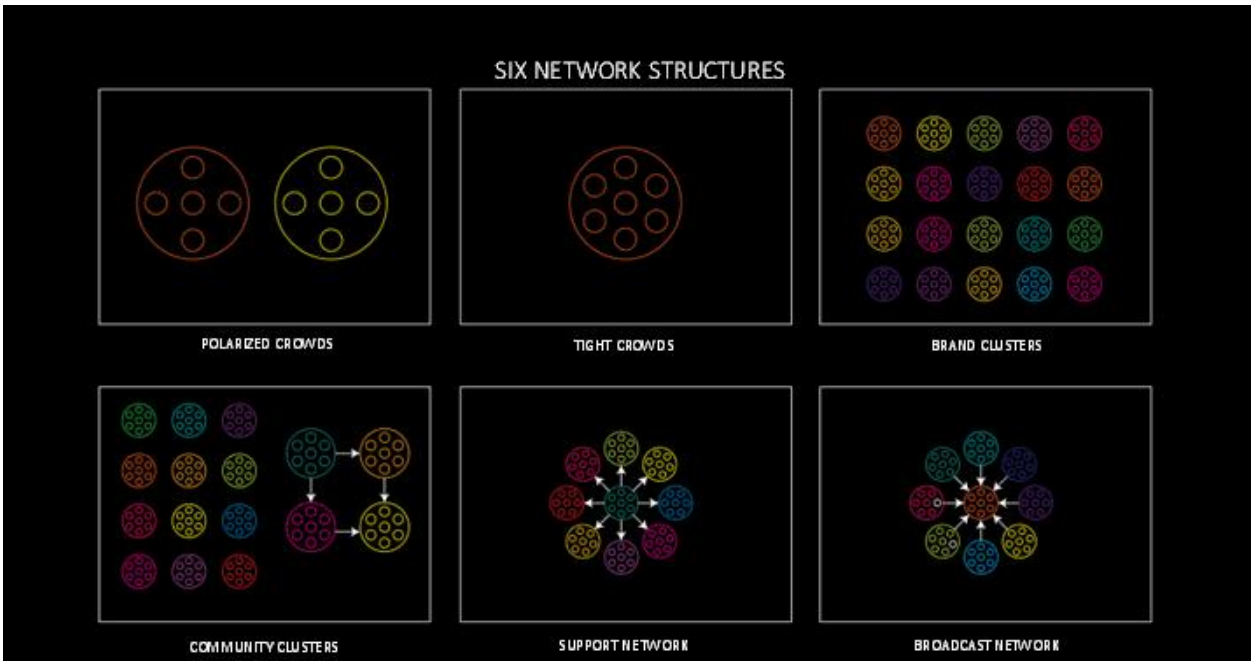


Figure 7. Diagram of the six network structures adapted from "Mapping Twitter Topic Networks: From Polarized Crowds to Community Clusters," by M. A. Smith, L. Rainie, B. Shneiderman, and I. Himelboim, 2014 *Pew Research Center, 20*, 8. Smaller circles represent “communities,” larger circles “crowds” (macro-communities).

Many social networks have parts, which operate independently from each other; and their characteristics are not always consistent. Therefore, not all networks can be perfectly characterised under the individual typologies proposed by Himelboim. Indeed, the result of following the classification process detected a hybrid system which has aspects of multiple types. An effective approach to addressing this variation is to locate where these key features change throughout the network to identify areas that operate as different types. This is done in individual steps by searching for key features which meet the criteria of each of the types, following the order of centrality, modularity, and fraction of isolates. Considering these features, the network seems to have a combination of features, as large interconnected groups and fragmented groups co-exist. Thus, the network is primarily five distinct crowds that resemble a *community cluster* and is interspersed with many disconnected communities that resemble *brand clusters.*

This network does not perfectly follow the structure of broadcast networks despite containing known celebrities (central persons). There are no well-defined central communities as many of them hold comparably high centralities ranging from 83 to 118. Rather than belonging to one prestigious community, these celebrities are dispersed throughout the network and their followers are weakly

connected to each other. Further investigation reveals that the celebrities in this network have a limited sphere of influence that relates to different ethnic groups such as Indian, Thai or Korean.

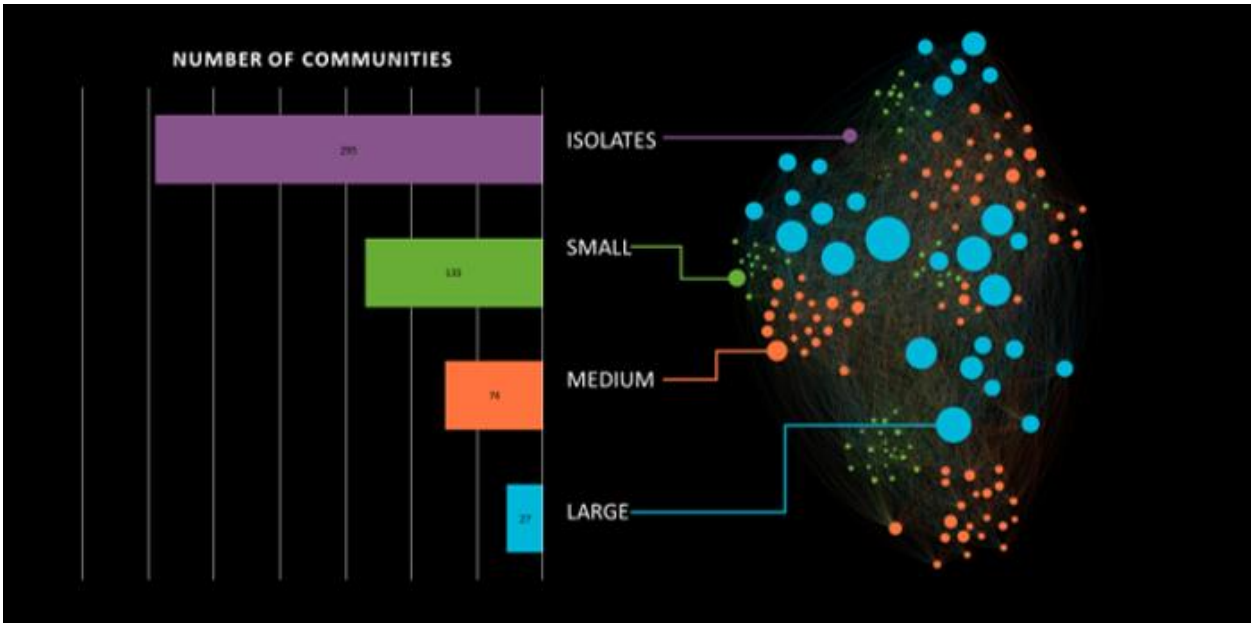


Figure 8. Key feature 2: Group size and count.

The network is highly interconnected forming five distinct crowds which include all 27 of the largest communities. Unlike *polarised crowds*, this network shares information from multiple sources and holds diverse opinions. As the public can openly participate in the network, the high interconnectivity does somewhat/slightly resemble a tight crowd which is made up of experts who specialise in one topic.

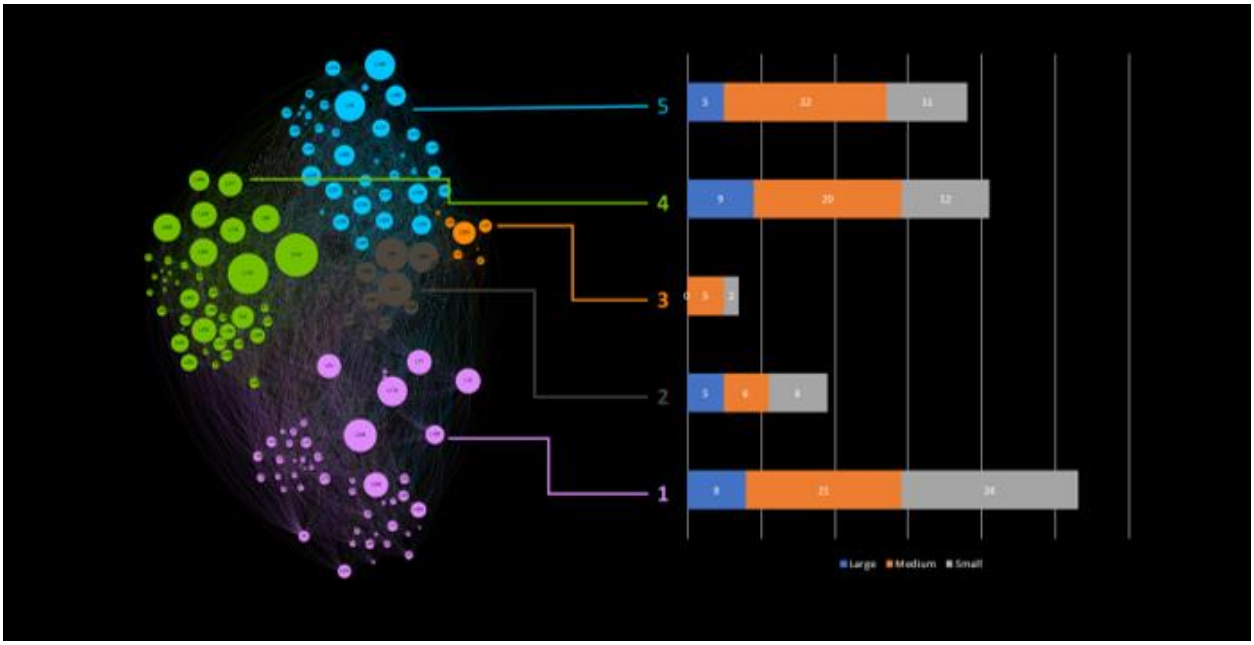


Figure 9. Key feature 3: Level of group interaction

The fragmented area of the network is typical of *brand clusters*. These fragmented populations are many small communities that are attracted to specific topics yet lack connections between each other. They remain separate as they do not have mutual connections that can form a bridge. To identify brand clusters, we find groups made up of isolate or small communities which show at least twice the average interest in only one major topic. The dominant interests of the 52 total communities which meet these criteria consist of: Art (4 %), Beauty (6 %), Places (13 %), Fashion (17 %), Events (19 %), Food (19 %), and Social (21 %).

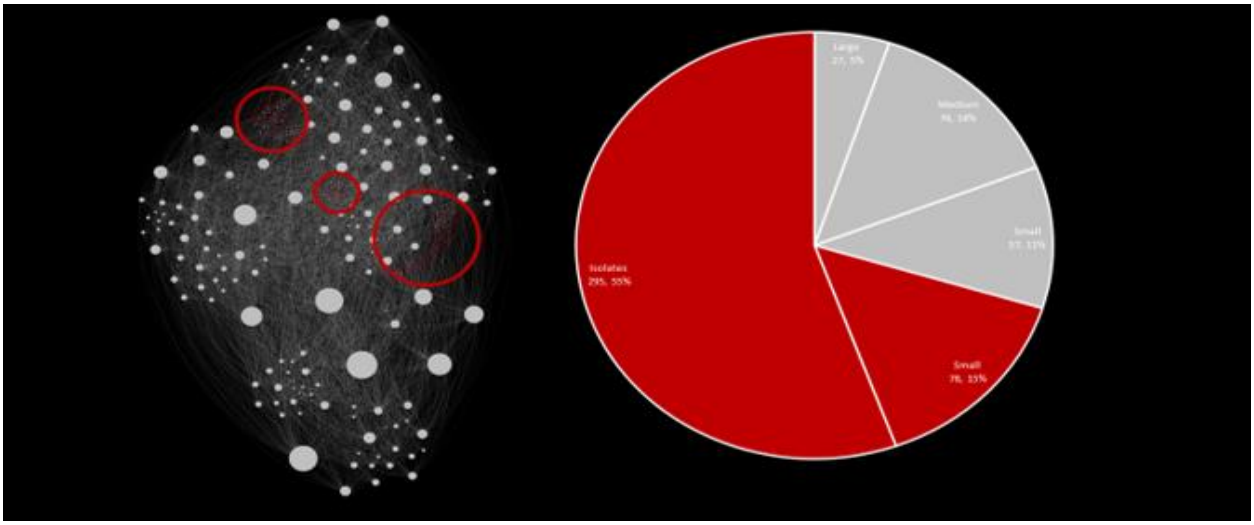


Figure 10. Key feature 4: Percentage and size bracket of communities excluded from macro-groups.

2.1.1.6 System Structure (Layered)

Networks can grow more unified or fragmented depending primarily on cluster size and level of engagement with other communities. The network may be thought of three dimensionally, as multiple layers that operate under different conditions. As the stack of layers collapses into a single layer, the network becomes more unified. In this case, three layers were identified.

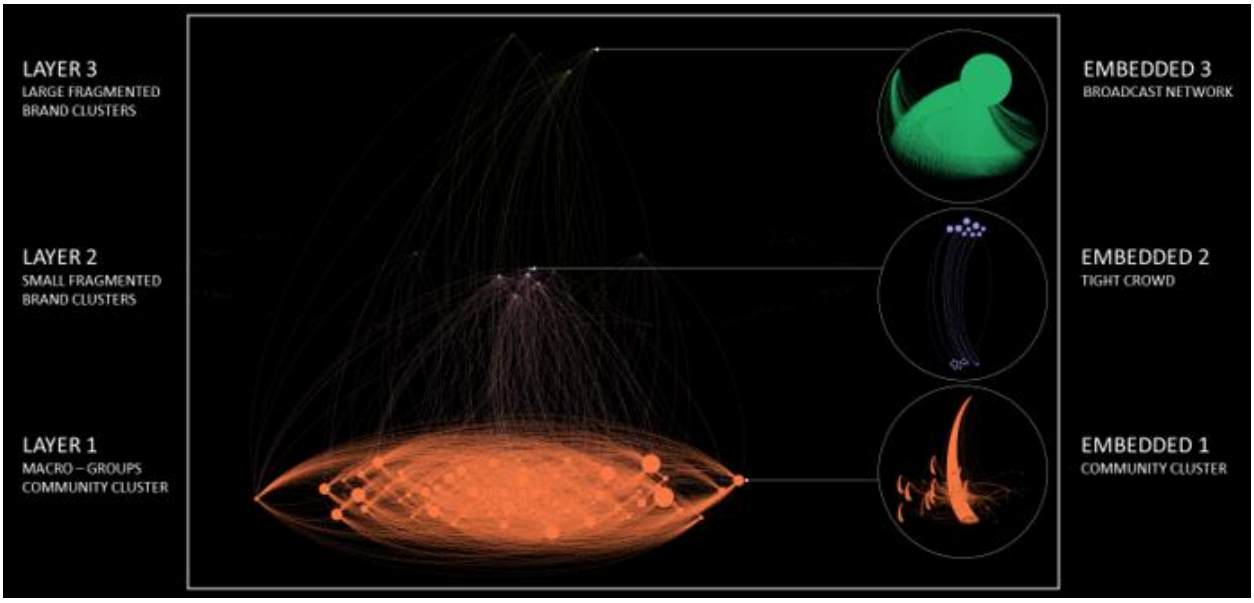


Figure 11. Layered system.

- **Layer 3: Fragmented large communities resembling brand clusters**
  Due to the hierarchical structure, spontaneous communities which form around celebrities enter the network as a single large and unified cluster. However, once the celebrities stop engaging, the whole community's activity drastically drops and becomes dormant.

- **Layer 2: Fragmented small communities resembling brand clusters**
  On the other hand, once smaller clusters enter the network, they operate differently depending on their level of engagement. At low levels of engagement with other communities, they remain small and isolated and have a limited sphere of influence. Conversely, they may interact with other communities where they become part of a crowd.

- **Layer 1: Integrated multi-sized crowds resembling community clusters**
  Community cluster are made up of a few distinct groups that have their own central figures and audience. They are tightly interconnected and easily reach different parts of the network.

Unification of these three layers can occur in several ways such as when small communities broaden their interaction beyond their own members and merge with larger communities (Layer 2 to Layer 1). This can also happen if a hub of a self-contained large community brings their followers when they join a crowd (Layer 3 to Layer 1), and, also, when the followers gain interest in the activity of other communities unrelated to the celebrities' activity (Layer 3 to Layer 1).

This would result in a crowd containing a mixture of large, medium and small communities. In this network, all large and medium-sized communities were members of the crowd, whereas only 42 % of small community were included. This suggests that community size alone is not what determines a more unified network, but the level of interaction specifically of small communities with large communities. This is because smaller communities tend to merge with larger communities rather than the other way around. Therefore, fewer large communities in a crowd also means fewer communities of all other sizes.

While it is possible to form a crowd with only interconnected medium and small communities, this leads to a drastically fewer members. Crowd 3 is 71 % medium communities and 29 % small, with no large communities at all. As a result, its size is considerably smaller with only seven communities compared to other crowds which contained between 19–53 members. Therefore, even if a crowd could eventually be formed by merging only small communities, it would only have access to a small percentage of the entire network.

2.1.1.7 Subsystems Analysis (Interactivity of Embedded Subgroups)

The way groups of people operate can be analysed at various scales: full networks, layers, crowds, or communities. A node may represent an individual, but it can be merged with others to represent a group of people. This merging process is done sequentially to identify the hierarchical structure of a community. As the number of steps of the process is extremely high (potentially as high as the total number of members), finding a meaningful scale for the analysis of subsystems requires careful consideration to identify the factors which influence the relationship between subsystems and layers. Once the structure of different communities is detected, communities can be compared to identify how they operate at different levels, and the patterns of their subsystems classified. Findings show the degree of openness/closure or extroversion/self-containedness of each group toward the rest of the network.

2.1.1.8 Finding Meaningful Subsystems

Subsystems are found by zooming deeper into a network so that it can be analysed in more detail. However, the focus on a subset of nodes becomes increasingly narrower. At a certain point, subgroups will be too small to assess key features as they will contain only a few connected individuals. Excessive subdivision results in groups which appear as miniature tight crowd or broadcast networks. This is because more complex structures such as community clusters require more people to be able to identify groups with distinct hubs. By considering this with our own dataset, modularity has been run once to identify subsystems, and a second time to identify crowds. This results in parts that are large enough to recognise key features and, also, closely connected so that only their direct impact is observed.

2.1.1.9 Relationship Between Subsystems and Layers

Parts of a network can be independent from each other or closely bonded. Subsystems become more unified when their members form a more complete structure through stronger and shorter connections. Conversely, a community can also split when it becomes more connected to members of other communities than members of the original community. Furthermore, a group can be more integrated within the overall system depending on its sphere of influence. This may happen when communities with dominant hubs have widespread influence.

The influence of a celebrity may be concentrated into a specific area of a network. This has been found to be particularly true with broadcast networks that revolve around international celebrities. By manually assessing the central figures of these communities, we have found that, out of 27 large communities, four of them were international celebrities with prestigious occupations such as singer, actor, or fashion designer. Seven communities revolved around local celebrities ranging from bloggers, athletes, and journalists. Sixteen of them did not have identifiable celebrities or celebrities with noteworthy occupations.

International celebrities had less engagement than local celebrities despite having a considerably larger following. Although the New Zealand celebrity with the largest following had a number of followers much smaller than the largest international celebrity, they had 43 % more incoming interactions per person. Furthermore, local celebrities had more posts spread over the whole year. By tracking whether central figures have posts in two non-sequential months, we can detect if posting behaviour is consistent or inconsistent. Only one out of the four international celebrities posted consistently (25 %), while 17 out of 23 local central figures posted consistently (74 %). The more sustained and less hierarchical structure of communities which contained local celebrities had diverse interests that were not as focused on one individual. This means that members from these communities have a greater impact on the rest of the network than international celebrities.

Additionally, the area of isolates or small/fragmented communities acted as tight crowds. Each of them enters as group of people rather than becoming more associated over time. Their sphere of influence encompasses their own group and has little effect outside of their own community.

On the other hand, the area made up of community clusters is large enough to be divided into smaller community clusters. They have a complex structure made up of various well-connected areas and multiple hubs. Unlike the other subsystems, they can be further divided into smaller parts if desired. They are also well connected as a self-contained group and to other communities.

2.1.1.10 Assessment of Full Networks with Non-Normal Distributions

At the most general level of analysis, averages are suitable for comparing differences in interaction for two sites without dividing the network into communities. This works well in cases when the data is

normally distributed, as all measures of central tendencies are close in value and give reliable results. Although these wide-encompassing values can describe entire datasets, they fail to do so when there are large variations between parts or individuals. Furthermore, social networks have been known to have asymmetrical distributions and have distorted averages due to outliers with extremely high or low levels of interaction. The collected dataset does not have a normal distribution even after removing commercial accounts with abnormal behaviour, and plotting each type of interaction separately. Furthermore, as we are concerned with only one site, alternative measures have been used to assess the whole network in a way which also overcomes issues with its non-normal distribution.

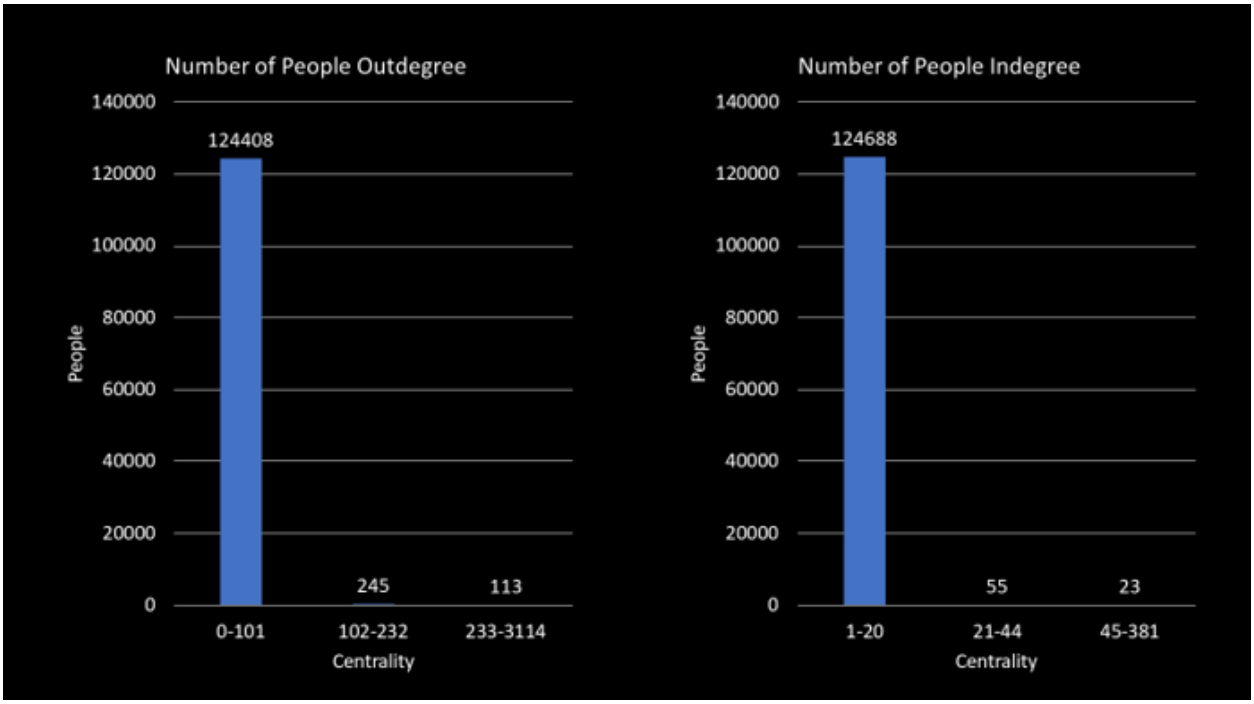


Figure 12. Number of outdegree and indegree people.

2.1.1.11 Distribution and Hierarchy

The distribution shows how a person who sends or receives a certain number of interactions fits within the higher or lower levels of engagement within a hierarchy. This is done for the total incoming or outgoing interactions and is also assessed by type. A hierarchy can be thought of as a structure with multiple tiers, each containing branches. A power law distribution tests whether a hierarchy exists; the distribution plot tells us how many branches are in each tier (width); its range of the indicates the number of tiers (height); and the difference in centrality between the ranks tells us how large the gaps are between tiers (thickness of tier).

2.1.1.12 Power Law Distribution and Hierarchy by Interaction Type and Direction

A power law distribution describes a network with a highly influential minority. This can be recognised by comparing the likes, comments, and posts with the number of people who hold a specific centrality value. As the number of interactions increase, the number of people who have that amount of interactions decreases. By the same logic, as the number of interactions decrease, the number of people increases. All forms of interactions in this dataset follow this power law pattern for both incoming and outgoing interactions. The network is hierarchical because people tend to interact with those who already receive a lot of interactions. Over time, the gap between the influential minority and majority increases.

2.1.1.13 Range of Distribution by Interaction Type and Direction

The range of the distribution is the number of unique values which fit between the largest and smallest centrality. It includes people with no received interactions if they have outgoing interactions. Indegree centrality ranges from 0–3,114 interactions with 307 tiers, versus outdegree centrality which ranges from 1–381 with 60 tiers. Since the number of received interactions is dependent on many people rather than an individual, the indegree range is significantly higher than the outdegree range, across all types of interactions. The highest number of indegree likes exceeds the highest number of outdegree likes by 2,441. Similarly, the highest number of indegree comments is 685, more than the highest number of outdegree likes. Dividing the range by the number of tiers provides a general distance between tiers. There is an average difference of 15 interactions for each indegree tier but only a difference of 6 for outdegree. However, the tiers at the bottom and top of the hierarchy can be focused on to be more accurately assessed. This is done by comparing the gap between the two highest, and the two lowest centrality values. For the gap between the two highest ranks, there is a 32 % difference in total outdegree centrality (92 interactions) compared to 9 % for total indegree (261 interactions). The biggest gap by interaction type is indegree comments which range from 85 to 702, a 726 % difference (617 interactions); the outdegree comments range from 10 to 17 (70 % difference or 7 interactions). The smallest gap by interaction type was indegree likes (17 %), followed by the second smallest which was outdegree likes (32 %). A single unit separates the two lowest ranks for comments, posts, and likes.

2.2 SEMANTIC ANALYSIS

2.2.1 Semantic Analysis of the Whole Network

2.2.1.1 Keywords and Topics of Interest

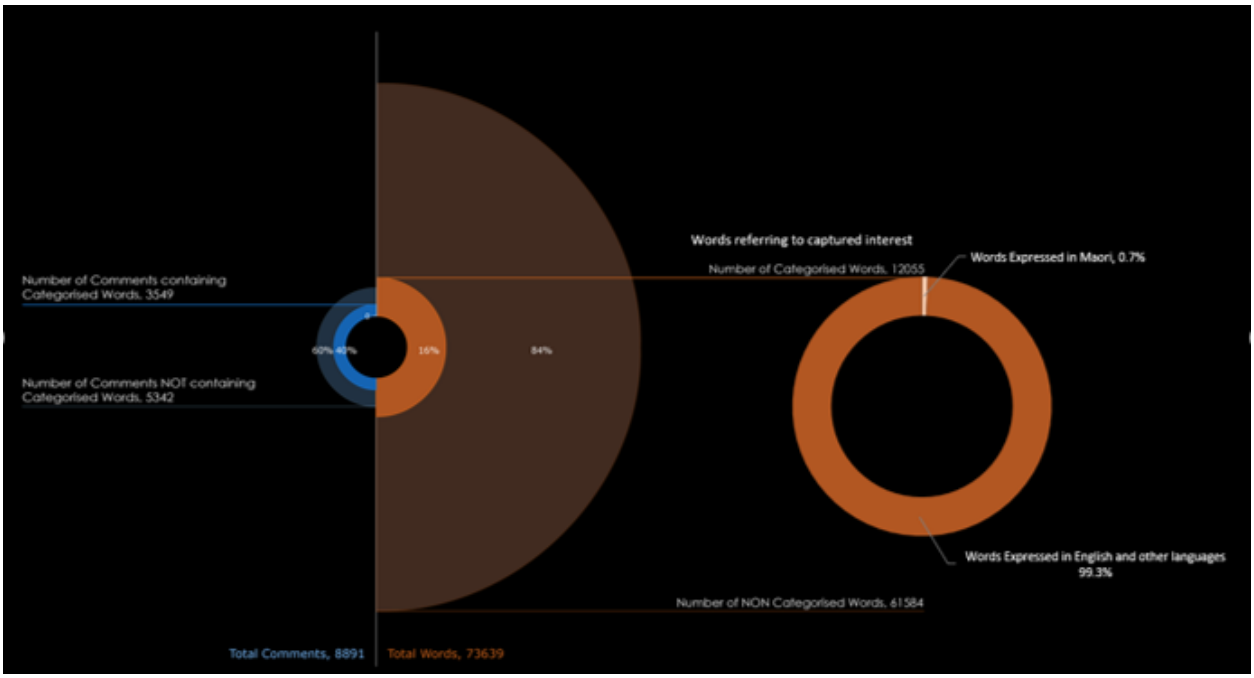


Figure 13. Number of comments and words in 2017.

Semantic analysis investigates textual elements like captions and comments of Instagram posts tagged at Sylvia Park Shopping Centre in 2017 to explore and understand public activities, events, and topics of interest which stimulate people's interaction within a public space. This analysis examines 73,639 words

from 8,891 captions and comments. A total of 40 % (3,549) of the captions and comments contain 12,055 keywords referring to captured interest which are analysed to inform the visitors' interest at Sylvia Park Shopping Centre throughout 2017. Moreover, analysis of the keywords allows an exploration of cultural significance and influence through languages used; 99.3 % of 12,055 keywords referring to captured interest are words expressed in English and other languages (using the Roman alphabet), and only 0.7 % are found to be words expressed in Māori.

2.2.1.2 Keyword Language

However, all 12,055 keywords are categorised into 10 different topics of interest including Animals; Art, Design, and Photography; Beauty, Sports, and Wellness; Events and Entertainment; Fashion and Style; Food and Drinks; Nature; Places and Architecture; Technology; and Social and People. This represents a total distribution of the public's interest at Sylvia Park Shopping Centre, as well as presenting the interest topics of keywords expressed in English and other languages in comparison to interest topics of keywords expressed in Māori. The topic with the most keywords expressed in English and other languages is Food and Drinks, with the highest percentage of 23.5 %, indicating that Food and Drinks is the topic which encourages the most communication between people. Additionally, the most used words of the topic are *coffee*, *lunch*, *chocolate*, *yum*, *burger* etc., suggesting predominant kinds of food, drinks, or meals, and perhaps events involving these foods and drinks at the Shopping Centre.

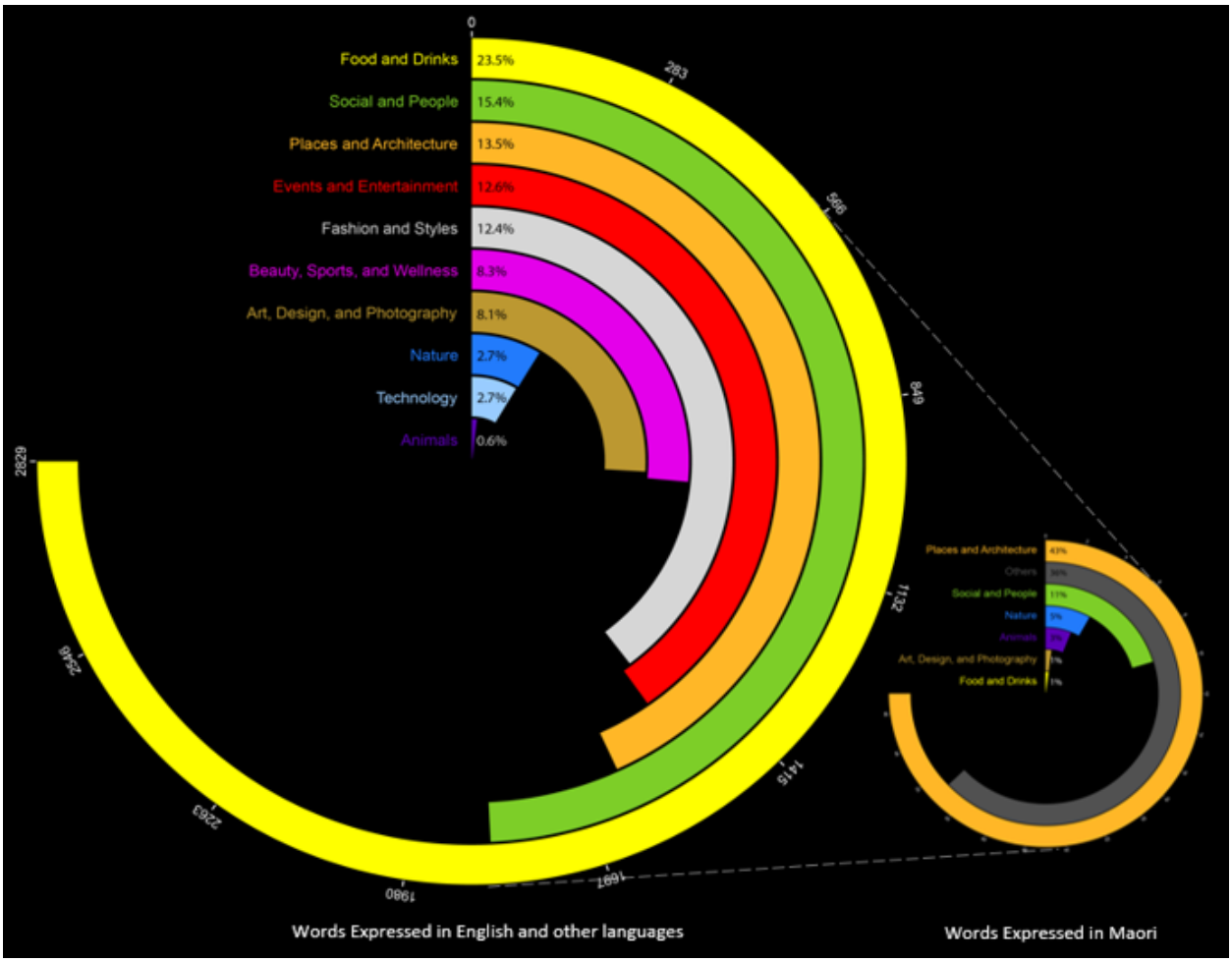


Figure 14. Topics of interest.

The next-largest interest topics are Social and People with 15.4 %, with the most used words being *bro, girl, guys, family, friends* etc.; Places and Architecture has 13.5 %, with the most used words including *Auckland, Sylviapark, Newzealand* etc., revealing that people are likely to mention others they are with or they know and places that they are in. Although there is less variety of keywords in Places and Architecture compared to smaller topics like Events and Entertainment, the most used word of the topic, *Auckland,* has the highest word frequency among all 12,055 keywords.

Events and Entertainment, with 12.6 %, and Fashion and Styles, with 12.4 %, have a similar percentage of interest level and are the second-most popular activities after Food and Drinks. Events and Entertainment consists of a variety of keywords; the most used words, including *shopping, Christmas, birthday, Xmas, Santa* etc., are significantly concentrated around one event like Christmas, indicating the powerful impact of Christmas at Sylvia Park Shopping Centre. Fashion and Styles includes most used words like *HM, Zara, fashion, dress, style* etc., revolving around particular clothing brands, thus the stores are the main attraction and destination for visitors interested in Fashion and Styles.

The next two topics with a similar interest level are Beauty, Sports, and Wellness with 8.3 % and Art, Design, and Photography with 8. 1%. Beauty, Sports, and Wellness's most used words include *hair, face, beauty, Lush, fitness, makeup* etc., suggesting words which perhaps involve a beauty event or beauty services within the mall as well as indicating a particular beauty store like Lush. Art, Design, and Photography's most used words include *photo, selfie, picoftheday, design, colour* etc. which are largely associated with photography, especially selfies.

Nature and Technology have an equal interest level of 2.7 % each. Their most used words are *flowers, sunset, and nature* etc. for Nature; and *Ford, cars,* and *iPhone* etc. for Technology; both topics consist of general terms which probably refer to a store like a florist within the mall or suggest a temporary event held at the mall.

Lastly, the lowest level of interest with 0.6% is Animals, with the most used words including *reindeer, panda, bat* etc. Noticeably, the keywords are linked to events like Christmas and perhaps movie names like *Batman* and *Kungfu Panda*. As pets like dogs and cats are prohibited within the public space of Sylvia Park Shopping Centre this may explain the lowest number of keywords in the Animals topic.

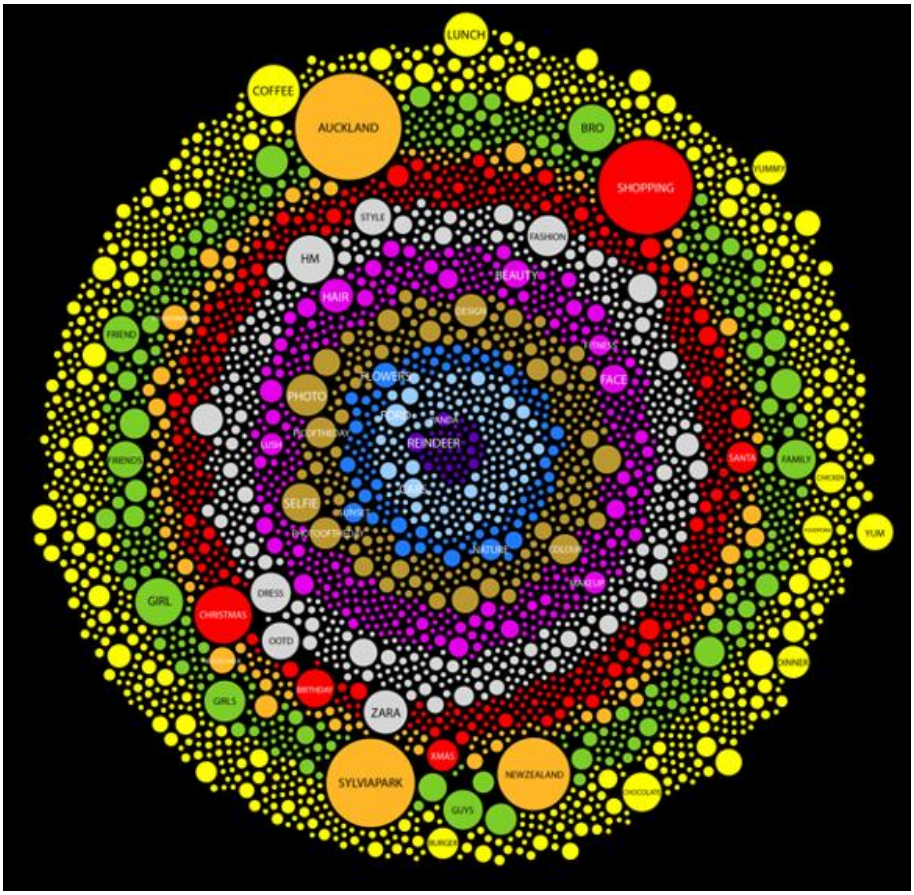


Figure 15. Most used words of each topic.

2.2.1.3 Keywords Expressed in Māori Language

Notwithstanding, the most interested topic of keywords expressed in Māori is Places and Architecture with a level as high as 43 %, of which the most used words are mainly place names such as *Aotearoa, Manukau, Onehunga, Orewa* etc., followed by 36 % of other uncategorised Māori words like *kiaora, ehoa, kaha* etc. Yet, the third most interested topic of keywords expressed in Māori is Social and People with 11 % of keywords including *Maori* and *whanau*, followed by 5 % of Nature with keywords such as *moana, paki, pohutukawa* etc., and 3 % of Animals with keywords including *pukeko, tui* etc.; the last two interest topics with 1 % each are Art, Design, and Photography and Food and Drinks, which include the keywords *moko* and *kaimoana*.

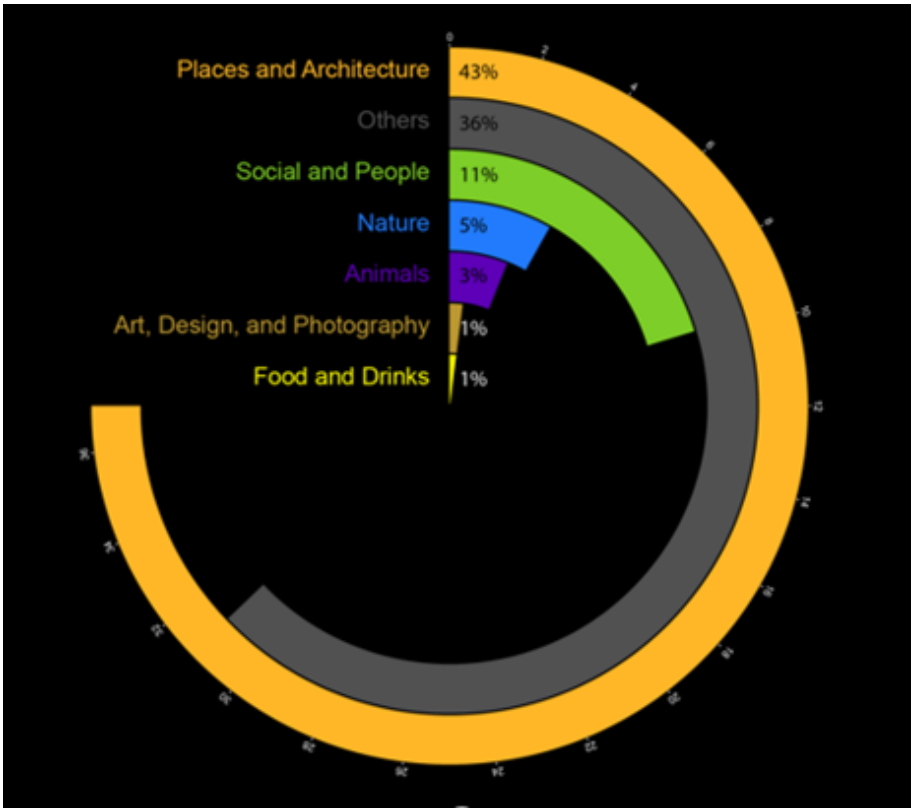


Figure 16. Words expressed in Māori language.

Certainly, keywords expressed in English and other languages present a different direction of interest topics compared to keywords expressed in Māori. The rank of interest topics derived from keywords in English and other languages shows more variety; topics are not only driven by personal interest but also influenced by the facilities and services available at the Shopping Centre, e.g., restaurants, cafés, cinema, clothing stores etc. Food and Drinks is revealed to be the most dominant topic and the leading activities that attract people and provoke a lot of communication at Sylvia Park Shopping Centre. While the rank of topics from keywords expressed in Māori indicates that the most used words tend to be the Māori names of places, objects, plants, or animals, those are widely used and known by the public—for example *Aotearoa, pukeko, pohutukawa* etc. Perhaps promoting food or products in Māori to the public could encourage more use of local language. Although this is an overall distribution of interest topics based on keywords expressed through comments at Sylvia Park Shopping Centre, it can be further investigated to understand the connection between people and these topics as well as the relationship between the topics.

2.2.1.4 Interest Network and Correlation of Interest Topics

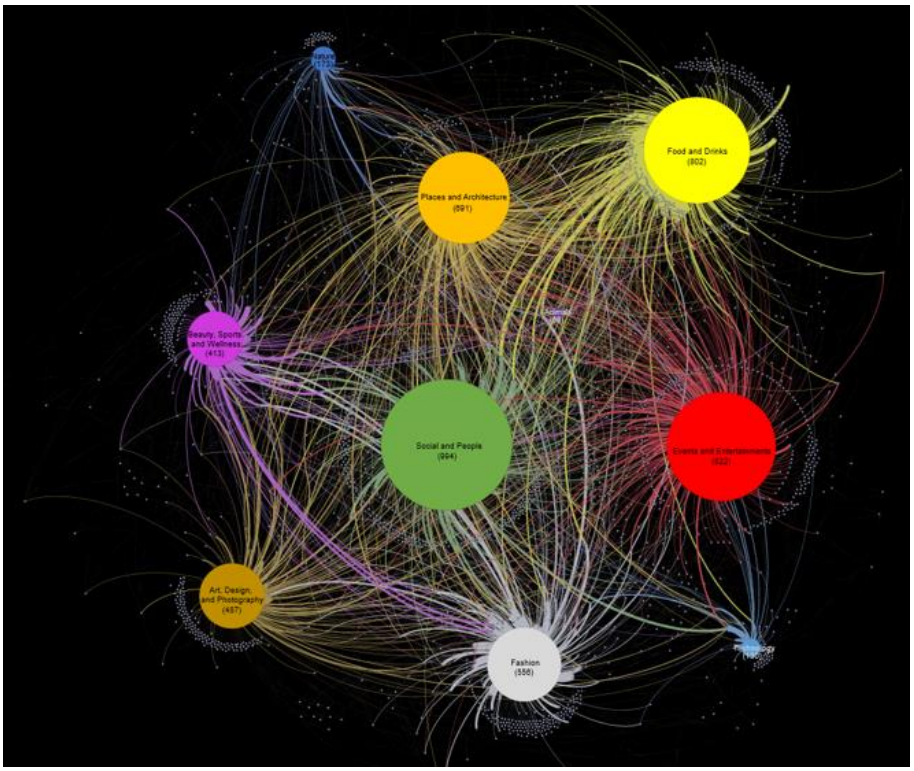


Figure 17. Correlation of interest based on keywords contained in comments.

	Animals	Art, Design, and Photography	Beauty, Sports, and Wellness	Events and Entertainments	Fashion	Food and Drinks	Nature	Places and Architecture	Social and People	Technology
Animals										
Art, Design, and Photography	0.041									
Beauty, Sports, and Wellness	0.039	0.112								
Events and Entertainments	0.115	0.175	0.075							
Fashion	0.018	0.157	0.120	0.106						
Food and Drinks	0.033	0.053	0.021	0.044	-0.011					
Nature	0.029	0.186	0.024	0.053	0.027	0.020				
Places and Architecture	0.075	0.261	0.087	0.290	0.157	0.278	0.181			
Social and People	0.033	0.150	0.089	0.208	0.116	0.060	0.027	0.156		
Technology	0.006	0.182	0.034	0.181	0.007	0.001	0.135	0.113	0.141	

Figure 18. Correlation matrix of various pairs of interest.

- Commenters

A network of interest represents a connection of people to the interest topics and shows how the topics can be correlated with each other. The data is analysed and visualised through a network of interest graph. The below graph shows small grey dots representing the commenters; linkage lines represent the comments containing keywords belonging to each topic, while the line thickness represents the number of keywords mentioned in the comments; and the 10 big circles represent the 10 topics of interest, where the circle size represents the number of people or commenters referring to the topic.

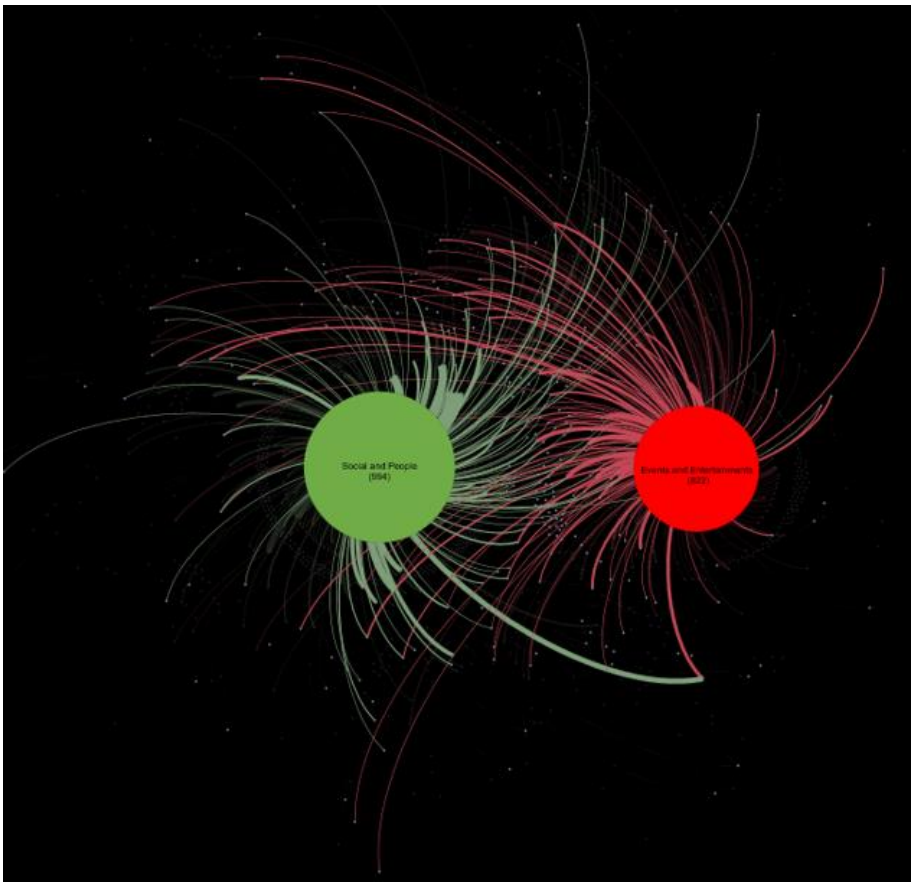


Figure 19. Correlation between Social and People and Events and Entertainment.

Although Social and People does not contain the largest number of identified keywords in the previous analysis, it is the largest circle in the network of interest graph, which means that the keywords in this category have been mentioned by the highest number of unique commenters (994 people) rather than a small group who make repeated connections. Furthermore, a correlation matrix of the topics informs that Social and People is highly correlated with Events and Entertainment. People are most likely commenting using keywords which refer to both Social and People and Events and Entertainment. Hence, since Events and Entertainment facilities are particularly aimed toward group activities, such as movies, mini golf etc, friends or family members are frequently mentioned.

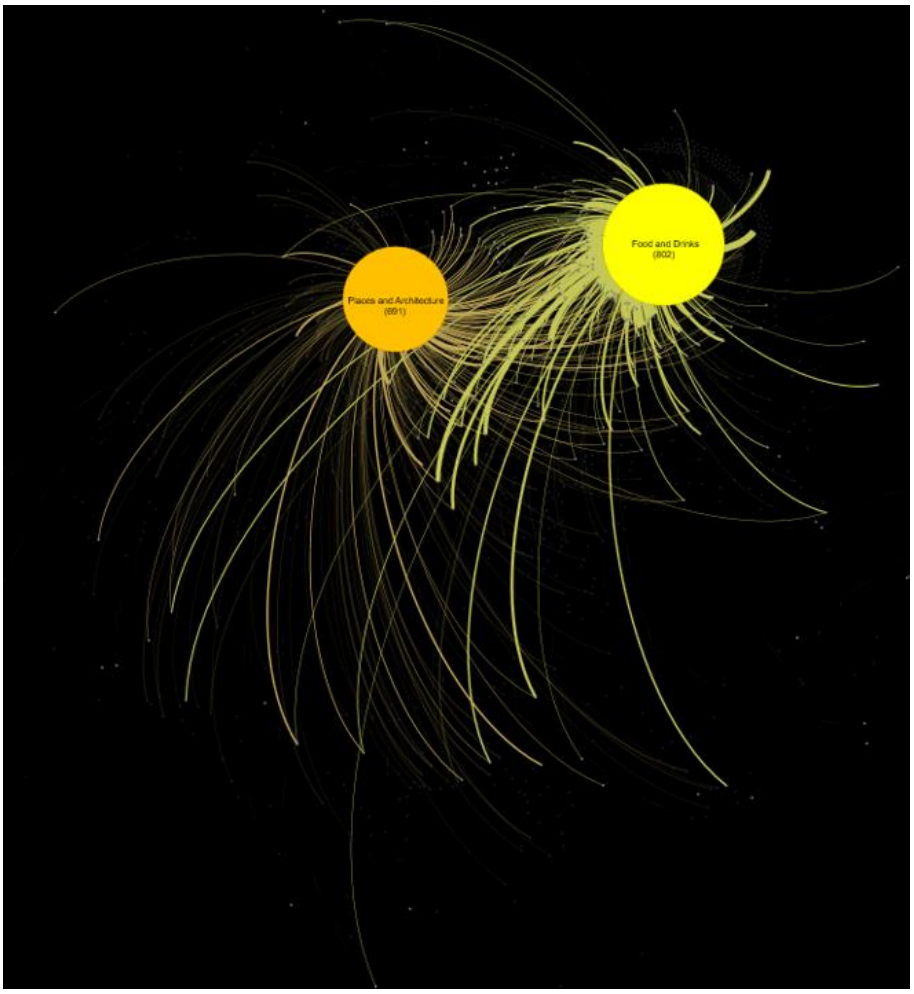


Figure 20. Correlation between Places and Architecture and Food and Drinks.

On the other hand, Food and Drinks, which contains the largest number of keywords (23.5 %), is the third most mentioned by the commenters (802 people) after Events and Entertainment (822 people). According to the network of interest graph, the connections of people to Food and Drinks include some single linkage lines but many dominant thick linkage lines. This significantly indicates a group of people who frequently used keywords from Food and Drinks in their comments. Thus, the large number of keywords identified *in Food and Drinks are* generated by a smaller number of people who are strongly interested in the topic and regularly refer to the topic more than once or use more than one keyword at a time. Moreover, Food and Drinks is highly correlated with Places and Architecture, indicating people's intention to use keywords in Food and Drinks along with a location or places where the Food and Drinks activities take place such as Auckland, Sylviapark, Newzealand etc., or the restaurant names e.g., Casablanca, Starbucks, Birdies etc.

Nevertheless, Events and Entertainment and Food and Drinks have a similar number of people referring to the topics (822 and 802 people), represented through the similar-sized bubbles shown on the diagram, although the number of keywords defined in Events and Entertainment (12.6 %) is much lower than keywords identified in Food and Drinks (23.5 %). Hence, people interested in Food and Drinks tend to use more than one keyword at a time, while a similar number of people interested in Events and Entertainment tend to use fewer keywords.

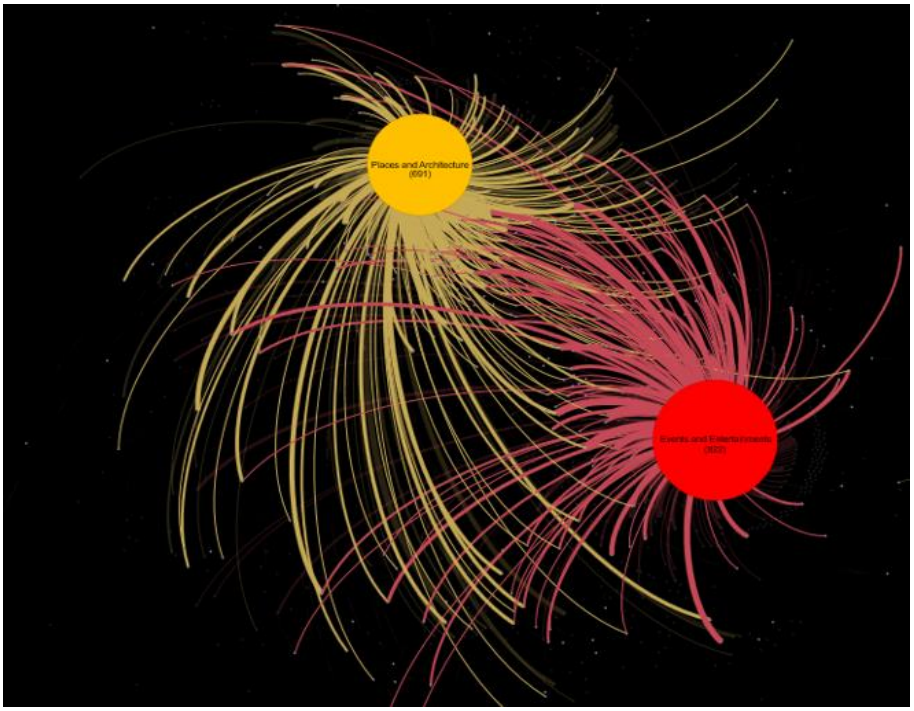


Figure 21. Correlation between Places and Architecture and Events and Entertainment.

Additionally, Events and Entertainment is another topic highly correlated with Places and Architecture, meaning people who comment about Events and Entertainment are also likely to mention Places and Architecture to share where the events take place. Moreover, Animals is highly correlated with Events and Entertainment, indicating a high chance of keywords in Animals and Events and Entertainment appearing in the same comments. Indeed, keywords identified in Animals (0.6 %) are mostly found related to events or entertainment rather than pets, as pets are not allowed in the Shopping Centre.

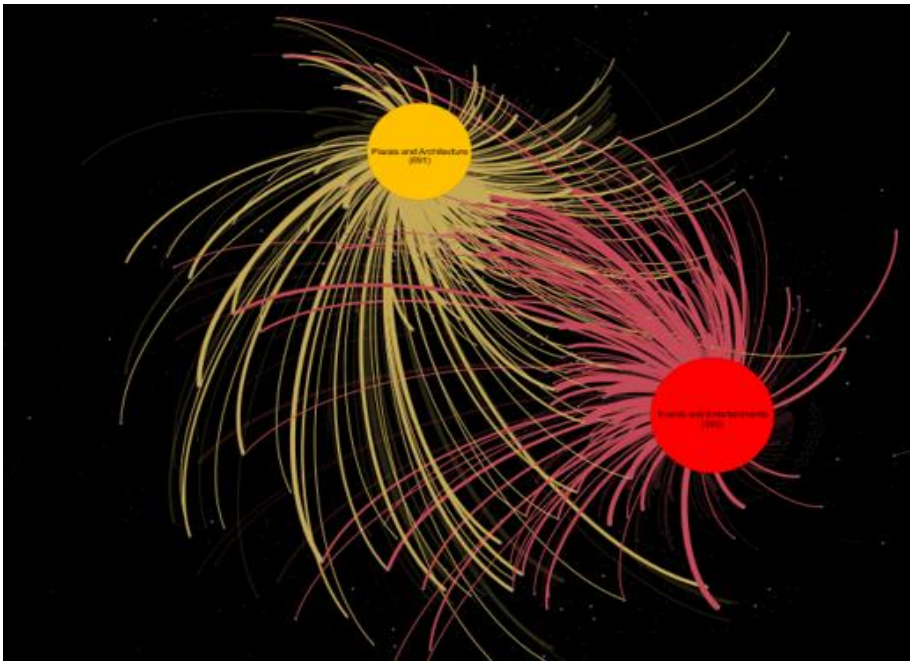


Figure 22. Correlation between Places and Architecture and Events and Entertainment.

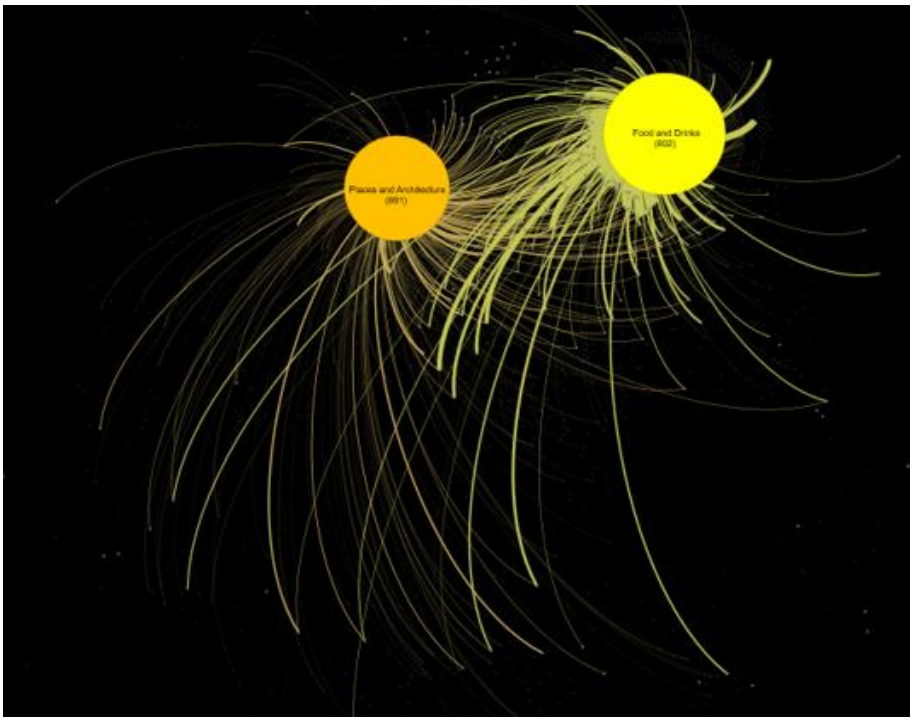


Figure 23. Correlation between Places and Architecture and Food and Drinks.

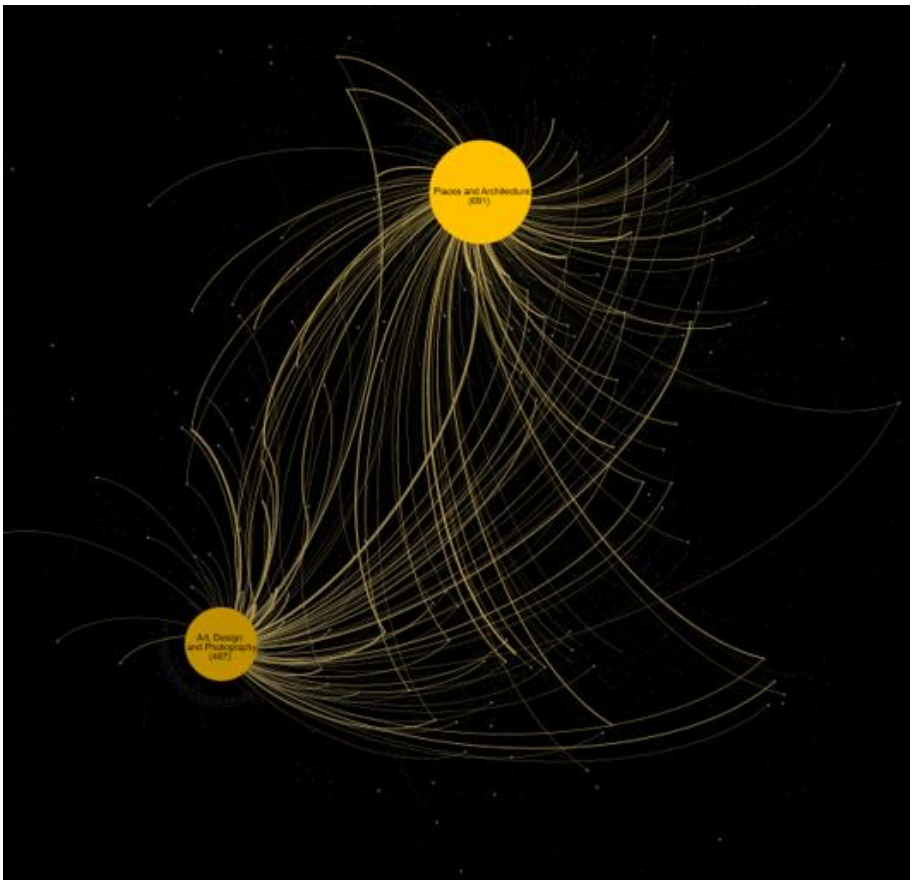


Figure 24. Correlation between Places and Architecture and Art, Design, and Photography.

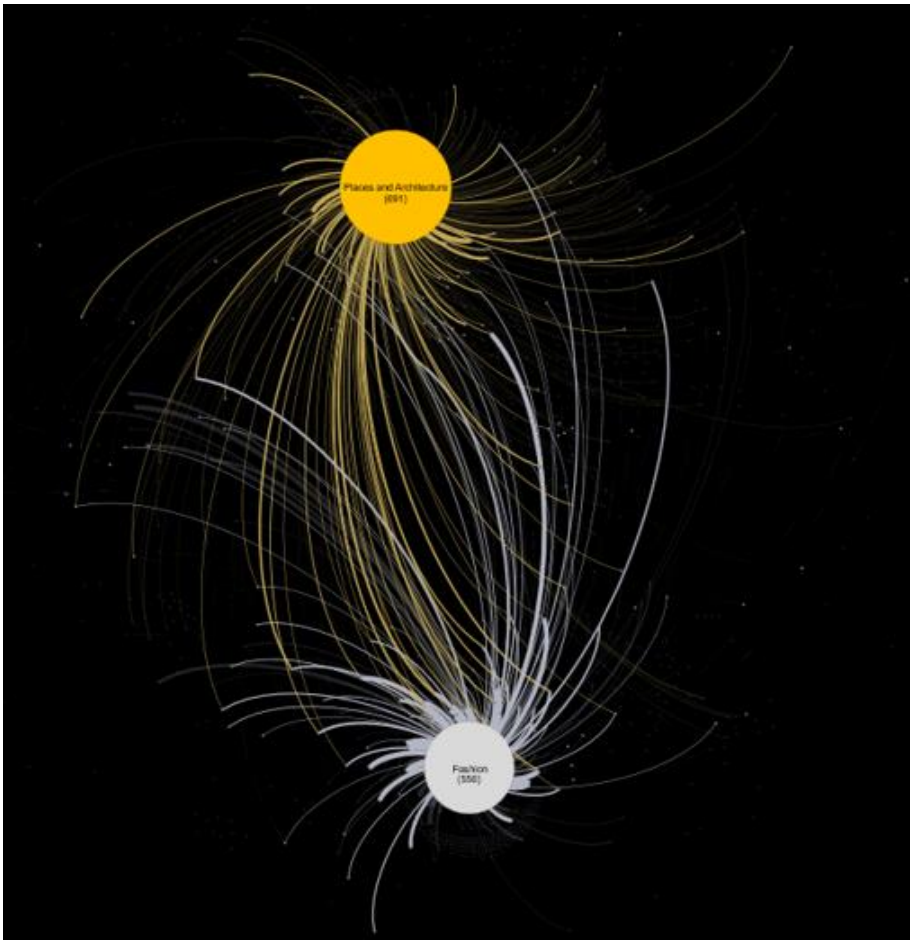


Figure 25. Correlation between Places and Architecture and Fashion.

Notably, Places and Architecture is highly correlated with four other interest topics including Events and Entertainment; Food and Drinks; Art, Design and Photography; and Fashion and Style, but Places and Architecture is the fourth largest topic in the network of interest. This shows that a smaller number of commenters (691 people) mentioned Places and Architecture and most Places and Architecture keywords are mentioned along with keywords from other topics, probably to promote the architectural features of the space or share the locations where the activities in various interest topics take place. Similarly, Art, Design, and Photography is another topic which highly correlated with multiple other topics. Other than Places and Architecture, these include Fashion and Styles, Nature, and Technology. This indicates a common use of keywords from Art, Design, and Photography to illustrate the design, aesthetics, or features of objects or places from the other topics.

Nonetheless, Fashion and Styles is equally correlated with Places and Architecture as well as Art, Design, and Photography, meaning people who comment on Fashion and Styles (556 people) are most likely to mention keywords in Places and Architecture, and in Art, Design, and Photography. Furthermore, Beauty, Sports, and Wellness is also highly correlated with Fashion and Styles, revealing that a major group of people interested in Beauty, Sports, and Wellness (413 people) tend to be interested in Fashion and Styles (556). Indeed, the graph shows the linkage lines between commenters of the two topics of interest as mostly thicker lines, indicating a high intensity of interest of people who repeatedly mentioned the two topics more than once or used more than one keyword from both topics.

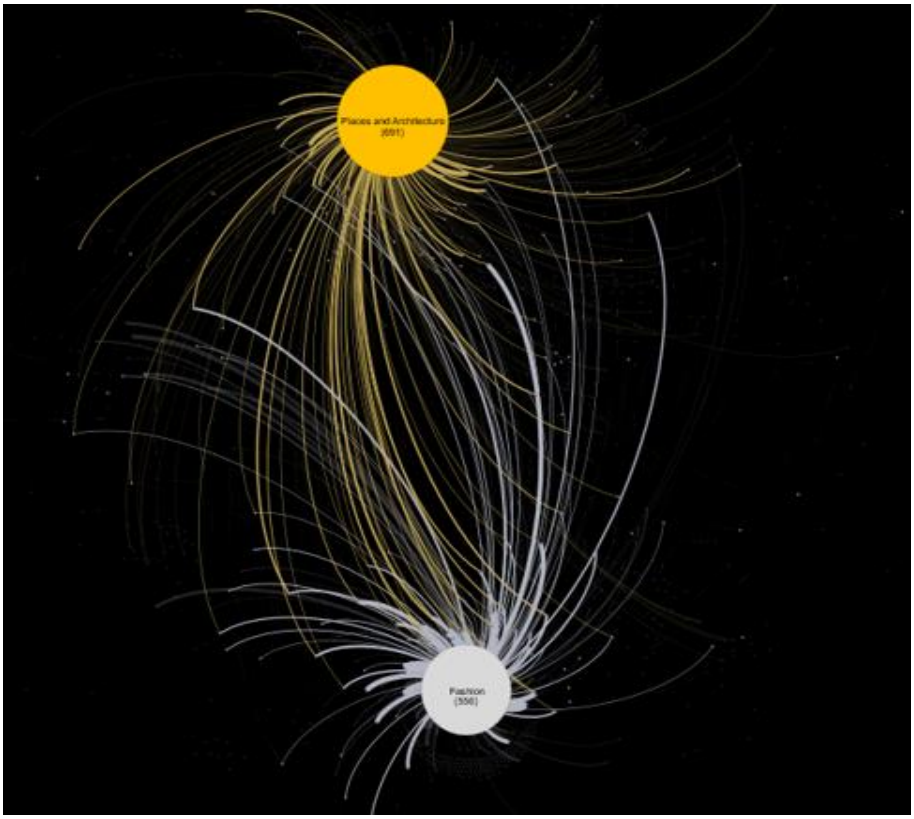


Figure 26. Correlation between Fashion and Places and Architecture.

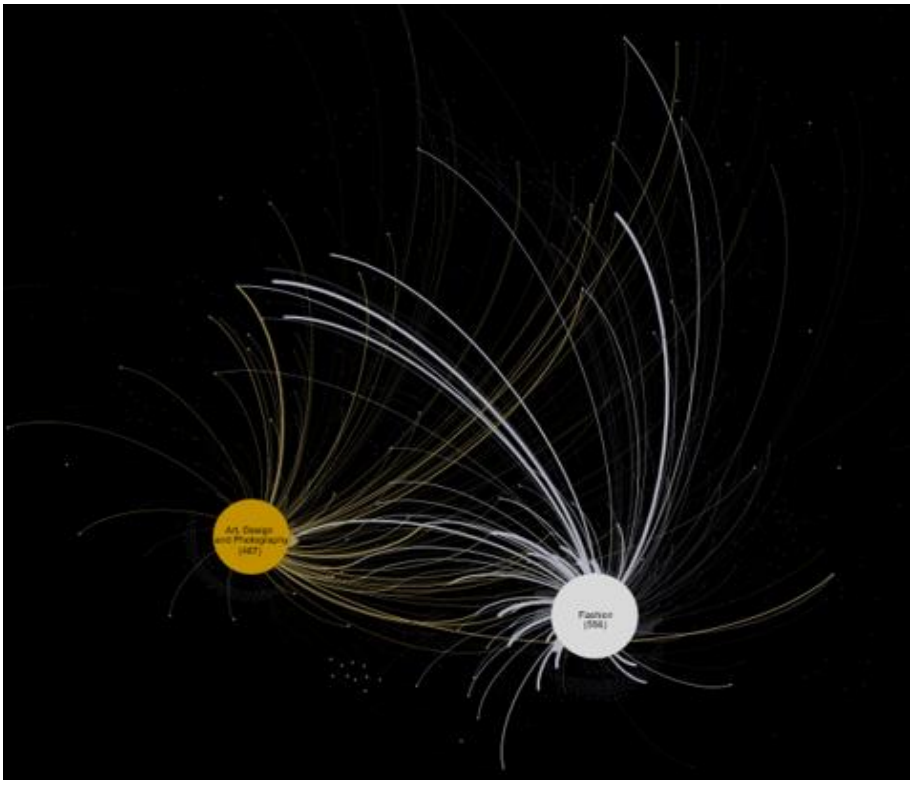


Figure 27. Correlation between Fashion and Art, Design, and Photography.

The network of interest reveals a significant relationship between interest topics and people; it shows both abundant activities in some interest topics and a lack in activities in others. Notably, keywords from Social and People are most commonly used by the highest number of people as these keywords can be used alone to mention friends or family members, and are often used along with keywords in other interest topics especially Events and Entertainment.

Significantly, Food and Drinks and Events and Entertainment are connected by a similar number of people. While Food and Drinks is the interest topic which provokes the highest use of keywords among all comments, Events and Entertainment encourages interaction between people such as friends and family, as the keywords from Events and Entertainment often appear alongside keywords in Social and People. Thus, Food and Drinks and Events and Entertainment are the leading topics of interest which stimulate most activities and interaction at Sylvia Park Shopping Centre.

Another important topic is Places and Architecture, even though it does not contain the highest number of keywords and is not connected to the highest number of people. Places and Architecture is significantly used along with keywords in many other topics to provide important information about locations, places, and features of architecture where the activities or events in other topics take place or feature.

Moreover, Fashion and Styles has a similar target group of people who strongly share the same interests in Beauty, Sports, and Wellness. Yet, Fashion and Styles emerges beyond its own topic by expanding some discussions over Art, Design, and Photography and Places and Architecture, perhaps mentioning and sharing locations, design of the clothes, or even the architectural design of the retail spaces. As modern retailers give more attention to the store design to attract more people, this generates more promotion through visitors' social media.

Animals, the interest topic with the lowest frequency of mentions, has most of the keywords highly influenced by other interest topics like Events and Entertainment. Perhaps the limited activity in this topic is related to the strong pet access limitations.

2.2.1.5 Timeline of Comments and Keywords

Month	Average words per comment
January	8
February	8
March	8
April	7
May	9
June	7
July	9
August	8
September	10
October	8
November	9
December	10

Figure 28. Average words per comment in each month.

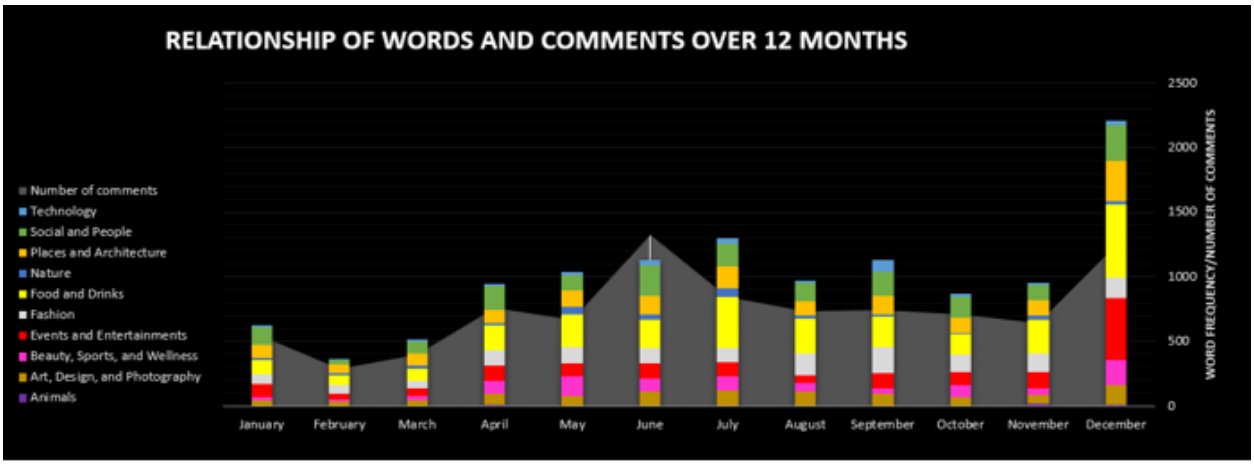


Figure 29. Relationship of keywords and comments over 12 months.

2.2.1.6 Relationship Between Comments and Keywords over 12 Months

The comments and keywords are segregated over 12 months to investigate the timeline of the comments in relation to the changes in the level of interest, based on the identified keywords, and understand the patterns of the interest topics throughout the year.

The number of comments in each month generally increases in the first half of the year and is more stable in the last 5 months before December; and the number of words used per comment ranges between 7 and 10 words on average. Also, the number of identified keywords has an overall increasing trend which mostly follows the changes in the number of comments. However, there are significant differences detected in the patterns of comments and keywords. The timeline of comments indicates two significant peaks, one in June and another in December, where the number of comments exceeds the general trend. The timeline of keywords has a slight peak in July and extreme peak in December, indicating a higher usage of keywords in July and an extreme use of keywords in December. Yet, the number of identified keywords, which is usually higher than the number of comments, falls below the peak of comments in June.

Furthermore, the comments and keywords are further analysed and visualised in two proportional bar graphs including: the proportion of comments containing keywords versus comments without keywords, presenting the changes in proportion of the comments which contain identified keywords in each month; and the proportion of unidentified words versus identified keywords, showing the changes in proportion of the total identified keywords of each month.

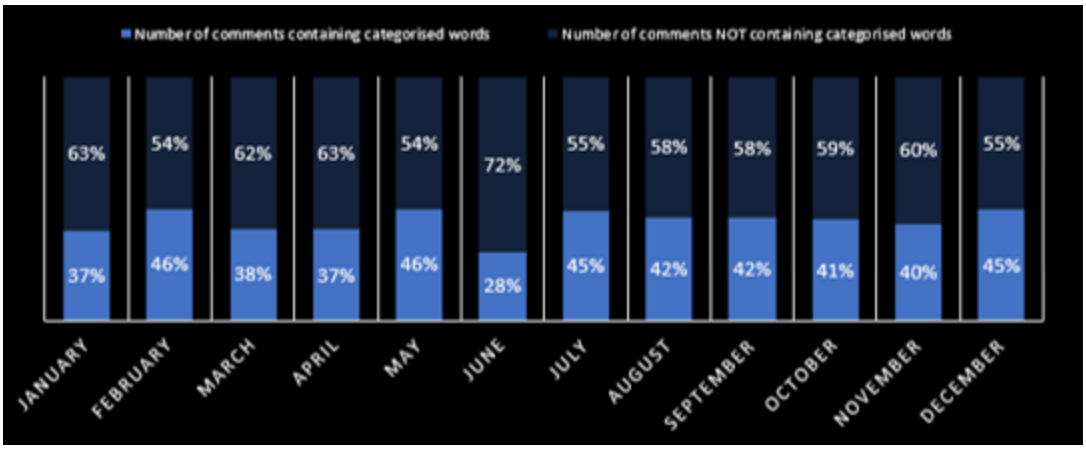


Figure 30. Proportion of comments containing keywords vs NOT containing keywords.

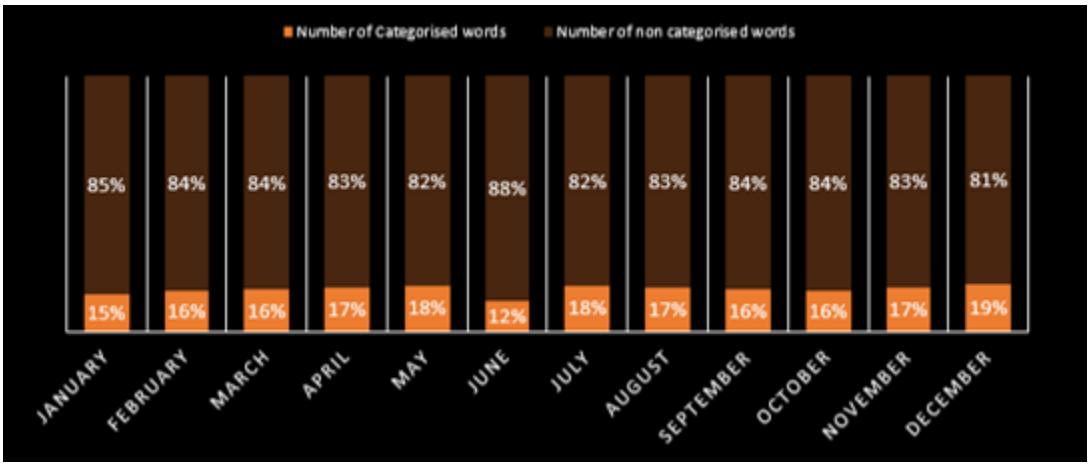


Figure 31. Proportion of categorised keywords vs non-categorised words.

The analysis reveals that February (46 %) and May (46 %) have the highest percentage of comments which contain identified keywords, followed by July (45 %) and December (45 %). Comments in December (45 %) contain the largest percentage of identified keywords (19 %), followed by May (18 %) and July (18 %), then April (17 %), August (17 %), and November (17 %).

Notably, the percentages of comments and keywords in most months correspond to each other but they have slightly different patterns from month to month, revealing various communication behaviours. For example, May has the highest percentage of comments with keywords (46 %) and a relatively high percentage of identified keywords (18 %).

February is another month with the highest percentage of comments with keywords (46 %), but it has a lower percentage of identified keywords (16 %) compared to other months; thus, there is a low usage of keywords in each comment. Moreover, December posts stimulate a high percentage of comments with keywords (45 %) and the highest percentage of identified keywords (19 %), meaning the highest usage of keywords per comment is in December.

Additionally, June shows the highest peak of the number of comments in the previous graph, but it has the lowest percentage of comments with keywords (28 %) and the lowest percentage of identified

keywords (12 %). The extremely low use of identified keywords in the comments may be because the comments are not focused on any of the interest topics

Although the investigation on the timeline of comments and keywords reveals notable differences in the number and percentages of comments and the level of keywords used in each month, a further exploration on significant dates and events throughout the year has been carried out to find possible influences causing the changes in each month.

2.2.1.7 Timeline of Events and Interest Topics Based on Identified Keywords

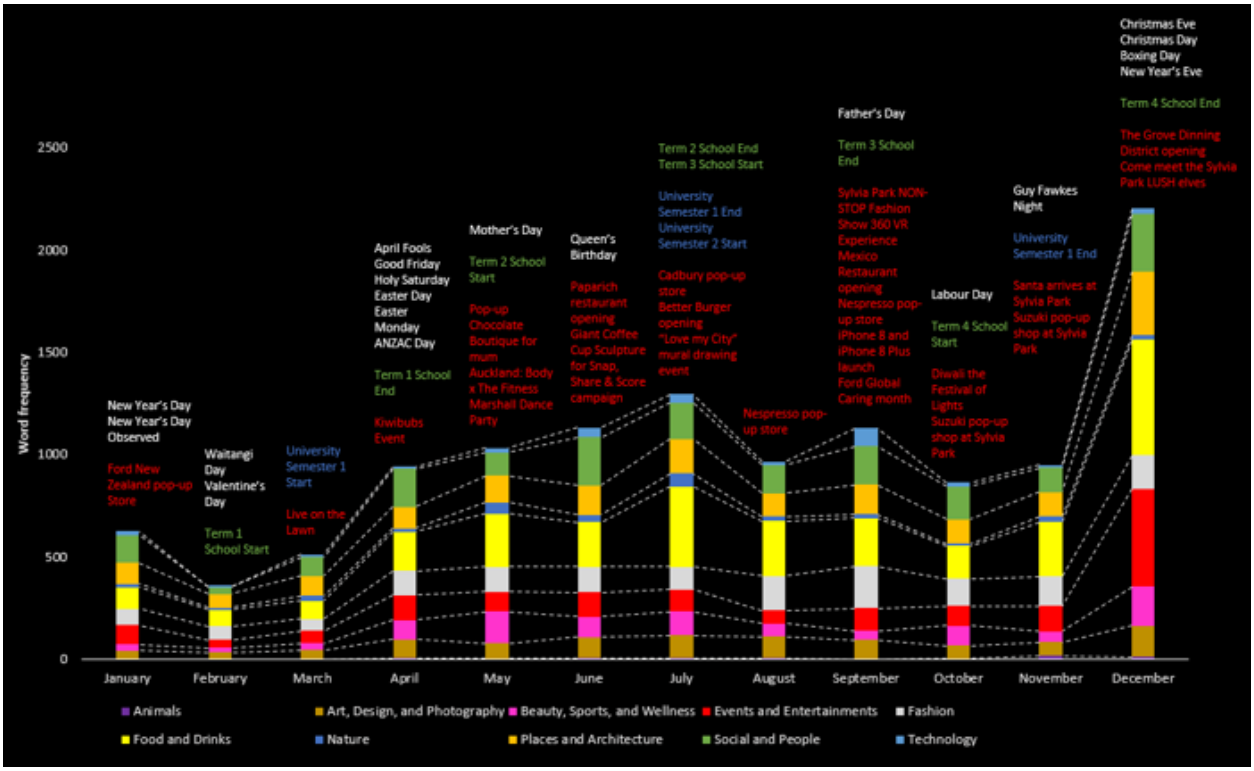


Figure 32. Timeline of interest and significant dates over 12 months of 2017.

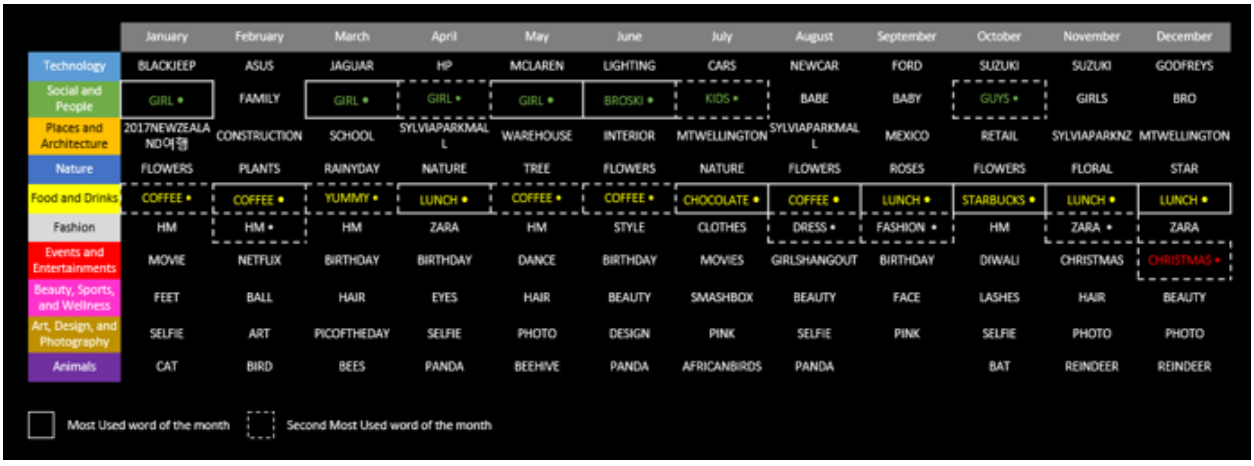


Figure 33. Most used words of each topic over 12 months of 2017.

This exploration focuses on the timeline of the topics of interest based on the identified keywords, visualised through a monthly bar graph showing the level of each interest topic, along with the most used keywords of each topic in each month, and highlighting the top two most used keywords of the month. In particular, it indicates the events including public holidays, school terms and holidays, university semesters and holidays, and public events held at Sylvia Park Shopping Centre throughout 2017. By separating the identified keywords of into the months they occur in, it presents the level of interest in each topic over 12 months and suggests possible causes of the changes in the interest level in relation to the events which occurred in each month.

Noticeably, Food and Drinks is the most active topic throughout the year, represented through the bold thickness in the bar chart and the keywords of Food and Drinks being the most or second-most used every month. The most common keywords are *coffee* and *lunch*; *coffee* appears as the most used keyword of the topic in 5 months (January, February, May, June, and August) and the most used keyword of the month in 2 months (February and August); *lunch* is the most used keyword of the topic as well as the most used keyword of the month in 4 months (April, September, November, and December). Perhaps, *coffee* and *lunch* are the most popular activities at Sylvia Park Shopping Centre among people interested in *Food and Drinks* due to the newly developed food precinct. However, the Nespresso Pop-Up Store in August could be another factor which makes *coffee* the most used keyword of that month. Similarly, another most used keyword of the topic and month such as *chocolate* could be influenced by the Cadbury Pop-Up Store in July. On the other hand, *lunch* tends to appear as the most used keyword of the topic and month during holiday periods as school holidays in April, September, November and December. Overall, Food and Drinks is the leading topic in stimulating communication between people that can be influenced by temporary events and holiday seasons, indicating a hint of higher activity levels during those periods.

Another active topic is Social and People, in which the keywords appeared as the most and second-most used of the month in 7 months (January, March, April, May, June, July, and October), with *girl* the most commonly used keyword of the month in 4 out of 7 months. Most of the keywords signify the younger generation, e.g., *girl, bro, broski, kids* etc., and tend to be prominently used around the start and end of school holidays.

Meanwhile, Fashion and Styles keywords are the second-most used keywords in 4 months: February, August, September, and November. These keywords include general terms such as *dress* and *fashion,* and clothing brands as *Zara* and HM. Their high volume is possibly influenced by their promotion and sale periods prior to (November) and after (February) the holiday season.

However, the second-most used keyword of September, *fashion*, is most likely to be affected by the fashion event Sylvia Park NON-STOP Fashion Show 360 VR Experience. Yet, the volume of the total Fashion and Styles keywords in September is also the highest of the year; therefore, it is probable that the fashion event encourages higher communication in the Fashion and Styles topic.

In addition, Events and Entertainment appear as the second-most used keywords in December and the keyword is *Christmas,* which has one of the largest word frequencies among all keywords. Indeed, Christmas has a strong impact on the communication between people as the number of keywords in all topics, especially Food and Drinks and Events and Entertainment, increase significantly during this period. Nevertheless, the influence of events on the level of interest and number of keywords stimulated is reflected through other keywords—for example, *Diwali,* from Events and Entertainment, influenced by Diwali, the Festival of Light in October; *Suzuki,* from Technology, originated from the Suzuki Pop-Up Store in November; and *dance,* in Events and Entertainment, influenced by Auckland: Body x The Fitness Marshall Dance Party in May.

Social and People is mainly associated with school holidays and activities involving the younger generation; Fashion and Styles is impacted by sale seasons and fashion-focused events; Events and Entertainment largely revolve around the Christmas season; and other topics like Technology; Places and Architecture; Beauty, Sports, and Wellness; Nature; Animals; and Art, Design and Photography are specifically provoked by events focused around their topics. Food and Drinks, which generates the highest number of keywords, is stimulated by people constantly referring to Food and Drinks in various occasions such as restaurant openings, temporary pop-up stores, and meals for celebrations or festive seasons. Certainly, public holidays, school holidays, and public events have a significant impact on the visitors' experience and participation as well as influencing the level of their communication and representation of their interest, activities, and spaces. As a result, temporary events of any interest topics held at Sylvia Park Shopping Centre stimulate higher communication with more identified keywords. Therefore, it is probable that events can be a solution to encourage activities in neglected areas of interest or provoke the utilisation of dead spaces especially during holiday periods.

2.2.2 Semantic Analysis of Communities

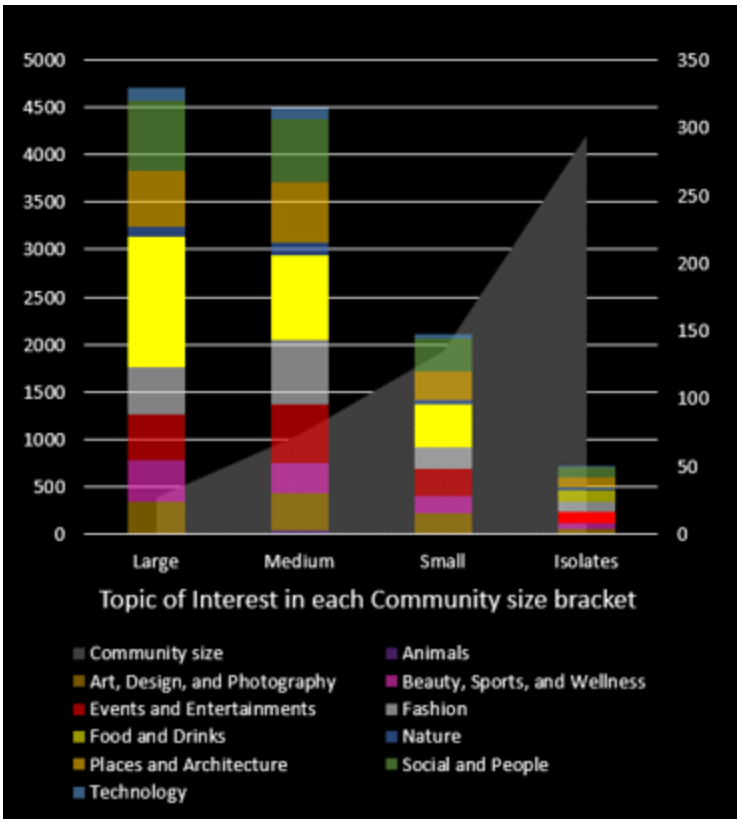


Figure 34. Topics of interest in each community size bracket.

2.2.2.1 Keywords and Interest Topics in Each Community Size Bracket

This analysis focuses on exploring the keywords and identifies the topics of interest of the communities in different size brackets (large, medium, small, and isolates) detected in the network analysis. This reveals the distribution and differences of interests in the communities of each size bracket.

The number of communities in each size bracket, from large to isolates, is inversely proportional to the number of keywords and interest level. The large-size bracket includes the smallest number of communities (27) but contains the largest number of identified keywords (4,737) and shows the highest volume of interest levels. In contrast, the brackets with the highest number of communities, the isolates (295) and small-size (135), include the smallest number of keywords, respectively with 721 and 2,131 keywords.

Although there are a higher number of communities in the medium-size bracket (74) than the large-community size bracket (27), they have a similar number of keywords as well as a similar number of total people: 4,522 keywords identified from 50,335 people in 74 medium communities, and 4,737 keywords from 51,308 people in 27 large communities. Indeed, people are more connected within the 27 large communities as opposed to the similar number of people which is more distributed in the 74 medium communities.

2.2.2.2 Distribution of Interest Topics over Different Community Size Brackets

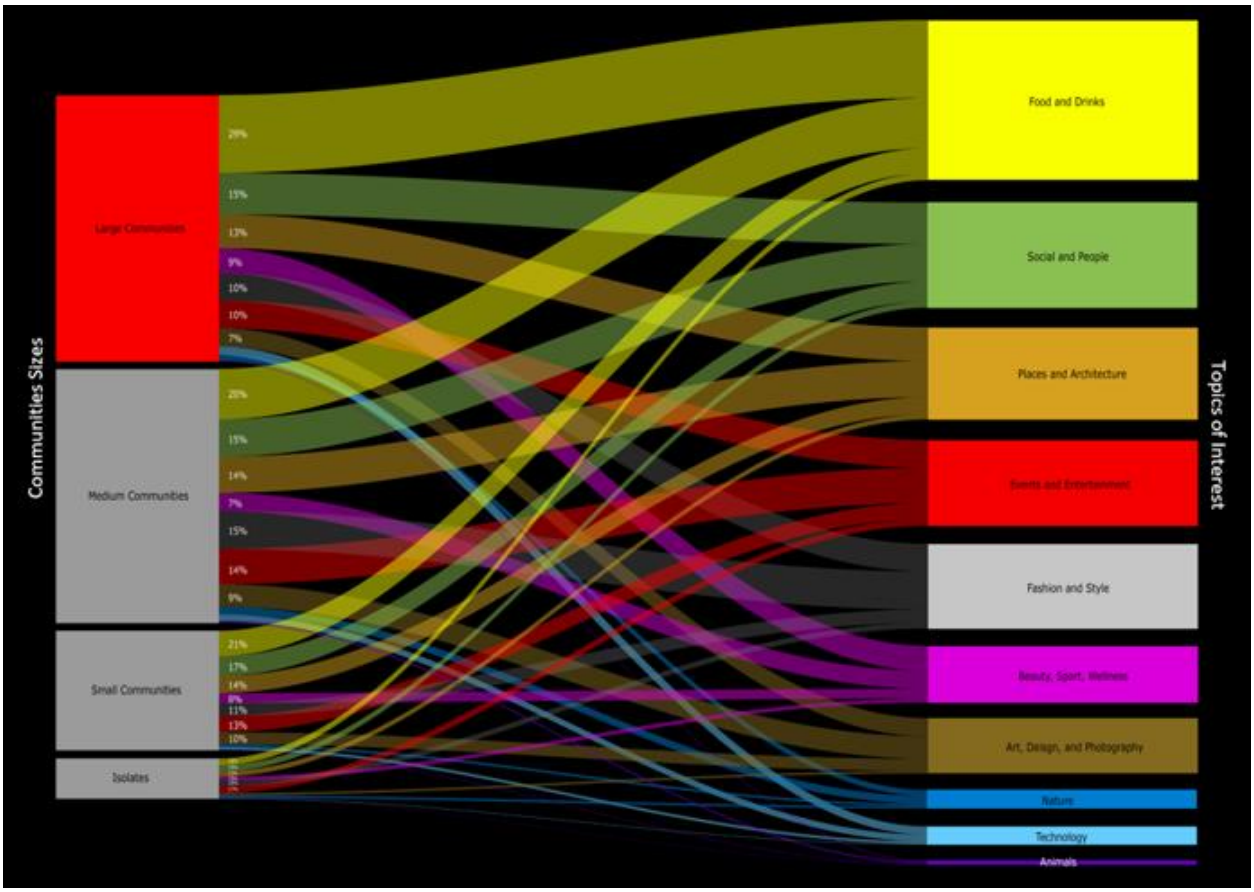


Figure 35. Distribution of interest in each community size bracket.

While the interests of isolates, small, and medium communities are distributed among all topics, Food and Drinks is the most dominant topic in all community size brackets. Large communities' interest in Food and Drinks (29 %) is more than a quarter of all interest topics. Yet, some topics in the large communities like Places and Architecture (13 %), Events and Entertainment (10 %), and Art, Design and Photography (7 %) are smaller than the same topics in the medium communities (Places and Architecture, 14 %; Events and Entertainment, 14 %; Art, Design and Photography, 9 %).

Significantly, the interest topics within the 27 large communities are shared among closely connected groups of people, for which the main interest is largely concentrated on Food and Drinks. Thus, these 27 largest communities can be further investigated to find the causes of the major interest in Food and Drinks as well as studied to understand the characteristics of each large community through their activities (comments, likes, and posts) and distribution of interest topics.

2.2.2.3 Activities, Interests, and Characteristics of 27 Largest Communities

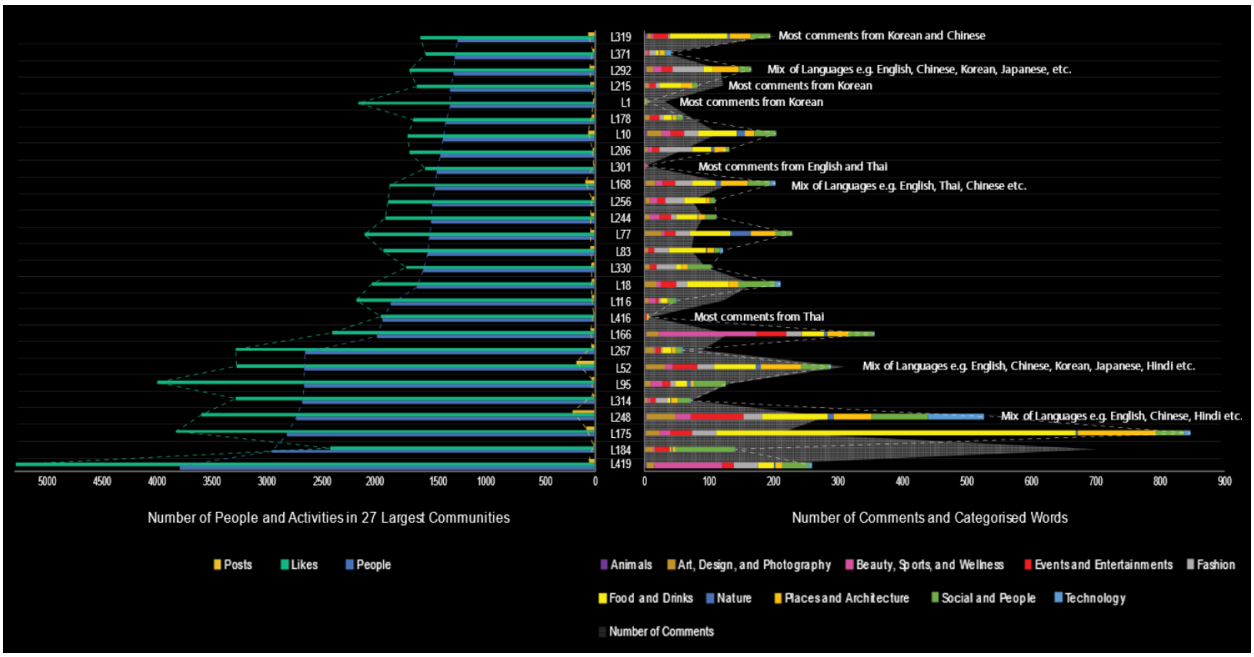


Figure 36. Comparison of activities and interest level of 27 largest communities.

2.2.3 Activities and Interest Topics of 27 Largest Communities

The 27 largest communities are individually analysed by looking at the number of people within the communities and the amount of activities including number of posts and likes in comparison to number of comments and level of interest based on number of identified keywords.

2.2.4 General Community Trend

Generally, the level of interest follows the size of the community (number of people) and the level of interaction between people (comments and likes); thus, the higher the level of interaction, the higher the level of interest. This can be seen in communities L319, L292, L10, L168, L77, L18, L166, L248, L175. However, community size, interaction level, and interest level do not always depend on the number of posts—for example community L52 has a high number of posts but does not necessarily have the highest number of people or level of interaction and interest.

2.2.5 Communities with Foreign Languages

Moreover, there are various characteristics specific to each community and some communities are distinguished from the general trend—for example, community L1, L301, L416, L184, L419 etc. Communities L1, L301, and L416 have a similar characteristic: they have a really small number of comments which are made in foreign languages, resulting in a small number of identified keywords and

very low interest level, although they have a higher number of people and number of likes than some other communities, especially community L1 which has a peak of likes higher than many larger communities. This type of community is connected by a close group of foreigners writing in the same languages and the communities are formed through likes rather than comments. Similarly, other communities like L319, L292, L215, L168, L52, and L248 contain comments in English and a variety of foreign languages. These communities have a higher number of comments, and, as a result, a higher number of identified keywords, showing higher interest levels. Thus, these communities are formed by a more balanced level of interaction of comments and likes.

The comments in foreign languages can be assumed to be made by either tourists or immigrants. The first three communities, L1, L301, and L416, are most likely to be a group of tourists with a lot of followers as there are a limited number of comments in foreign languages but a large number of likes. The second type of community, like L319, L292, L215, L168, L52, and L248, is more likely to be immigrants commenting on the same topics, as there are a mix of languages, including English, used in the comments.

2.2.6 Distinctive Communities

Significantly, community L184 is the second largest community and has the highest number of comments but it has a very low number of keywords and a drop in the level of activities of posts and likes. Thus, this community is formed through a large number of comments rather than likes. It is probable that the posts which stimulate the comments involve public figures rather than other interest topics, as the dominant interest of the community is Social and People. In contrast, the community L419 is the largest community with the highest number of likes but it has fewer comments and lower interest levels than some smaller communities. Thus, the largest community L419 is largely connected through likes of people who share the same interest, presumably Beauty, Sports, and Wellness, as it is the most dominant topic of the community. Another notable community is L175, which has the highest level of interest— the highest number of keywords, significantly in Food and Drinks, is identified from a moderate number of comments. Nevertheless, this community has a high number of interactions especially through likes, thus, community L175 is formed through likes and comments with a core interest in Food and Drinks.

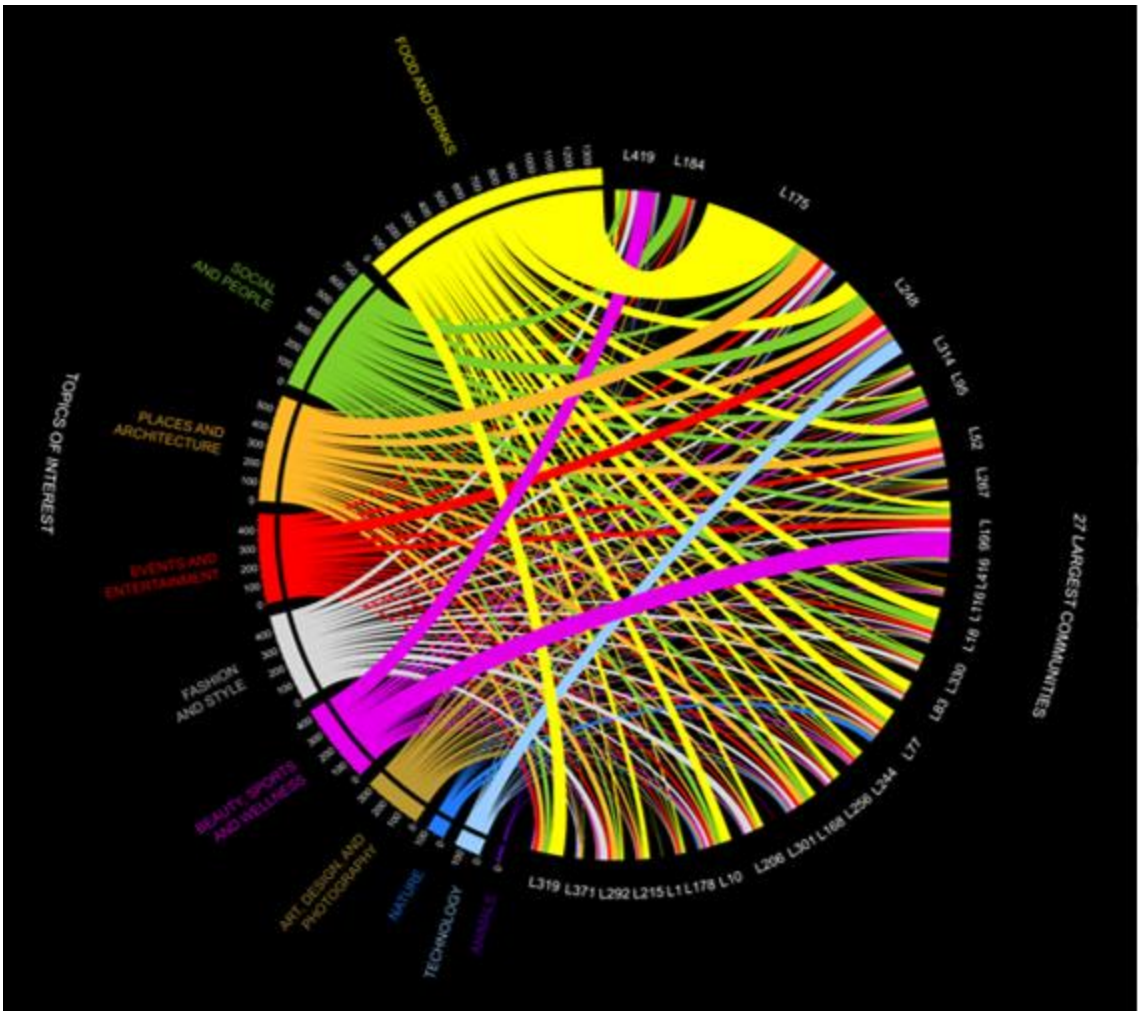


Figure 37. Topics of interest distribution of 27 largest communities.

2.2.7 Distribution of Interest Topics over 27 Largest Communities

Noticeably, the interests in each of the 27 large communities are spread over various topics but some communities have a strong concentration of interest on certain topics. Therefore, distribution of interest topics and communities' dominant interests are further explored and visualised in a circular diagram. This presents two significant interest topics—Social and People and Food and Drinks, which appear in almost every community.

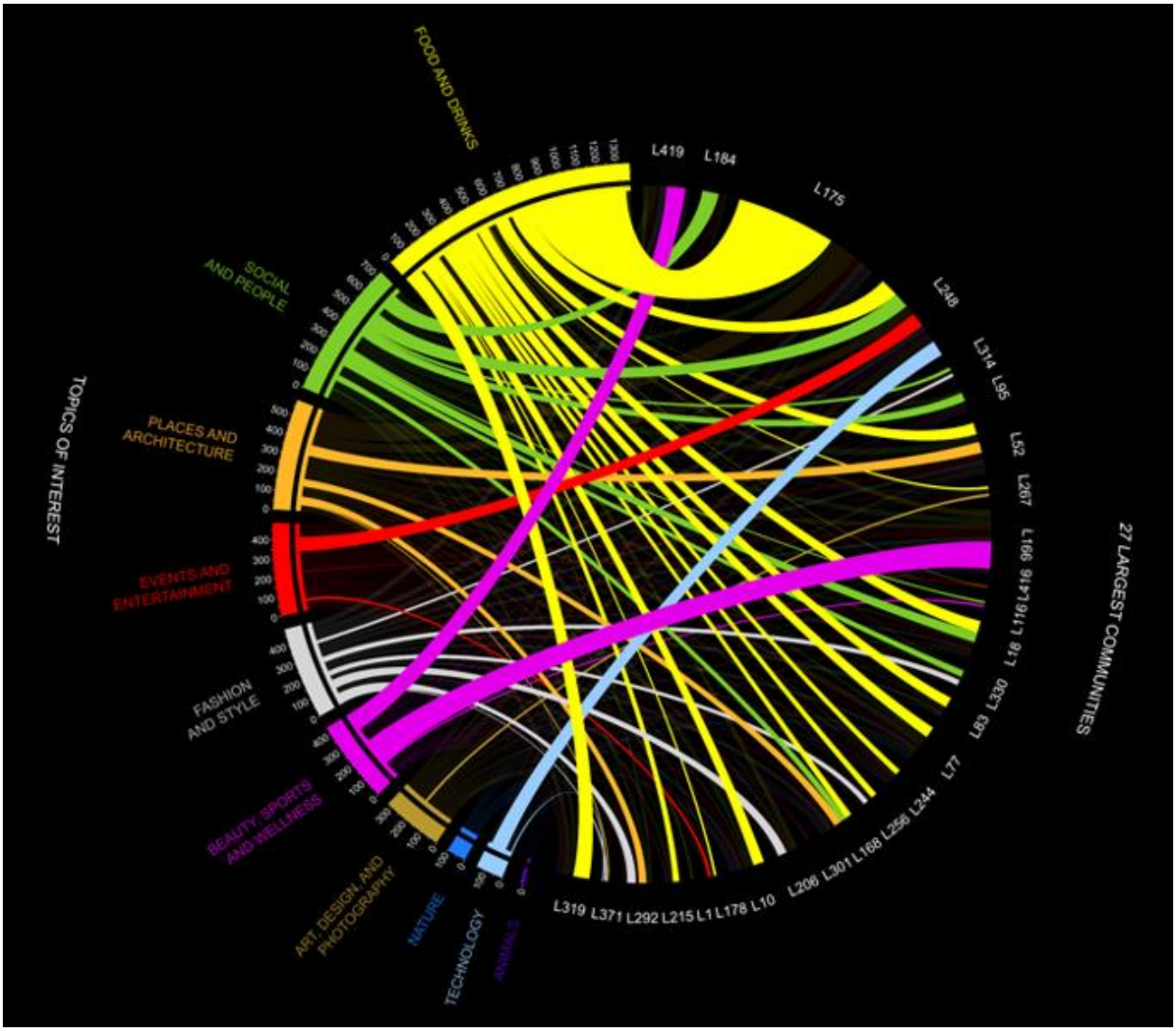


Figure 38. Isolation of dominant interest topics of 27 largest communities.

2.2.8 Significant Distribution of Social and People and Food and Drinks in the 27 Largest Communities

The interest in Social and People is presented in 26 communities and is the most dominant topic of eight communities including L184, L248, L314, L95, L116, L18, L330, and L1. Usually, Social and People is a joint-dominant topic with others, but it is the most dominant topic in community L184. Thus, community L184 is formed around a topic of a group of people or a person, perhaps a public figure due to the large amount of attention received from likes and comments.

On the other hand, interest in Food and Drinks is distributed over 23 communities and is the most dominant topic of 14 communities including L175, L248, L52, L267, L18, L83, L77, L244, L168, L256, L10, L178, L215, and L319. Significantly, Food and Drinks is the topic most commonly presented as the communities' dominant interest; sometimes it is a joint-dominant interest with other topics but often it is the sole dominant interest topic of the community. Community L175, especially, has the largest level of interest in Food and Drinks. It stimulates the highest number of identified keywords and is the most

extreme interest among all other communities' interest in every topic. Indeed, the high intensity of interest in Food and Drinks appearing in the large-community size bracket is influenced by the extreme interest in Food and Drinks of community L175—the third biggest community of the network.

2.2.9 Distribution of Other Interest Topics in the 27 Largest Communities

Additionally, Places and Architecture receives the largest interest in community L175 after Food and Drinks, suggesting a relationship between the interest in Food and Drinks and Places and Architecture. Correspondingly, the earlier analysis reveals that keywords in Food and Drinks are highly correlated with keywords in Places and Architecture, thus, a high level of interest in Food and Drinks influences a higher level of interest in Places and Architecture in community L175.

Nevertheless, Beauty, Sports, and Wellness is the most dominant topic of communities L419 and L166. Although community L419 has a lower overall interest level than community L166, these two communities share the exact number of interest topics.

Furthermore, although the interest in Events and Entertainment and Technology is the predominant in community L248, this community has an almost equal distribution of four major interest topics (Food and Drinks; Social and People; Events and Entertainment; and Technology).

While Fashion and Styles is not a powerful topic with a remarkable amount of interest, it is the most dominant topic in six communities: L314, L330, L256, L206, L292, and L371. It often appears as a dominant topic along with Social and People, Places and Architecture, and Food and Drinks, except in community L206 where Fashion and Styles is the only dominant topic of the community. Hence, the community is formed around a particular interest in Fashion and Styles.

The analysis of the distribution of interest topics and communities' dominant interests reveals an impact of interest on communities' formation with distinct characteristics in different communities. Remarkably, some communities have a strong concentration in one specific interest topic, suggesting the dominant interest which is shared by the people within the communities, while other communities have a few interests distributed over several topics, and they share an equal interest between these topics. Additionally, different communities which are not connected can have the exact same interests and dominant interest topic. Perhaps these communities are formed by different types of influence such as a special event or an influential person.

2.2.10 Comments Distribution with Monthly Events and Interests Timeline of Each Community

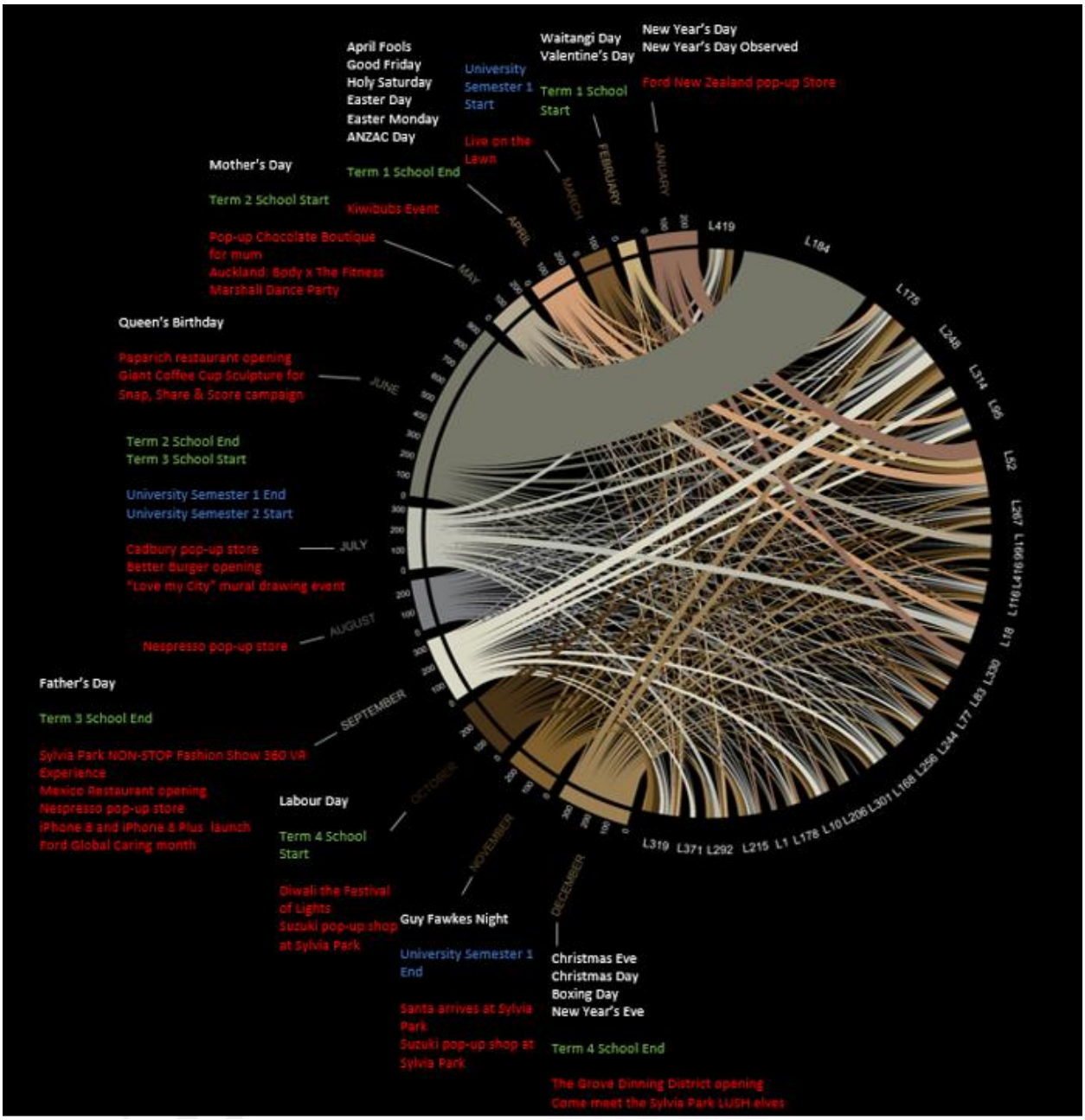


Figure 39. *Comments Distribution over 12 months of 27 largest communities.*

Exploring the monthly distribution of the comments in each community enables a trace of significant events which may have a tendency to initiate the interests within the communities. In this case, the dates of the posts which the comments are intended for, are assigned as the time stamp instead of the actual commented dates.

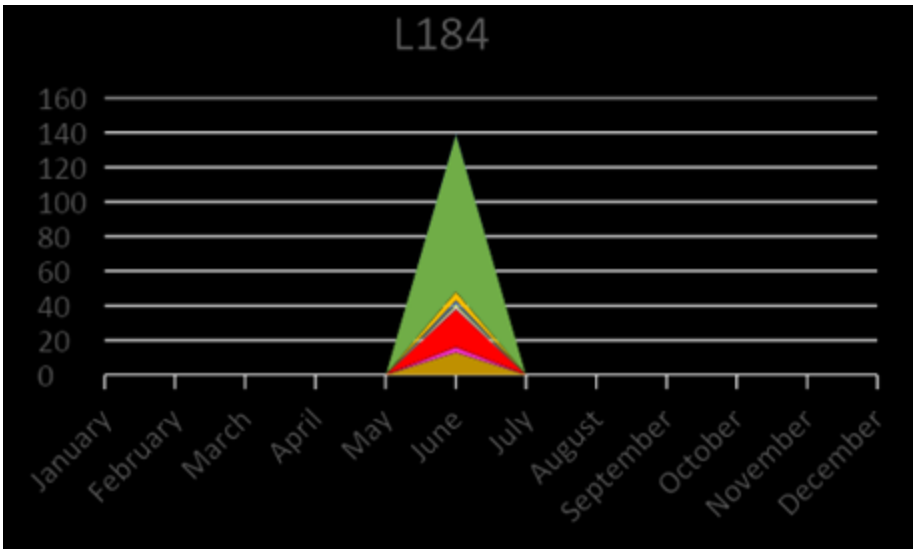


Figure 40. Timeline of interest of community L184.

2.2.11 Communities' Interests not Influenced by Events

Notably, community L184 is the only community in which all the comments are directed toward the posts in June; therefore, there are interests based on keywords detected in the comments shown only in June. The significant dates in June are Queen's Birthday and Food and Drinks events, but the community's dominant interest is largely associated with Social and People with very little interest in Food and Drinks. Hence, community L184's dominant interest is formed around specific posts in June and is influenced by the posts' content rather than events occurring in June.

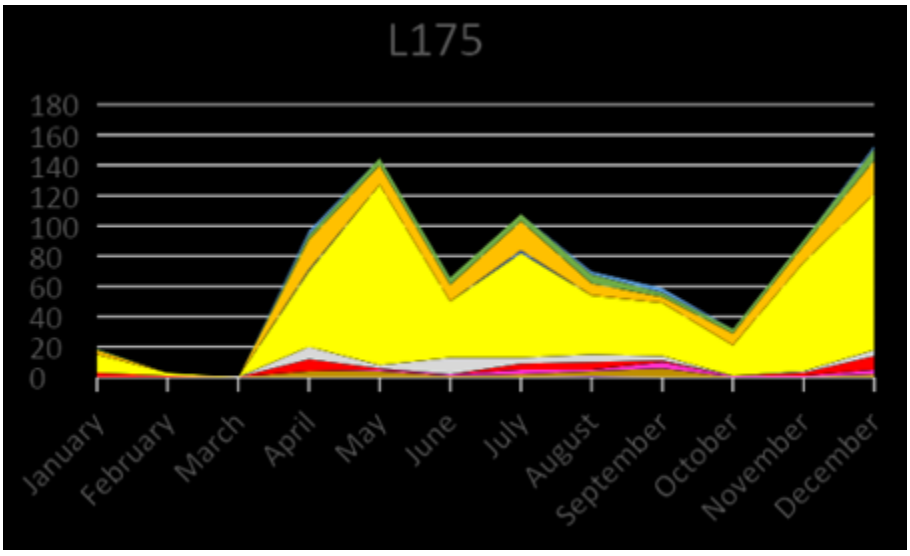


Figure 41. Timeline of interest of community L175.

2.2.12 Communities' Interests Driven by Events Related to a Specific Topic

On the other hand, community L175 has comments distributed over 11 months throughout the year, except for an absence of comments in March. Although the number of comments is low in the first 2 months—January and February, the number of comments from April onward are relatively similar before

a rapid rise of comments in December. Looking at the distribution of interests over the year, Food and Drinks, which is the most dominant topic of community L175, is the leading topic of the community in every month. Nevertheless, there is a relationship between the interest in Food and Drinks and events held at Sylvia Park Shopping Centre; the peaks of interest in Food and Drinks in May, July, and December align with Mother's Day and the Pop-Up Chocolate Boutique for Mum in May, the Cadbury Pop-Up Store and the opening of Better Burger restaurant in July, and the opening of The Grove Dining District in December. Thus, community L175 is largely driven by the interest in Food and Drinks, and events involving Food and Drinks held at Sylvia Park Shopping Centre have a significant impact on the communication between people within community L175.

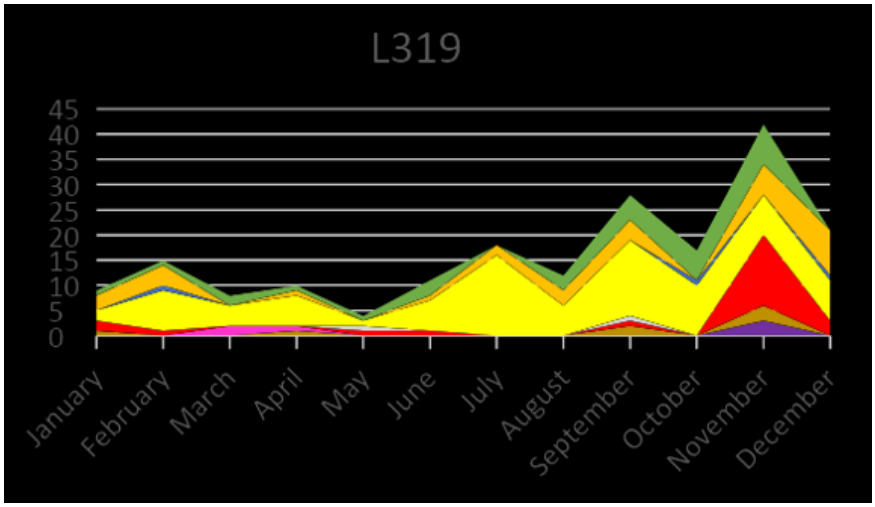


Figure 42. Timeline of interest of community L319.

2.2.13 Communities' Interests Affected by School Terms and Holidays

Similarly, community L319 has comments distributed in every month and Food and Drinks is the most dominant topic of the community. The interest level in Food and Drinks of community L319 is much smaller than community L175, but there are peaks of interests which are aligned with the events occurring at the mall. The peaks of interest in Food and Drinks in July and September parallel the Cadbury Pop-Up Store and the opening of Better Burger restaurant in July, and the opening of Mexico restaurant and the Nespresso Pop-Up Store in September. There is an additional peak of interest in Events and Entertainment parallel to the Christmas period and Boxing Day events in December. Noticeably, the months with peaks of interest tend to be during school holidays like the end of Term 2 in July and the falls of interest levels are most likely to happen when school terms start as in February, May, August, and October. This reveals that community L319 is mainly influenced by events of Food and Drinks but also affected by other big events happening at Sylvia Park Shopping Centre like Christmas as well as school terms and holiday seasons; perhaps the members of community L319 are mainly students.

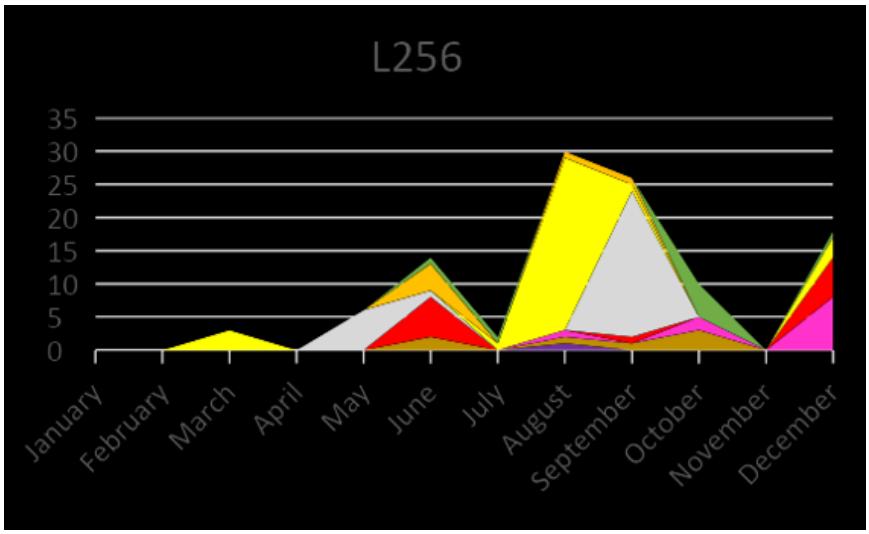


Figure 43. Timeline of interest of community L256.

2.2.14 Communities' Interests Affected by Different Events in Various Topics

Furthermore, community L256 has comments directed toward posts in 8 months: March, May, June, July, August, September, October, and December. The community shares various topics of interests, of which the most dominant topics are Food and Drinks and Fashion and Styles, influenced by the events held at the mall. The highest rise of interest in Food and Drinks in August parallels the Nespresso Pop-Up Store, and the rise of interest in Fashion and Styles occurred during the Sylvia Park NON-STOP Fashion Show 360 VR Experience in September. The two most dominant topics are not always the leading topics and do not continuously present every month, instead, they only appear a few times with the peaks in the 2 months. Thus, community L256 is not driven by specific interests but sometimes influenced by events in various topics.

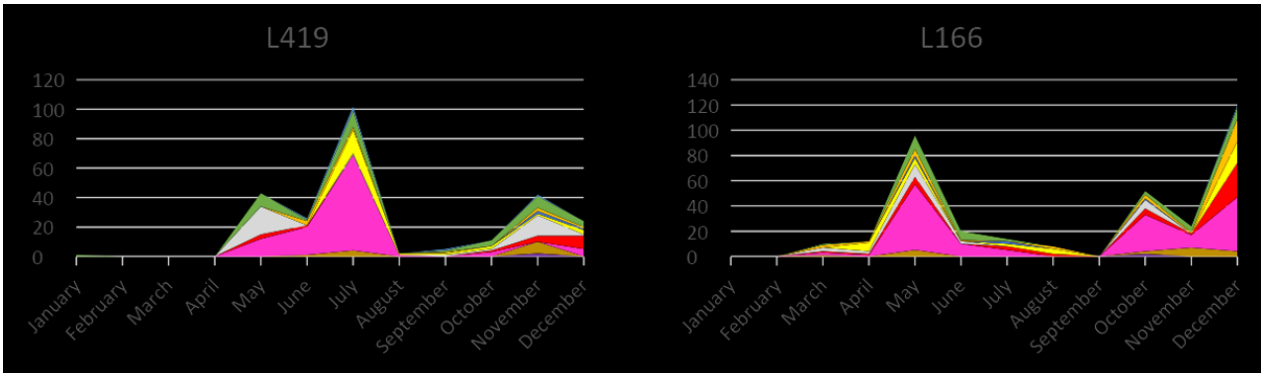


Figure 44. Timeline of interest of community L419 and L166.

2.2.15 Comparison of Communities with the Same Interests

While communities L419 and L166 are driven by the same interest in Beauty, Sports, and Wellness, they are totally disconnected from each other. Both communities L419 and L166 have comments distributed in different 9 months; community L419 has comments in January, May, June, July, August, September, October, November, and December, of which the highest number of comments is in July; and community L166 has comments in March, April, May, June, July, August, October, November, and December, of

which the most comments are in May followed by October and December. Thus, different attractions provoke the increase of comments in different months between the two communities. A high volume of comments in July from community L419 results in a high number of keywords and high interest level, especially in the most dominant topic of the community, Beauty, Sports, and Wellness; however, significant dates in July do not involve events related to the topic. Therefore, the increase of interest in Beauty, Sports, and Wellness during July is not affected by an event, but perhaps by the posts' content or an influential poster.

In contrast, the increase of interests, especially in Beauty, Sports, and Wellness, in May and December within community L166 is parallel to the occurrences of Auckland: Body x The Fitness Marshall Dance Party and Come Meet the Sylvia Park LUSH Elves events held at Sylvia Park Shopping Centre. Although communities L419 and L166 share the same dominant interest, they are separated from each other because they are driven and influenced by different causes. Community L419 is formed by a group of people who share the same interests and interact toward the same posts or posters despite the influence of a special occasion, while community L166 is formed by another group of people who share the same dominant interest topic and the level of interest is stimulated by special events held at Sylvia Park Shopping Centre.

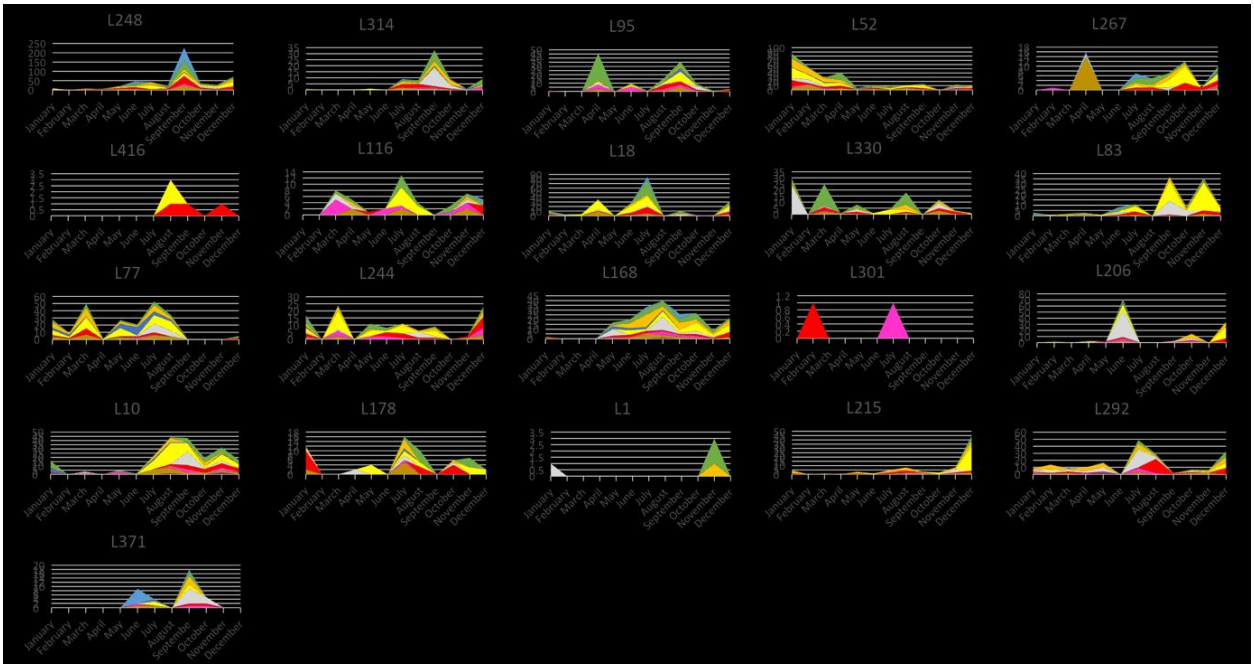


Figure 45. Timeline of interest of other large-sized communities.

Tracing significant dates and events held at Sylvia Park Shopping Centre reveals the impact these events have on the changes of communities' interests. Even though some communities are not affected by the events, there are communities which are strongly influenced by events related to specific interest, especially their community's dominant interest. Otherwise, the communities can be influenced by different events in various interest topics. Additionally, some communities are affected by school terms and holidays as their interest level decreases during school terms and increases during school holidays. Therefore, sometimes, communities which are driven by the same interests can be isolated from each other as they are influenced by different factors.

2.2.16 Interest of the Central People of 27 Largest Communities

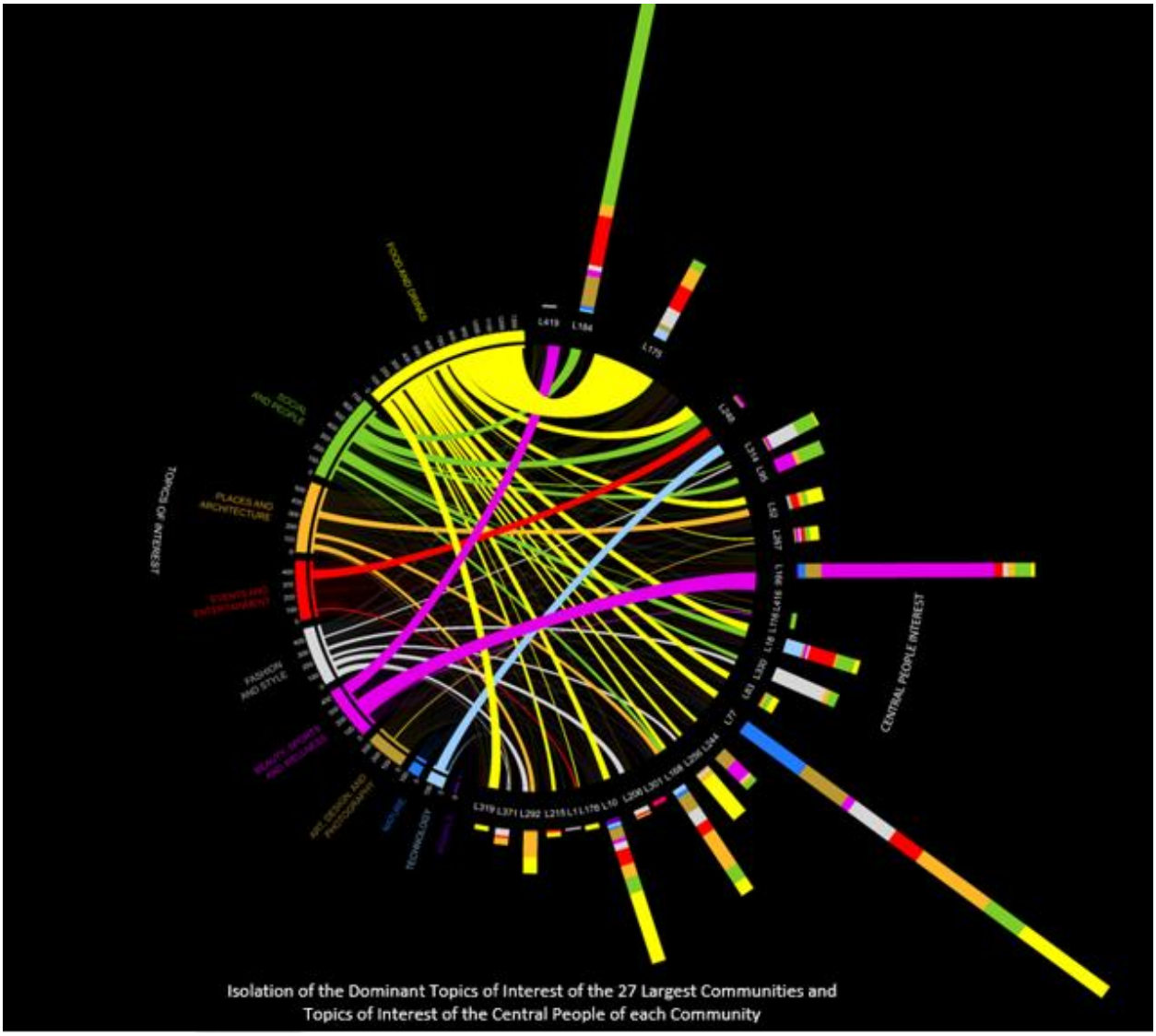


Figure 46. Isolation of dominant interest topics of 27 largest communities and topics of interest of the central people of each community.

The smallest scale of the network is an individual and this individual person can be another factor which influences a community. Every community has a central person who is the most active person and has the highest activities among others in the same community. Investigating the interests of these central people in relation to the communities' interests allows a discovery of their influences to the communities.

2.2.17 Central Persons Who Have Different Interests from Their Communities' Dominant Interest

The central people from five communities, L419, L175, L248, L416, and L244, have no influence on their communities, as they do not share their dominant interests with majority of the people within the communities. For example, the dominant topic of community L419 is Beauty, Sports, and Wellness, but the central person of the community has only one interest detected, in Fashion and Styles. Some central

people have a variety of interests but none of their interests align with the dominant topics of the communities—for example the central person of community L175 is interested in Social and People; Places and Architecture; Events and Entertainment; Fashion and Styles; Art, Design, and Photography; and Technology, but the most dominant and significant topic of the community is Food and Drinks.

2.2.18 Central Persons Who Have Some Interests Aligned with Their Communities' Dominant Interest

Some central people from other communities such as L52, L267, L18, L83 and L206 etc. share some interests aligned with the communities' dominant topics but their level of interest is moderate and distributed in various topics. For example—community L52's dominant interests are Food and Drinks and Places and Architecture, but the central person's interests are sequentially spread over Food and Drinks; Events and Entertainment; Places and Architecture; Social and People; Art, Design, and Photography; and Technology.

2.2.19 Central Persons Who Have Influenced Their Communities

Remarkably, the central persons of community L184 and L166 show significant interest topics which align with their communities' dominant interests. The central person of community L184 shows nine interests including Social and People; Events and Entertainment; Art, Design, and Photography; Places and Architecture; Beauty, Sports, and Wellness; Fashion and Styles; Nature; Food and Drinks; and Technology, but the most significant topic is Social and People which is also the community dominant interest.

Similarly, the central person of community L166 shows interests in nine topics including Beauty, Sports, and Wellness; Social and People; Art, Design, and Photography; Events and Entertainment; Places and Architecture; Nature; Fashion and Styles; Food and Drinks; and Animals, and the largest topic, which aligns with the community's dominant interest, is Beauty, Sports, and Wellness. The significantly large number of keywords in Social and People and Beauty, Sports, and Wellness, used by the central persons of community L184 and L166, align them with the dominant interest topics of the communities. These central persons are the leaders of the interest topics, stimulating and influencing other community members to interact on the same topics, thus, Social and People and Beauty, Sports, and Wellness become the most dominant interest topics of their communities.

Central persons of the communities are the most active people who highly interact with others through posts, comments, or likes. Yet, the central persons can be highly active in one form of interaction but low in another. Hence, the lower number of keywords used by a person who interacts a lot by likes but less by comments is often associate to lower interest levels. Although some central persons can be highly active via comments, their comments are not always relevant or do not contribute toward their communities' dominant interests; these central persons have no or few interests aligned with their communities' dominant interest. There are, however, significant central persons in some communities where the dominant interests of the central persons and the communities are aligned, and where the central person leads the stimulation of a high number of keywords of the communities' dominant interest. Thus, these central persons are influential, and they have a significant impact on their communities' interests.

2.2.20 Influences on Community Interests

Investigation of the 27 largest communities reveals various ways communities behave and different causes affecting the patterns of interests in each community. Community members interact with each

other differently from one community to another. While some communities have a high interaction through likes, another can have a high interaction through comments, or be highly active in both forms of interaction. However, if a community has a limited number of comments, the number of keywords detected is also limited, resulting in a low interest level.

Communities are influenced and driven by different factors, of which the three main factors are the interests, significant dates or events, and influential persons. Some communities have dominant interests which connect different members who share the same interest, and sometimes different communities are interested in the same topics, but they are driven by different causes. There are communities which are purely influenced by the posts' content related to a specific interest topic despite the recognition of the posters or events related to the topic.

Other communities, which are interested in the same topic, are influenced by the occurrences of events involving their interest held at Sylvia Park Shopping Centre. Additionally, some communities' interest levels fluctuate according to school terms and holidays, signifying the kind of members in the communities. Nevertheless, communities' interests are sometimes affected by influential persons or the central persons of the communities. An influential person can be someone with a strong interest like food bloggers or beauty bloggers. The members of communities with these influential persons share the same interest as the central persons and are influenced by the central person's activities. On the other hand, the central person can be a public figure, which means the dominant topic of the community is about the central person themselves rather than other interest topics.

2.3 NETWORK DYNAMICS

2.3.1 Dynamic—Full Network

2.3.1.1 Growth in Activity

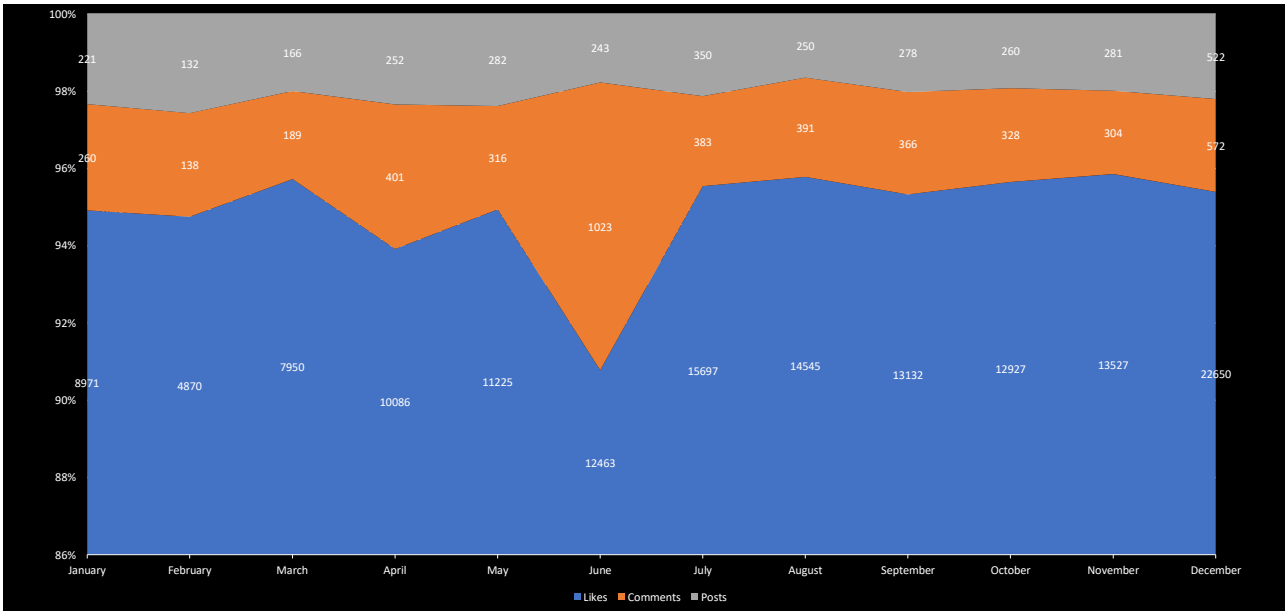
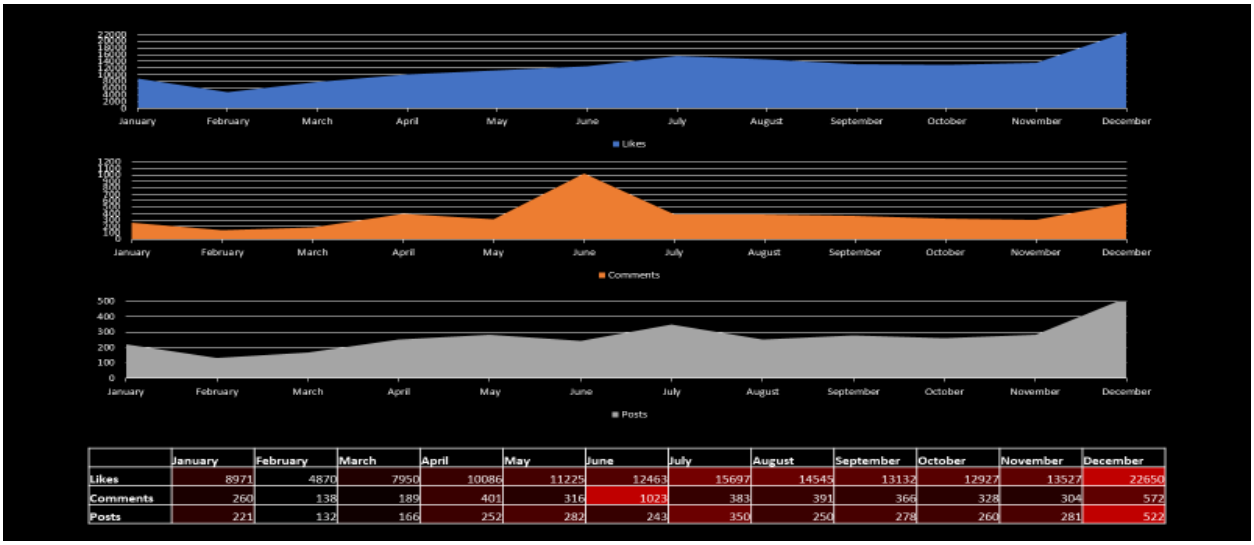


Figure 47. Activity by month relative values.



	January	February	March	April	May	June	July	August	September	October	November	December
Likes	8971	4870	7950	10086	11225	12463	15697	14545	13132	12927	13527	22650
Comments	260	138	189	401	316	1023	383	391	366	328	304	572
Posts	221	132	166	252	282	243	350	250	278	260	281	522

Figure 48. Activity by month by type.

Tracking the overall activity of the network tells how well a physical place like Sylvia Park is performing over time through each communication type. This tracks the number of likes, comments and posts that were made within each month, rather than the cumulative amount. This recognises a trend of growth across the year to determine the general direction as upward or downward. Furthermore, peaks and dips are used to find a pattern in the variance of growth to recognise cyclic patterns related to semesters, months, and seasons; and anomalies such as events, store openings, and pop-ups.

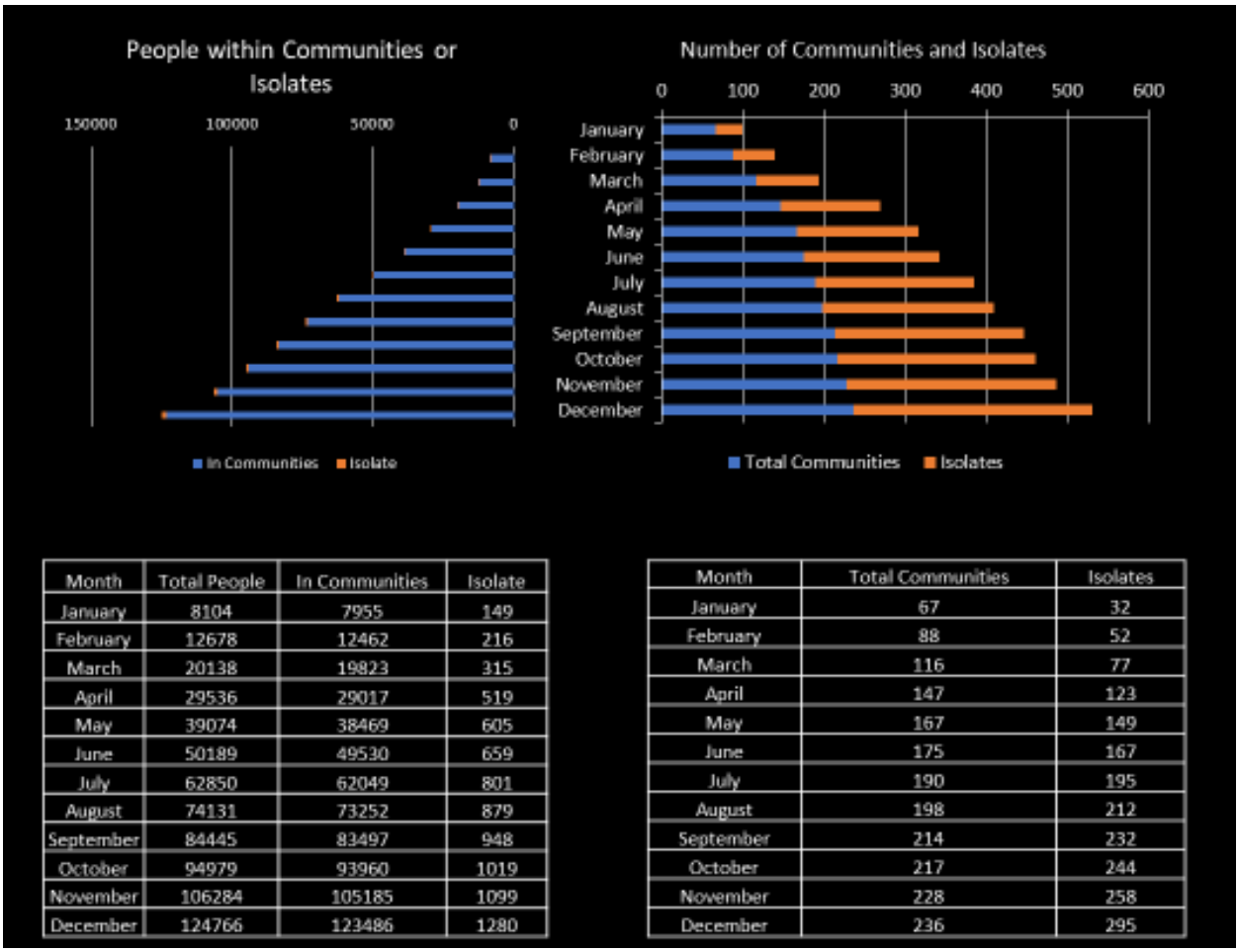
2.3.1.2 Trend and Variance of Growth

Over the year, all types of activity had continual and sustained growth; posts grew at 120 % and comments at 136 %, while likes grew at a rate of 152 %. By comparing the amount of change in each month we found that likes and posts had relatively stable growth. The likes had a variance of 10 % and posts had a variance of 12 %, while commenting behaviour was much more volatile with 64 % variance.

2.3.1.3 Patterns and Anomalies in Activity Growth

Variation are caused by major peaks or dips in the data that may fall into a pattern or have irregular fluctuations. As we are analysing 1 year of data, it is harder to study patterns as we are limited to short cycles. All types of interactions have peaks at the end of the year in December, and have a secondary peak in the middle of the year between June and July. For all types of interactions, the growth in the first half of the year is more rapid, but after the boom the growth is more stable until December. All types of interactions start with a dip, which is the end of the December boom from the previous year. It is much lower than the rest of the months in 2017 because of the continual growth which means that the activity from the previous year follows the same pattern but is scaled down.

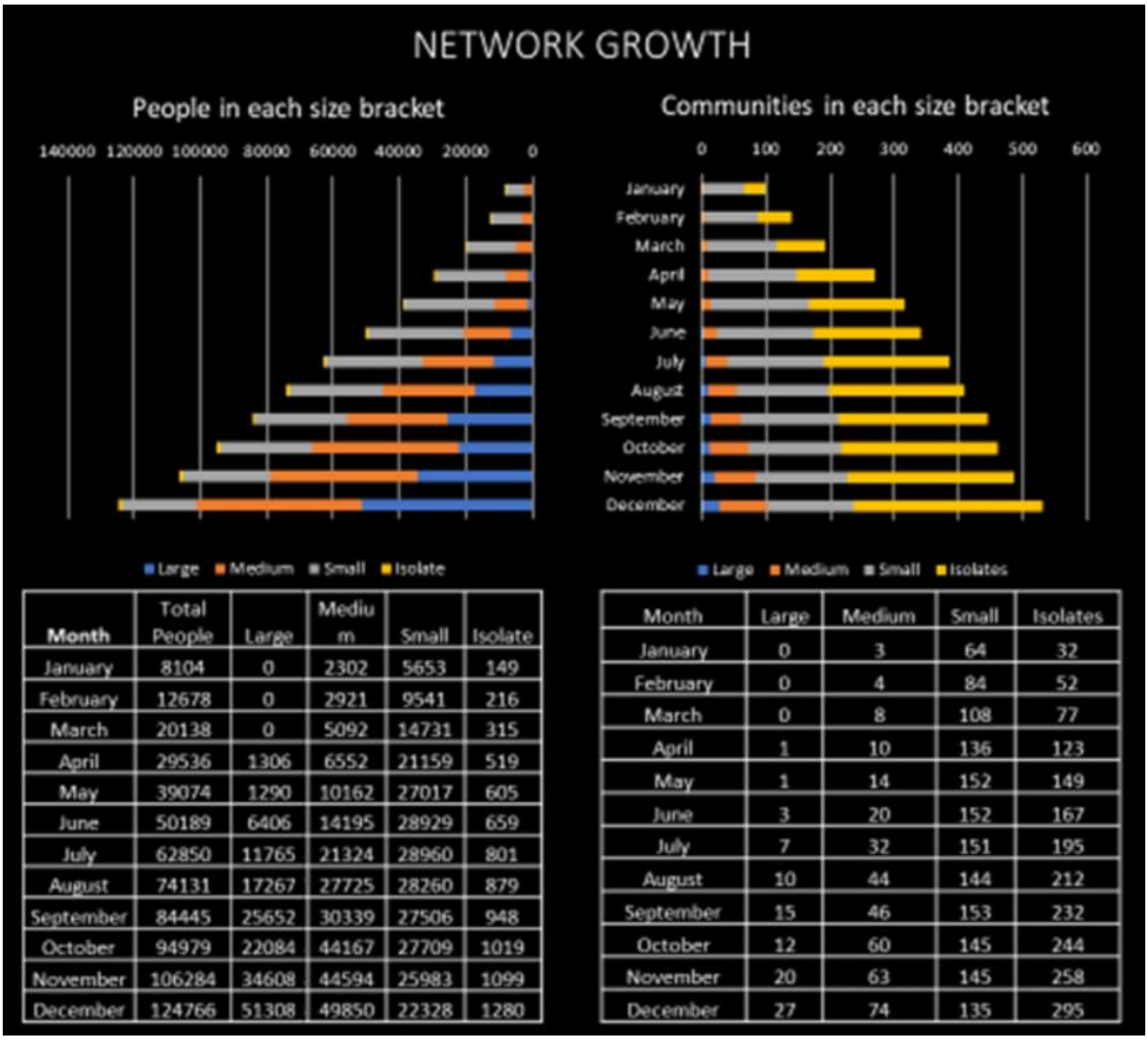
2.3.1.4 Growth in the Number of People and Communities



Month	Total People	In Communities	Isolate
January	8104	7955	149
February	12678	12462	216
March	20138	19823	315
April	29536	29017	519
May	39074	38469	605
June	50189	49530	659
July	62850	62049	801
August	74131	73252	879
September	84445	83497	948
October	94979	93960	1019
November	106284	105185	1099
December	124766	123486	1280

Month	Total Communities	Isolates
January	67	32
February	88	52
March	116	77
April	147	123
May	167	149
June	175	167
July	190	195
August	198	212
September	214	232
October	217	244
November	228	258
December	236	295

Figure 49. Network growth: People within communities or isolates and number of communities and isolates.



Month	Total People	Large	Medium	Small	Isolate
January	8104	0	2302	5653	149
February	12678	0	2921	9541	216
March	20138	0	5092	14731	315
April	29536	1306	6552	21159	519
May	39074	1290	10162	27017	605
June	50189	6406	14195	28929	659
July	62850	11765	21324	28960	801
August	74131	17267	27725	28260	879
September	84445	25652	30339	27506	948
October	94979	22084	44167	27709	1019
November	106284	34608	44594	25983	1099
December	124766	51308	49850	22328	1280

Month	Large	Medium	Small	Isolates
January	0	3	64	32
February	0	4	84	52
March	0	8	108	77
April	1	10	136	123
May	1	14	152	149
June	3	20	152	167
July	7	32	151	195
August	10	44	144	212
September	15	46	153	232
October	12	60	145	244
November	20	63	145	258
December	27	74	135	295

Figure 50. Network growth: People in each size bracket and communities in each size bracket.

Without measuring size, the number of communities is not a good indication of growth because it cannot differentiate between splitting communities or introduced communities. A large community that splits into many smaller communities would give the false appearance of growth. As a response we count both the number of people and communities within each size bracket and, also, track people shifting between communities and newly introduced members.

2.3.1.5 Amount of Growth

Both semesters are compared to track growth. Large communities have considerably higher growth than the other size brackets. They increased by 800 % in the second semester compared to medium communities which increased by 270 %; small communities dropped by -11 % and isolates grew by 77 %. Except for small communities, all had positive growth which gradually decreases as the size brackets shift from large to isolates.

2.3.1.6 Pattern of Growth

Small communities follow a different pattern of growth from the other community sizes. Its growth peaks near the middle of the year then decreases near the end of the year. Conversely, all other community sizes gradually increase so they peak near the end of the year. Furthermore, large communities have a higher variance in each month with 45 % compared to medium and isolates, which each have 4 %, and small, which has 2 %. This is because large has a major boom in growth during June and July. Furthermore, in October, the growth for large communities dips as medium and small communities increase by 5 %.

2.3.1.7 Formation of Large Communities

Large communities did not form until April. The next large communities would not appear for another 2 months, in June. The wide gap between the formation of large communities indicates the difficult conditions for them to arise. People enter the network as a large group only if tied to an influential individual or event which brings attention to the site. Conversely, multiple small groups may merge into a single large community if they build stronger ties gradually over time.

2.3.1.8 Shifting Between Communities and Merging Tendencies

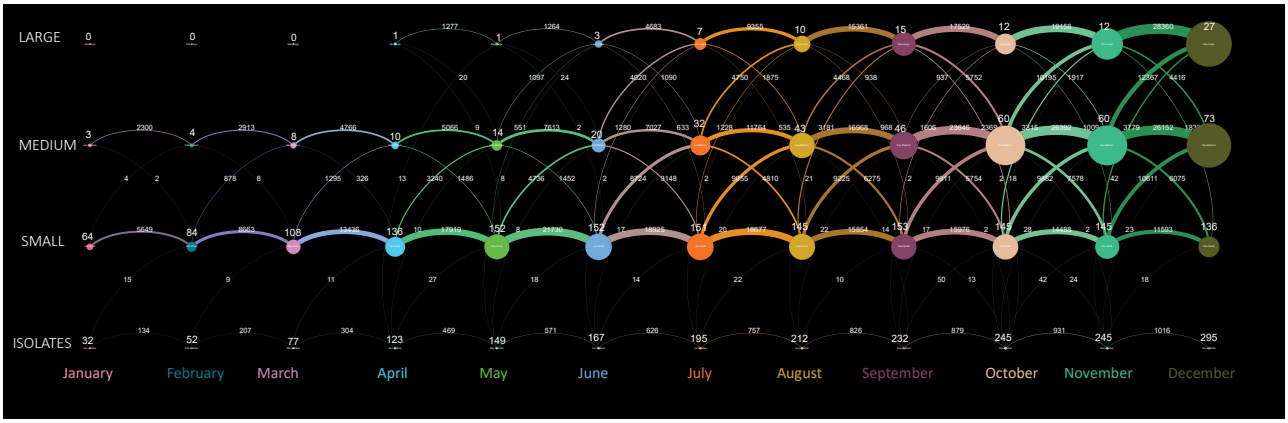


Figure 51. Shifting between different-sized communities.

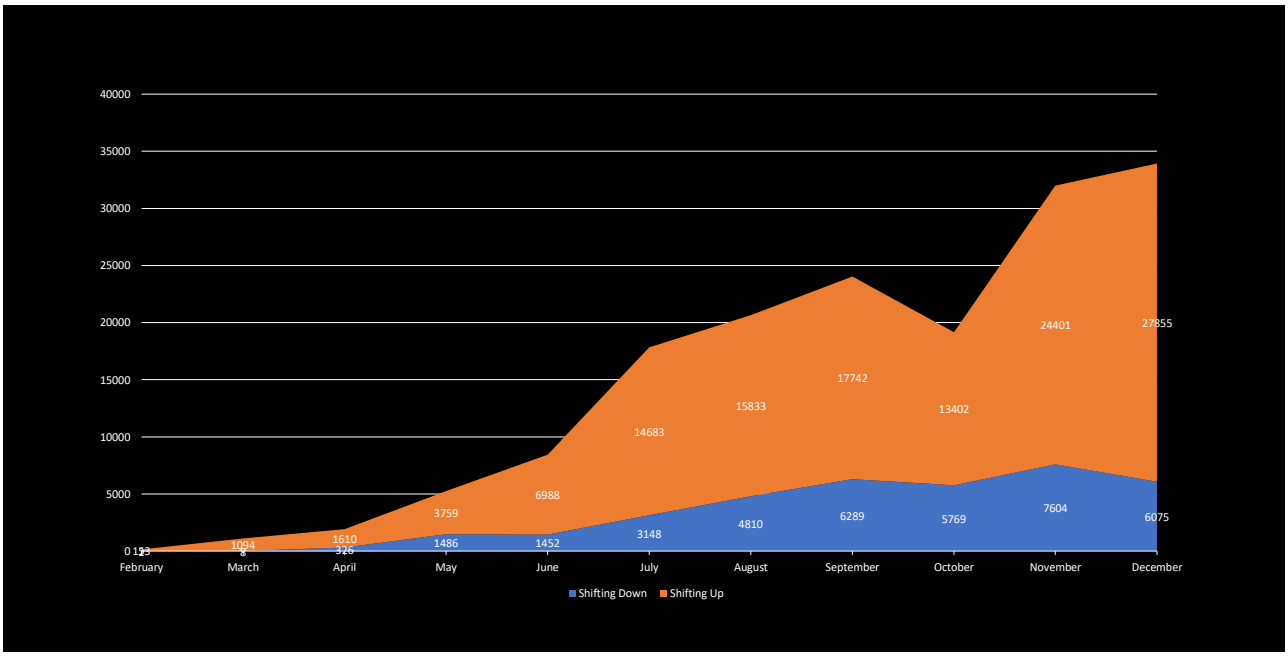


Figure 52. Number of people moving into larger communities vs smaller communities.

The movement diagram uses line thickness to represent the number of people moving between communities, while nodes are arranged from left to right by month and from top to bottom in descending size brackets. The size of the nodes relates to the number of people in the size bracket. Tracking people's movement between communities indicates whether a community has appeared spontaneously or was developed over time. People shift communities by forming stronger bonds with members in another community compared to their original community. For every month, we track the source and target community of each person.

Splitting and merging follows a pattern, with some minor exceptions, throughout the year, and depending on which size bracket is being observed. As people are introduced consistently throughout the year, the shifting pattern becomes more defined in each month. This can be seen in the diagram; lines get thicker and more differentiated each month. For each size bracket, the number of people moving has a set of tendencies which flows from strongest to weakest.

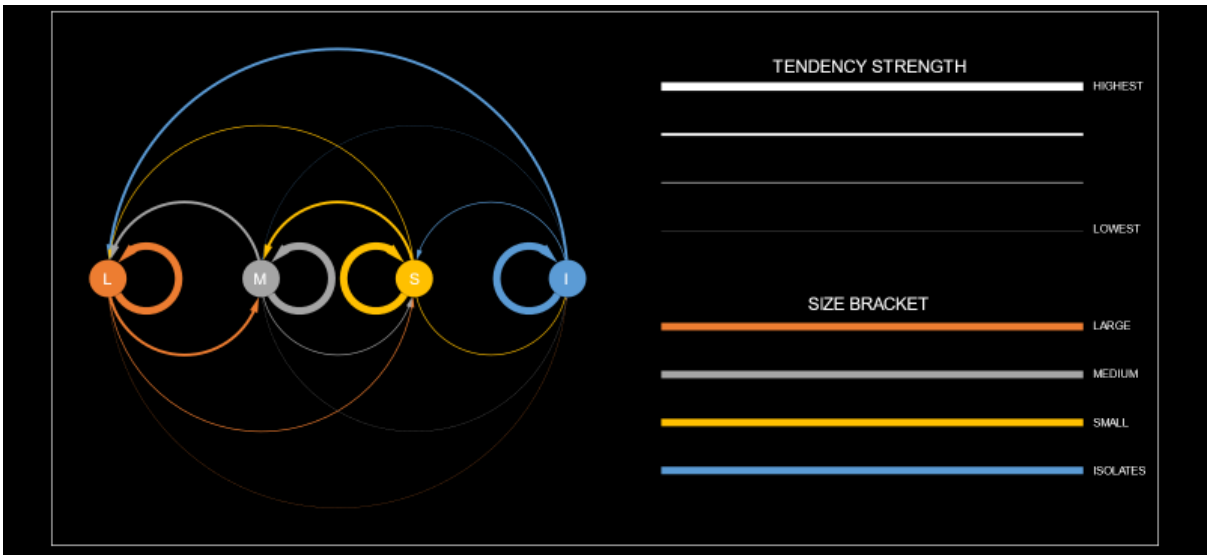


Figure 53. Shifting between different-sized communities.

2.3.1.9 Remaining in the Same Size Bracket

A fraction of people shifts between communities; however, most people tend to remain in the same size bracket. The effect is best seen at the end of the year from November to December, where 63 % (67,121) remain in the same size bracket while 37 % (39,163) do not. Out of the 39,163 who remain in the same size bracket, 42 % of them are in large communities, 39 % of them are in medium communities, 17 % are in small communities, and 2 % are in isolates.

2.3.1.10 Moving to the Nearest and Largest Size Bracket

The second strongest tendency that people follow is to merge with the nearest size brackets, in favour of the larger size bracket. Specifically, this is people moving from small to medium communities, and also from medium to large communities. However, as large is the maximum size bracket, members who shift from a large to an even larger community remain in the same size bracket. Furthermore, isolates are the exception to the nearest size-bracket rule. People from isolates tend to go to larger communities rather than small. This could be because small communities are not as capable of absorbing people as larger communities. This also suggests more merging instead of splitting, as they tend to shift up rather than down.

2.3.1.11 Moving to the Most-Distant Size Bracket

The third strongest tendency is for people to move into the most-distant size bracket which is not isolates. Specifically, this is people moving from small to large communities, or from large to small communities. Fewer people merge with the distant size brackets than the nearest size bracket. It is more difficult to merge communities with large differences in size, particularly from a larger size bracket to a smaller size bracket.

2.3.1.12 Moving to the Isolate Size Bracket

In all cases, the weakest merging tendency is with isolates. Only a few people from small communities merge into isolates, while all other communities do not merge with isolates at all. The number of people shifting into isolates is negligible, with the highest per month being 24 people.

2.3.1.13 Introduced into Network

The blue node at the top represents people who are external to the network and who are being introduced over time. The thicker line represents how many people are being introduced. Each comment, post, and like contains a timestamp and the target person of each interaction. This allows us to record when a person first enters the network, the people they are connected to when they enter, the size bracket they enter into, and how many are recurring Instagram users.

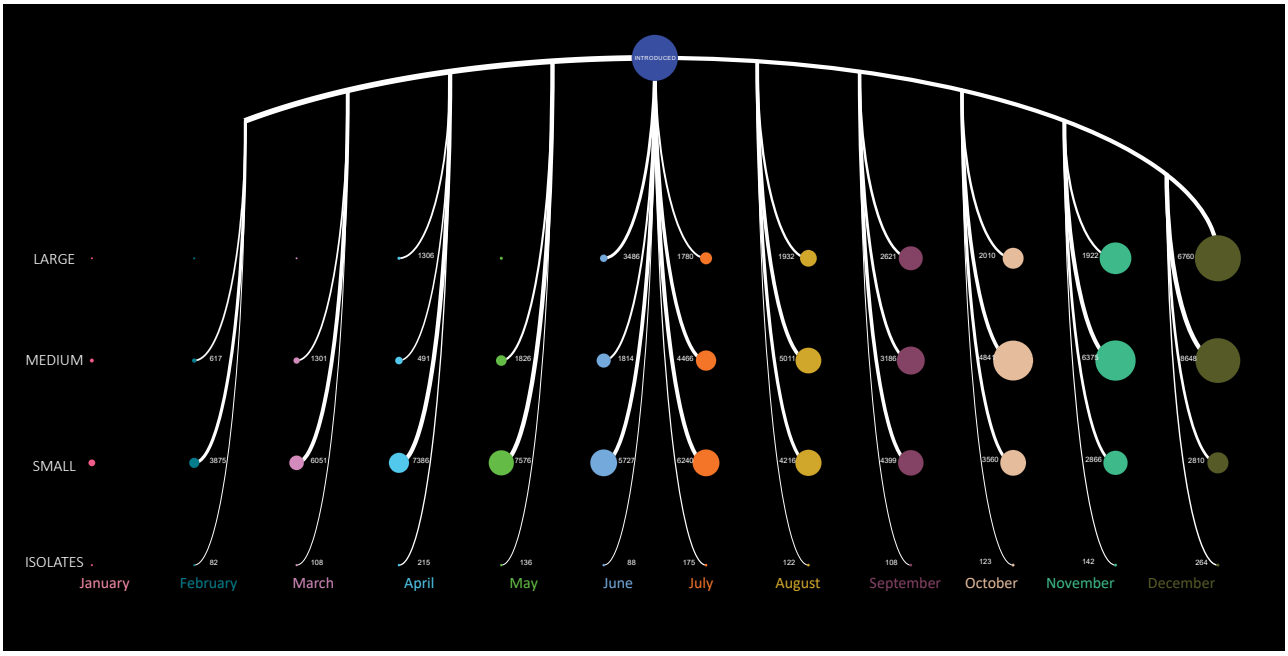


Figure 54. Distribution of growth: Shifting between communities + introduced people.

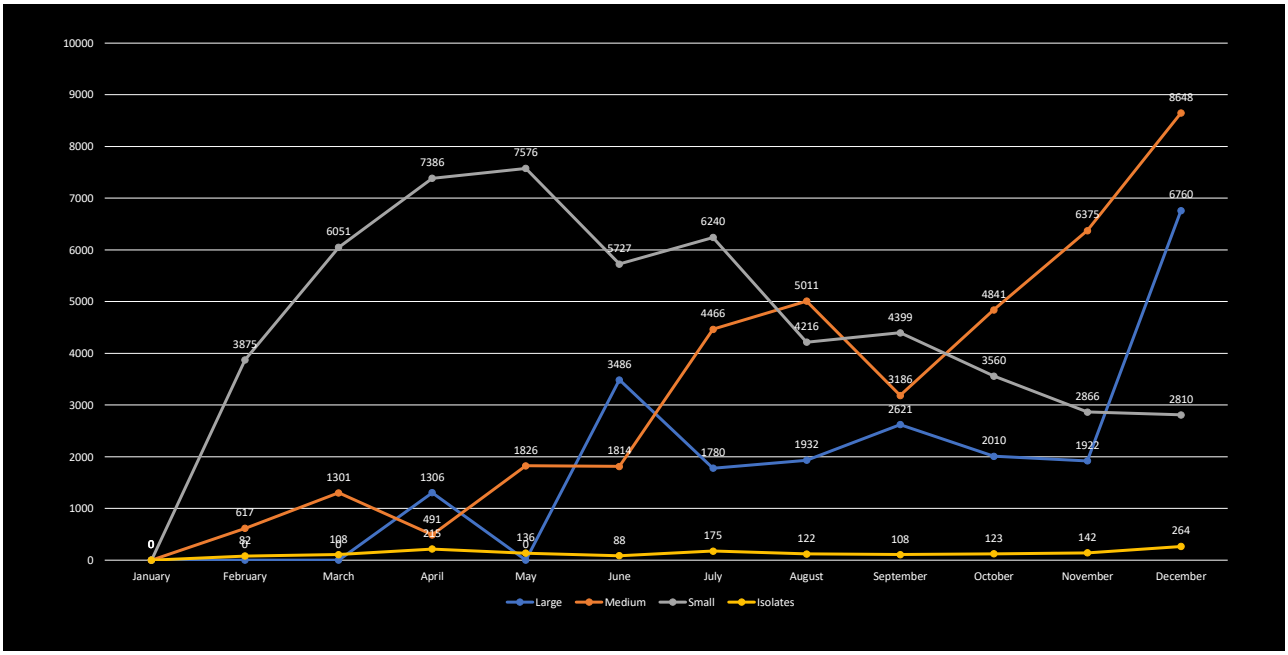


Figure 55. People introduced into each size bracket per month.

2.3.1.14 Entering Different-Sized Communities

It is important to track the size bracket people are entering as some parts of the network grow faster than others. In descending order, people enter into small, medium, and large communities and then isolates. However, in the second semester, more people enter into medium communities over smaller communities. This means that people have more access to the network as soon as they enter, rather than being absorbed into a medium community over time.

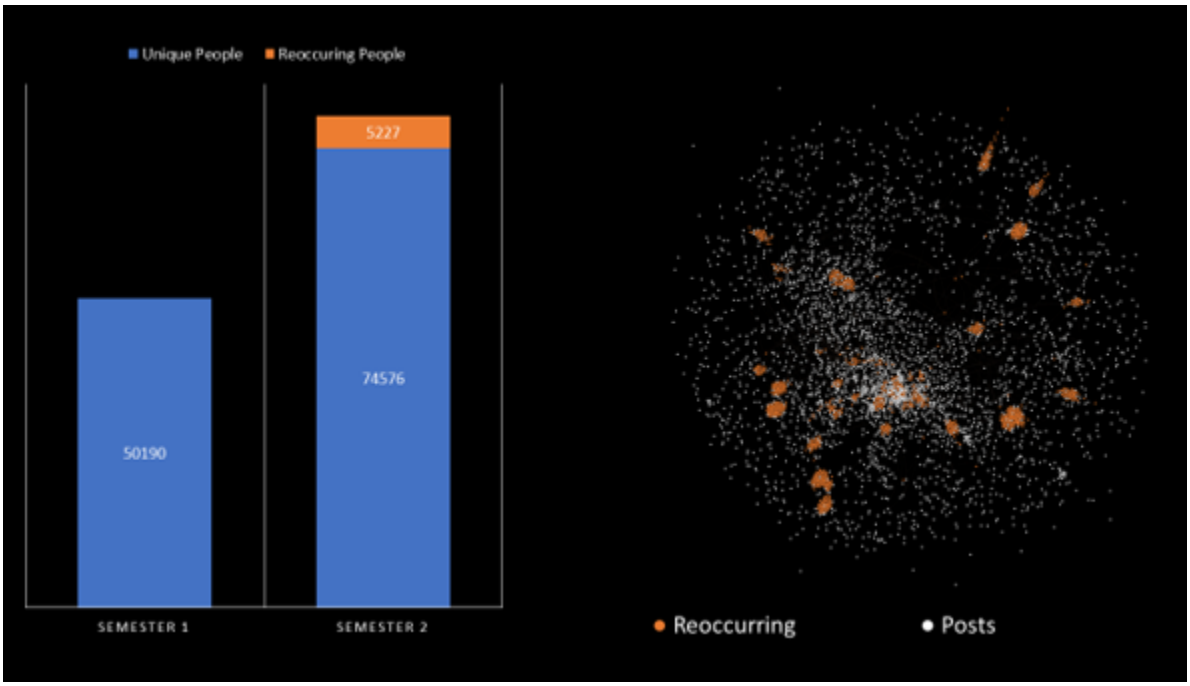


Figure 56. Introduced vs reoccurring.

2.3.1.15 Reocurring vs Unique

People who appear in both semesters are labelled as reoccurring. Those who are not are analysed to find which semester they first appear in. A network that has both a few reoccurring people and a few newly introduced people indicates a high-turnover rate. All people would be spontaneously and temporarily entering the network.

In contrast, a well-performing network would be high in both recurring and newly introduced people. This dataset shows 50,190 people in the first semester and TK people in the second semester. This is an increase of TK %. A total of 5,227 or TK % of people who were introduced in the first semester also continued posting in the second semester.

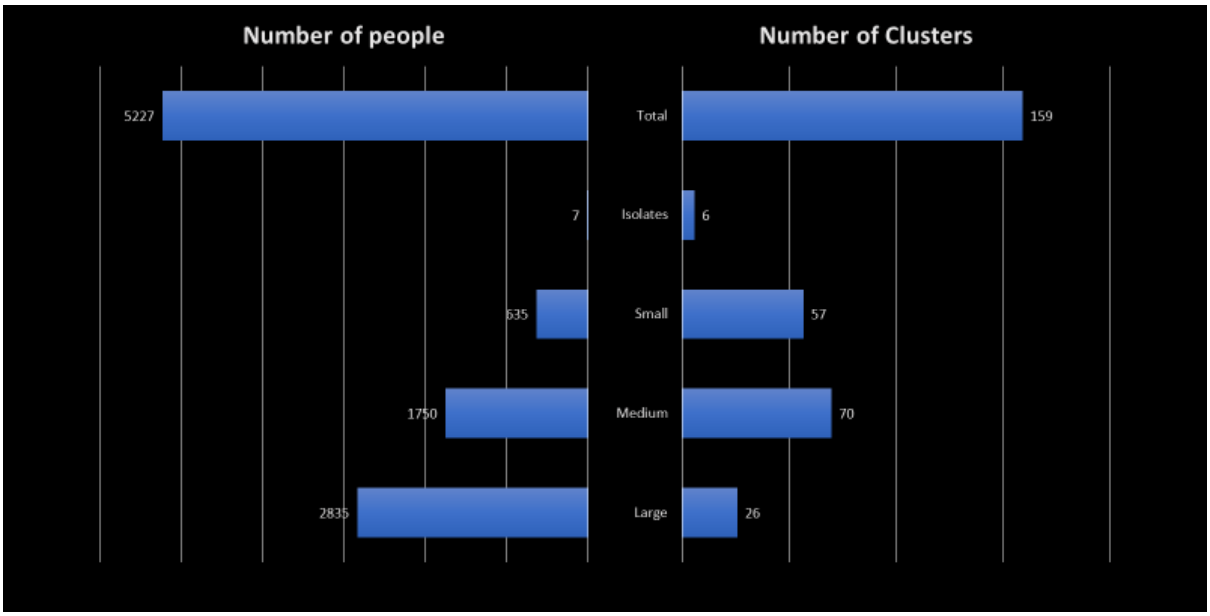


Figure 57. Rather than dispersed individuals, people enter the network in clusters.

2.3.1.16 Entering as Clusters or Individuals

People coming from different sources combine to form communities or join onto preexisting ones. Depending on the size of the community they are entering, some enter the network as interconnected groups while others are individuals. Those who enter as bigger clusters also enter larger communities. This is because smaller clusters attach onto bigger and more established groups, not the other way around. Therefore, the larger the core of each community, the greater the potential for absorbing more groups and bigger groups. This means that people attaching onto smaller or less defined cores, tend to form smaller communities. For instance, isolate groups are either made up of one small cluster that did not change, or a few individuals who become connected overtime. On the other hand, larger communities can absorb multiple clusters ranging in sizes from large, medium, small, isolates, and individuals.

However, it is important to focus on people who are active throughout the year to accurately measure different-sized clusters entering the network. This excludes spontaneous communities and prevents the data being distorted as we already know they enter as large groups. The 5,227 people who were recurring entered the network as TK clusters in total. Only seven people entered the network isolate groups. This means that the six isolate clusters comprised five disconnected individuals and one pair. In contrast, most people entered as a large cluster, despite there only being 26 clusters of this size. Most of the clusters were medium sized, followed by small clusters.

2.3.2 Dynamic Communities

A static community is a dense cluster of people in only one instance of time. On the other hand, a dynamic community strings together a series of these clusters where a person belongs to many different communities at different points in time. In our case, people belong to a maximum of 12 communities, one for each month.

2.3.2.1 Process of Measuring Growth and Community Interests

To measure growth, we must determine a stable group of people that best represents a community for a whole year, then identify people changing communities and entering the network in different parts of the

year. Thus, a dynamic community is made up of two elements: the core members (who act as a framework) and the transient members (people who attach to the core group). The core group has a fixed number of members which makes up a fraction of the community; the remainder is made up of transient members who are constantly changing over the year. In other words, as the core group is stable, it is the transient members that causes the increase or decrease in growth of a community. While the transient members can be measured easily by subtracting the core group, the core group must be identified by finding a subset of people with the most important lineage. This can be done with the steps described in the following sub-sections.

a. Assigning Nodes into Communities for Each Month

Community detection is done for each month to create a 12-part string which represents each person's lineage. Some people have shorter lineages because they have not been part of the network since the beginning of the year. Each part contains a letter, from *A* to *L,* to identify the month and, also, a number to identify the members of the community. For example, all people with *A12* in their lineage belong to the same community in January. A community splits occurs when a lineage of two or more people shift from a matching community in one month to different communities in the following month. By the same logic, merging between communities occurs when a lineage of two or more people shifts from different communities in one month to the same community in the next.

b. Identifying Core Groups

As the final community is connected to multiple lineages, a single lineage must be used as a point of reference to measure growth. To identify a stable group which best represents their community, different core detection methods were tested. The most reliable results were produced by identifying the largest lineage first, then, in the event of multiple lineages of equal size, the longest lineage is taken as the core group. Within each community, this finds the group of people who have the highest percentage of matching lineage across 12 months.

c. Identifying Transient Members

For each month we find members who have blocks of lineage which match the core group. These people are temporarily adopted into the same community as the core group but may move to another community in different months. The number of times each block matches with the core lineage is the number of months a person belongs to the same community. Wide spaces between the matching blocks mean a person has split from the community in one month, then is reintroduced much later in the year. Conversely, people with matching blocks, with no gaps between them, have more stable connections with one specific community.

d. Measuring Growth and Strength of Interest

Growth is measured by counting the transient members for each month plus the number of people in the core group. An increase in growth occurs when the number of people who have blocks that match the core group is higher than previous months. When no growth occurs at all, the absolute minimum size of a community equals the number of people in the core group. This is typically the case when a group forms spontaneously, as most of its members enter the network at the same time because they are closely connected to an individual or an event. Therefore, a community can only be measured once a core group has been detected. In some cases, this happens in January, meaning the group can be tracked for 12 months; in other cases, this happens in December, meaning the community can only be tracked for 1 month. Subsequently, interests are isolated by community ID for each month then each comment is given a value to indicate how strongly it expresses interest in a topic. The results from the community interests and growth are combined into a single graph.

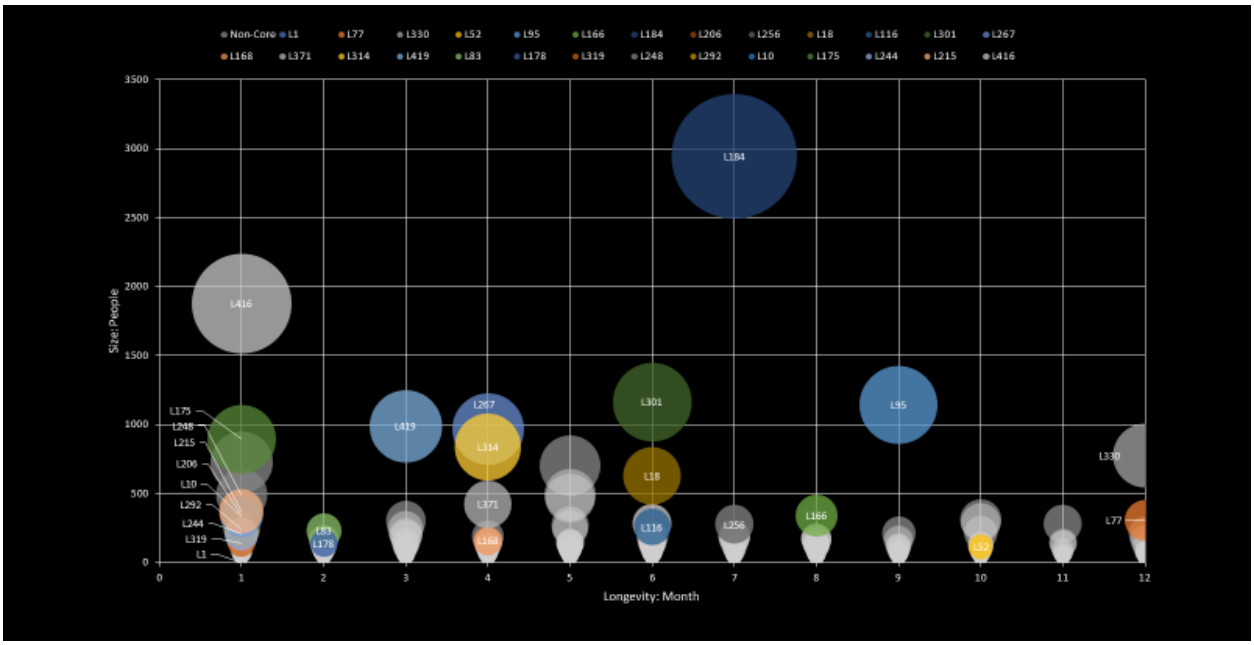
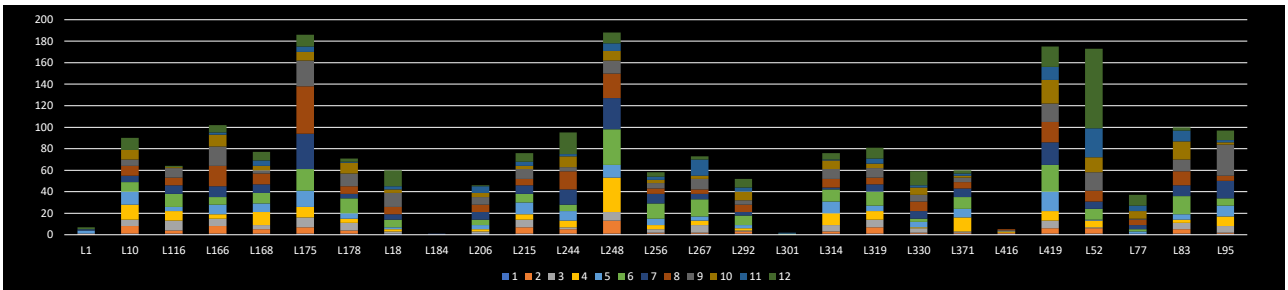


Figure 58. Core groups of largest communities.

2.3.2.2 Distinct Origins and Points of Introduction

Genealogy of a community can be traced back in time further than growth because it considers all lineage within a fixed community rather than only those who are temporarily attached to core lineages. The number of unique lineages and the date of their inception are used to evaluate the diversity and age of large communities. Diverse communities have a wide genealogy made up of many distinct origins, while a unified community has very few. A community with a wide genealogy has multiple descendants who were not associated at an early point in time, then gradually merged into a single community.

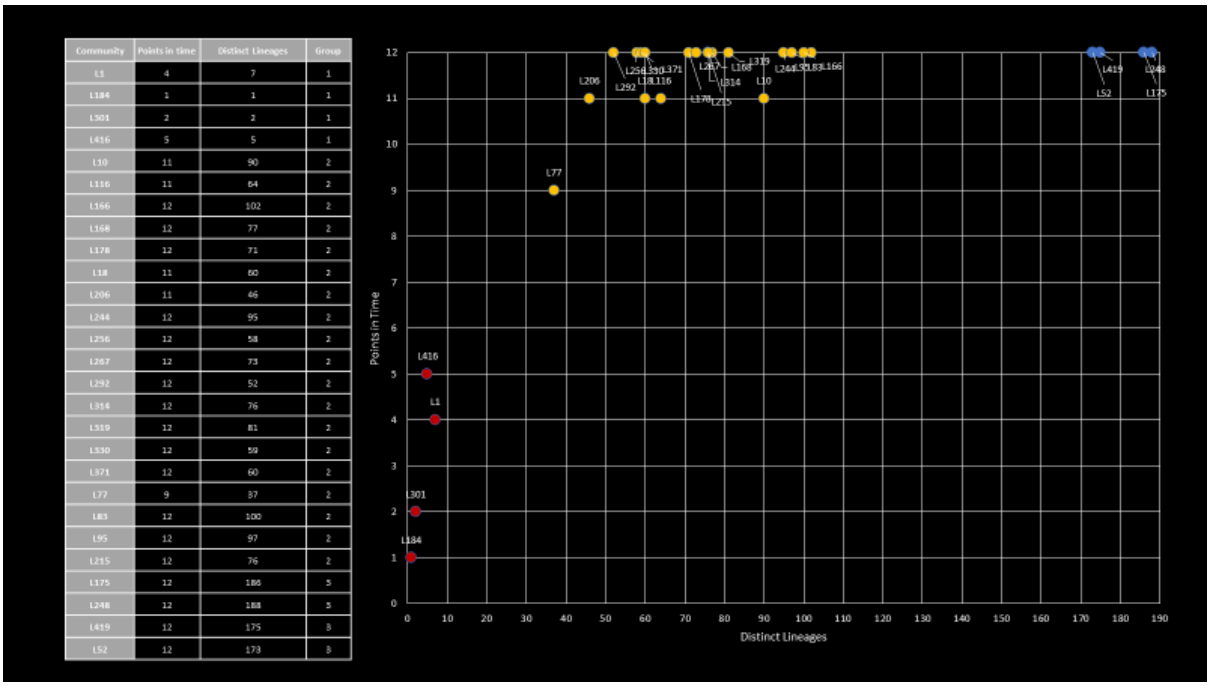


Community	1	2	3	4	5	6	7	8	9	10	11	12	Width	Max width	Max length	Average length
L1	0	1	0	0	3	0	1	0	0	0	0	2	7	3	0	5
L10	1	7	6	14	12	9	6	9	6	9	0	11	90	14	0	4
L116	1	3	9	9	4	12	8	7	9	1	0	1	64	12	0	6
L166	1	7	7	4	9	7	10	19	18	11	2	7	102	19	1	8
L168	1	4	4	12	8	10	8	10	3	4	5	8	77	12	1	4
L175	1	6	9	10	15	20	33	44	24	8	5	11	186	44	1	8
L178	1	3	7	4	5	14	4	7	12	10	1	3	71	14	1	6
L18	1	0	2	2	2	7	5	7	13	3	3	15	60	15	0	12
L184	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	7
L206	1	0	2	2	4	5	7	7	7	4	6	1	46	7	0	7
L215	1	6	7	5	11	8	8	6	9	3	4	8	76	11	1	5
L244	1	4	2	6	9	6	14	17	4	10	2	20	95	20	1	12
L248	1	12	8	32	12	33	29	23	12	9	7	10	188	33	1	6
L256	1	1	3	4	6	14	9	5	5	3	3	4	58	14	1	6
L267	1	1	7	4	4	16	5	4	10	3	15	3	73	16	1	6
L292	1	2	1	2	3	9	3	7	4	8	4	8	52	9	1	6
L301	0	0	0	0	0	1	0	0	0	0	1	0	2	1	0	6
L314	1	2	6	11	11	11	2	8	9	8	1	6	76	11	1	4
L319	1	6	7	8	5	13	7	6	9	4	5	10	81	13	1	6
L330	1	1	4	1	5	3	7	9	6	7	2	13	59	13	1	12
L371	1	1	1	13	8	11	8	6	4	2	2	3	60	13	1	4
L416	1	1	0	1	1	0	0	1	0	0	0	0	5	1	0	1
L419	1	5	7	9	18	25	21	19	17	22	12	19	175	25	1	6
L52	1	5	1	6	1	10	7	10	17	14	27	74	173	74	1	12
L77	1	0	0	0	2	2	4	5	1	7	5	10	37	10	0	12
L83	1	4	6	3	5	17	10	13	11	17	10	3	100	17	1	6
L95	1	1	6	9	10	7	16	5	29	2	2	9	97	29	1	9
Total	24	83	112	171	173	270	233	254	239	169	124	259	2111	74	0	6

Figure 59. Number and age of distinct lineages within each community over 12 months.

Each community comprises many people who enter the community in different points in time. The diversity and distribution of origins within a community is recognised by the number of unique lineages in parallel with the points in time at which they were introduced. This observes whether a community has many or few lineages and if their introduction was concentrated in a few points in time or over a long time span.

Out of 2,111 distinct lineages found for large communities, 61 % or 1,278 of the communities belong to the second semester, while only 39 % or 833 belong to the first. The community with the widest genealogy is L248, with a total of 188 lineages, followed by L175 (186 lineages) and L419 (175 lineages). Most of the communities have lineages that originated in different parts of the year. Of all large communities, 23 or 85 % had lineages from over 8 different points in time.



Community	Points in time	Distinct Lineages	Group
L1	4	7	1
L184	1	1	1
L301	2	2	1
L416	5	5	1
L10	11	90	2
L116	11	64	2
L166	12	102	2
L168	12	77	2
L178	12	71	2
L18	11	60	2
L206	11	46	2
L244	12	95	2
L256	12	58	2
L267	12	73	2
L292	12	52	2
L314	12	76	2
L319	12	81	2
L330	12	59	2
L371	12	60	2
L77	9	37	2
L83	12	100	2
L95	12	97	2
L215	12	76	2
L175	12	186	3
L248	12	188	3
L419	12	175	3
L52	12	173	3

Figure 60. Distribution of distinct lineages over 12 months.

Analysing the relationship between points in time and number of lineages, we find that communities with many unique lineages tend to enter the network over a long time span, while communities with fewer lineages grow sporadically. A scatter plot, also generated from the data, shows three. One cluster containing four communities has a few lineages that entered their communities within a few points in time. They have between 7 and 5 lineages that is broken up into 4–5 moments. In contrast, a second cluster of 19 communities has many more lineages over a much wider timespan. These communities contain between 37 and 102 lineages of 9 or more months. A third cluster made up of four communities has considerably higher lineages which originated across the entire year; between 173–188 entered the community every month.

2.3.2.3 Growth Patterns

Key features of spontaneous or sustained communities are identified by differences in their growth. This focuses the gradient of trendlines to show the rate of growth and direction. In addition, active parts of the year are identified by the date when communities were established and peaks in growth. Furthermore, stability of communities is rated by comparing the number of transient and core members.

2.3.2.4 Start of Growth and Peaks

Each community is established at different points ranging from January to December. The number of established communities within each month indicates which parts of the year stimulate the most growth. Many more communities formed in the second semester but they are more concentrated to specific months. This leads to wider gaps between when communities are established. In the first semester, only six communities are formed. In the second 21. Despite this, February is the only month in the first semester with no communities, while August and October have no communities for the second semester. Therefore, most of the communities are formed in December with nine in total. June, July, September and November also have considerable activity receiving three communities each, while March, April and May only receive one community.

Additionally, by counting the number of peaks and their position in the year we find that most of the communities form late in the year and only have one peak. The 16 communities with one peak all form during December and November. The nine communities with peaks form around the middle of the year between May and September. Only L330 and L95 have more than two peaks and they form in January and March. Therefore, a well-performing site would have many communities that form early in the year, because they tend to experience multiple bursts of growth and activity.

2.3.2.5 Trendline

The gradient of the trendline describes the growth of a community using a single positive or negative value to show direction and steepness by the size of the value. Sixteen communities have growth (55 %), four communities have a loss (14 %); nine communities could not be calculated as they start in December (31 %). Communities that form near the start of the year have a gentle gradient but are all positive. Communities that form near the middle of the year have more variation between the direction of growth as some are positive and some are negative. Near the end of the year, there is limited data, so the gradient cannot always be recorded. Gradients are exaggerated near the end of the year as there are only 2 months of data for communities that start in November. This because many communities have a peak in growth soon after they are formed; indeed, this is the case for 12 out of the 27 communities.

2.3.2.6 Difference of Core and Transient Members

Stability is found by the difference between average transient members and core people within a community. This value is given as a positive or negative percentage change over a 12-month period rather than for the communities' life-span as it accounts for communities that start both near the end or the beginning of the year. Otherwise, communities that start near the end of the year would be rated better than ones that have formed in the beginning of the year, even though they have more transient members. Therefore, larger values indicate well-established communities that have continual growth and activities. This dataset ranges from the highest value of 78 % increase for L52 to the lowest value of -589,100 % for community L184. While this method points out that communities which form earlier in the year are better established, the method has limited accuracy when communities have slow growth over a long period, or large growth over a very short period. Community L1 should be well established as it has been continually active since January; yet it has a low value of -307 % because only a few new members join the community during this time. The second limitation is pointed out by L83, L11, and L12 as they have a high value but formed near the end of the year. As there is currently only 2 months of data available for these specific communities, the results would improve as more data is collected on them for following years.

2.3.2.7 Spontaneous vs Sustained

By looking for a set of key features we can identify which communities are short lived or sustained. A spontaneous community experiences a short burst of activity and growth which forms around a single moment in time. Thus, they have a shallow gradient of growth and their activity can be contained within 1 month. They also have a large core group, as members of spontaneous communities tend to be introduced at the same time and do not exist in the network prior to an event. Furthermore, they must be formed before September to accurately determine if activity has ceased after the event. This provides enough time after the community has been formed to observe for delayed or second peaks in activity. Because of this, eight of the communities could not be measured because they start too late in the year. However, both L184 and L301 have been identified as spontaneous communities as they hold these key features. L184 has a gentle gradient of -0.11, much like L301 which has a gradient of -0.60. Both have a wide gap between core and transient members, with -589,100 % for L184, and 725 % for L301. Furthermore, they also established in the middle of the year: in June for L184, and July for L301.

2.3.2.8 *Interest of Dynamic Communities*

Investigating dynamic communities' interests allows a trace of interest patterns during and after the official establishment of each community. The communities are separated and assigned to different periods of establishment including communities established in January, communities established in the first half of the year (February–May), communities established in the second half of the year (June–November), and communities established in December. This division enables a clearer comparison of interest patterns of communities established within the same period, as well as a comparison of communities established in different time periods. This leads to recognition of differences in communities' characteristics and interest patterns to find out whether the communities are established by a certain interest.

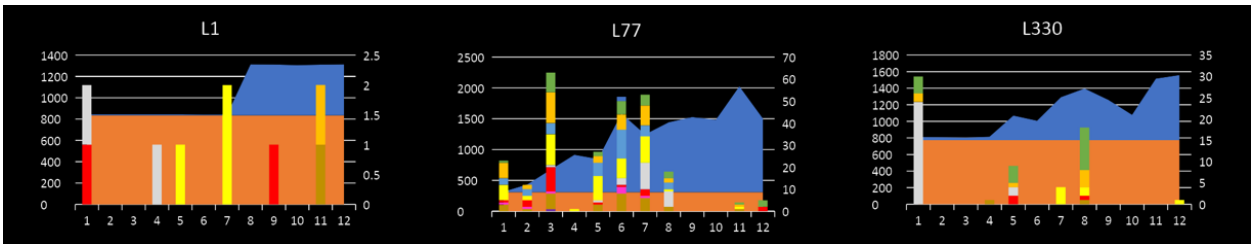


Figure 61. Timeline of dynamic interest of community L1, L77, and L330 established from January.

2.3.2.9 *Communities Established in January (L1, L77, L330)*

Communities L1, L77, and L330 are officially formed since January 2017, hence, they are the longest lasting communities of the 27 largest communities. Community L77 is the most active among the three communities. It has a higher level of interests with more consistent interests presented in almost every month. The interest pattern shows peaks in March, June, and July and an absence of interests in September and October. Although the interest level in community L77 rises in the first half of the year and decreases in the second half of the year, the number of community members is constantly increasing toward the end of the year. Thus, the community does not have a significant influence from an interest topic; it is growing through both forms of interaction, starting with comments on a variety of topics in the beginning of the year and continues growing through likes in the second half of the year.

In comparison to community L77, community L1 and L330 have lower interest levels and bigger gaps between months with interests presented, although they have more specific interests, suggesting the topics of interest to the community's members and the possible causes of the community's transformation. For example, the rise of Food and Drinks in June perhaps leads to an increase of the members in community L1 in July. Yet, the dominance of Fashion and Styles in January in community L330 could be the influence of the community establishment; however, it is probable that community L330 formed before January. Therefore, the interest topics detected in January are not necessarily what influenced the community's establishment.

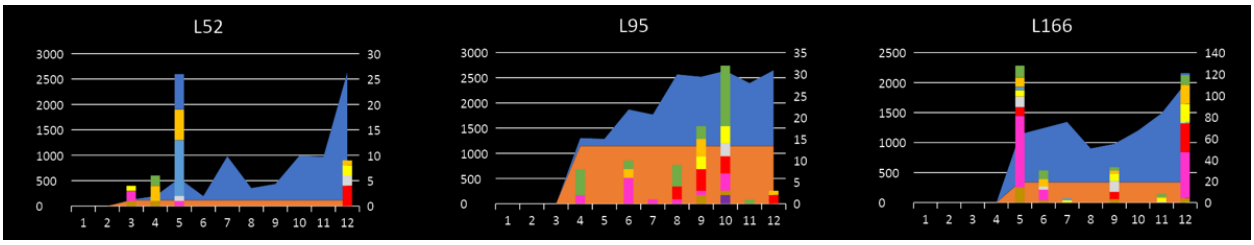
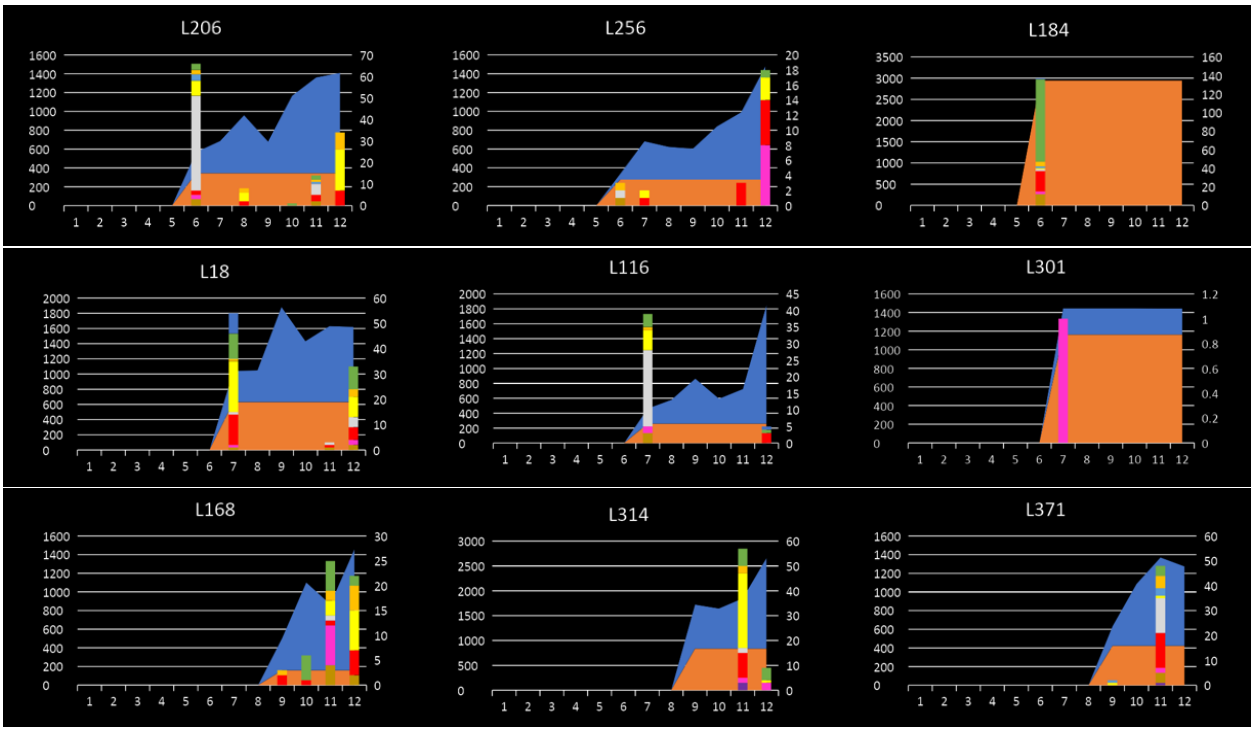


Figure 62. Timeline of dynamic interest of community L52, L95, and L166 established within the first half of the year.

2.3.2.10 Communities Established After January Within the First Half of the Year (L52, L95, L166)

Communities L52, L95, and L166 are officially established in March, April, and May. Hence, significant interests detected during the establishment of these communities are more likely to be an influence on the communities' official formation. Community L166, especially, has the highest level of interests during its establishment in May and Beauty, Sports, and Wellness is the most predominant interest topic of the month. The interest level in the following months is lower until another peak of interests in December, although Beauty, Sports, and Wellness remains the most significant topic. Thus, the significance of Beauty, Sports, and Wellness continues as community L166 is grows.

Communities L52 and L95, however, have much lower interests and an absence of prominent topics during their establishment. Instead, community L52 has a peak of interests with a dominant interest in Nature in June; and community L95 has an increasing trend of interests from August, September, and a peak in October with a dominant topic in Social and People. Hence, these two communities can be driven by or affected by certain interests for a period of time, but the interests do not have a significant impact during the communities' establishment.



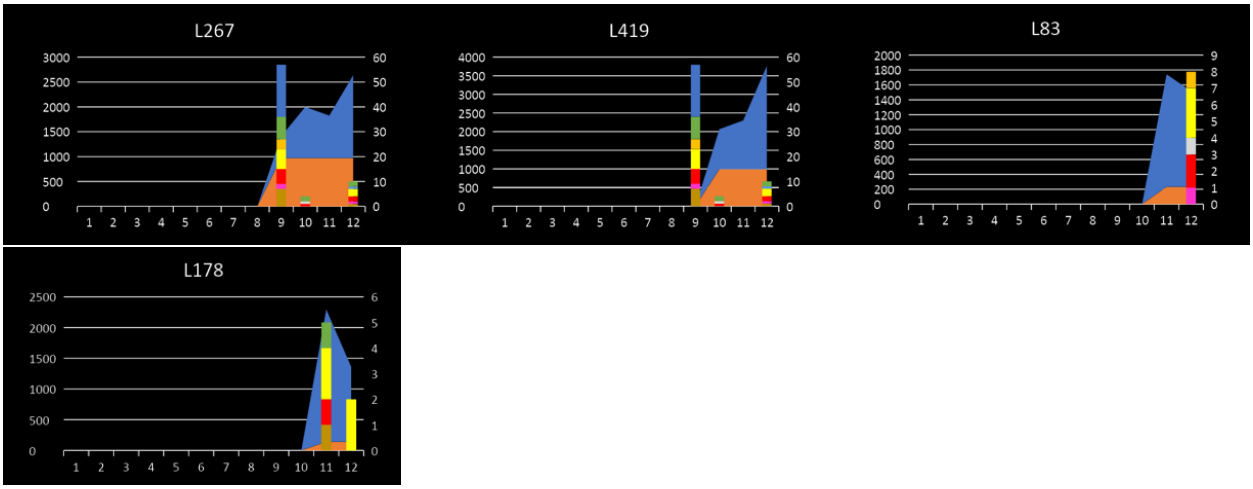


Figure 63. Timeline of dynamic interest of large-sized communities established in the second half of the year.

2.3.2.11 Communities Established in the Second Half of the Year (L206, L256, L184, L18, L116, L301, L168, L314, L371, L267, L371, L267, L419, L83, and L178)

There are 15 communities established in the second half of the year: L206, L256, L184, L18, L116, L301, L168, L314, L371, L267, L371, L267, L419, L83, and L178. Five communities have significant interests recognised during their establishment; these are communities L206, L184, L18, L116, and L267.

Notably, community L206 and L116 share the same significant interest of Fashion and Styles in the month of their establishment in June and July, which are also the highest peaks of their dynamic communities' interests. Yet, the patterns of interests for the two communities after their establishment are also similar. The interests of community L206 decrease and discontinue in some months before the presence of a low interest level in November includes Fashion and Styles as the dominant interest of the month, followed by another moderate rise of interests in Events and Entertainment, Food and Drinks, Places and Architecture, in December. Community L116 has a longer absence of interests after the community establishment in July, with a last small rise of interests in Events and Entertainment, Social and People, and Technology appearing in December.

Communities L18 and L267 have significant amounts of interests during their establishment but the interest level is more distributed in different topics rather than predominantly concentrated on one topic. Community L18 has most interest in Food and Drinks, closely followed by Events and Entertainment. Community L267 has a dominant interest in Technology then an approximately equal distribution of interests in Social and People, Food and Drinks, Events and Entertainment, and Art, Design, and Photography. Additionally, the interest patterns of the two communities are similar to each other as there are absences of interests in some months and low interest levels in other months after their establishment, although community L18 has another notable rise in December when Food and Drinks and Social and People become significant interests of the community again.

Although community L184 is another community with significant interest, highly concentrated on Social and People, during the community's establishment, the interests are absent afterwards as the community stops growing; hence, community L184 is only active in June.

Otherwise, other communities formed in the second half of the year have low interests at their establishment; the interests then gradually increase toward the end of the year, for community L168, or suddenly increase in one of the months as seen in community L371.

Communities L314 and L83 do not have any interests expressed during their establishment but they have a peak of interests in November and December. This suggests that the communities are established through likes rather than comments as no keywords are detected to inform the interests during their establishment period.

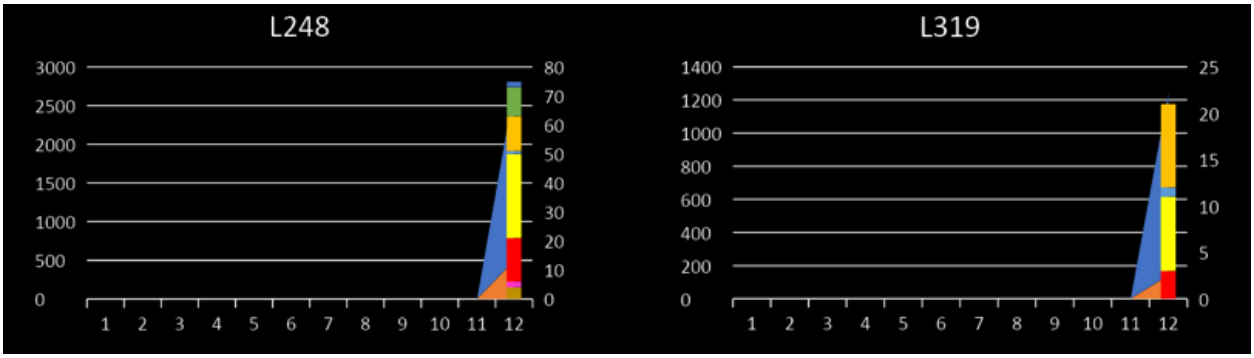


Figure 64. Timeline of dynamic interest of community L248 and L319 established in December.

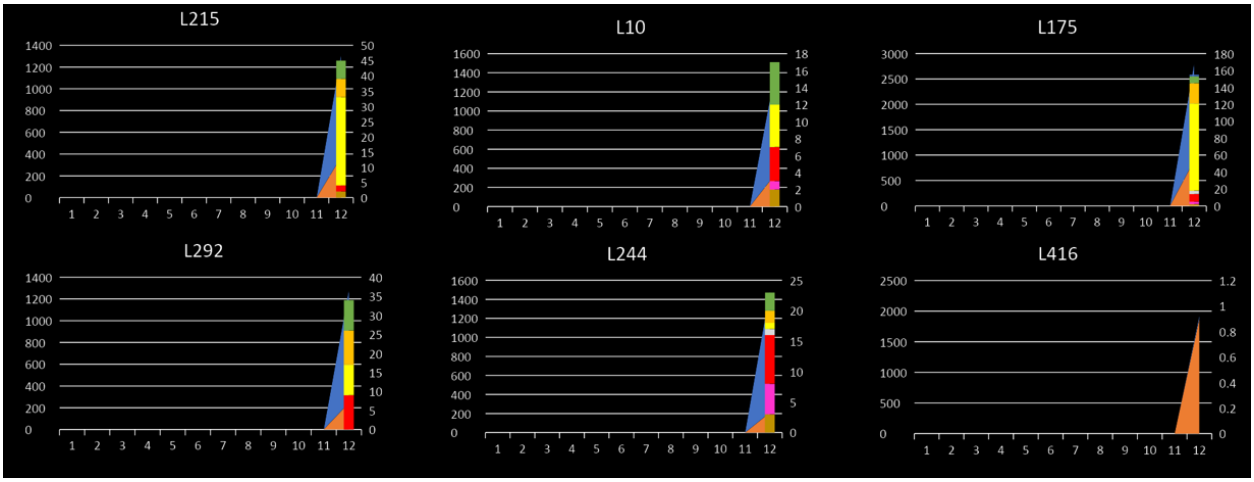


Figure 65. Timeline of dynamic interest of other large-sized communities established in December.

2.3.2.12 Communities Established in December (L248, L319, L215, L10, L175, L292, L244, L416)

Significantly, eight communities, L248, L319, L215, L10, L175, L292, L244, and L416, are established in December. This results in a lack of community interests growth patterns in the months following the communities' establishment.

However, most communities established in December reveal moderate to high interest levels, except community L416, which has no keywords detected, hence, no interest expressed. Communities with moderate interest levels, including communities L319, L10, L292, and L244, are likely to have more distributed interests over multiple topics—for example, community L10 has an almost equal distribution of interests over Social and People, Food and Drinks, and Events and Entertainment. Communities with high interest levels, including communities L248, L215, and L175, tend to have a single prominent

interest—for example, community L215 has the highest strength of interest *in Food and Drinks*, which is higher than a combination of all other interest topics.

Notably, the three communities with high interest levels have the same predominant interest in Food and Drinks, thus, the establishment of these communities is influenced by the same interest topic. Additionally, December is a festive and holiday season, therefore, most communities established during this time are influenced by a form of celebration. While communities with high interest levels are initiated by an interest in Food and Drinks, communities with moderate interest levels are influenced by diverse forms of celebration in various topics such as Food and Drinks, Events and Entertainment, Social and People etc. Accordingly, communities established in December are most likely initiated by celebration and holiday events.

Communities established in the same time period have a propensity to carry a similar pattern of interests due to a similar time range or events happening within the time span. Communities established in January have the longest history, showing the most complete interest patterns throughout the year. Although the communities show various interest levels across different communities, they reveal absences of interests in some months and a decreasing trend toward the end of the year. It is probable that the communities found established in January are formed prior, thus, the communities' interest patterns are decreasing as the communities are growing. Additionally, communities established in December cannot show growth and interest patterns after their establishment (most of them are established with a moderate to high interest levels). Perhaps communities established with a high interest level in December then continue growing in January of the next year as the interest level is decreasing.

The three most common interest topics, which are found to influence communities' establishment, are Food and Drinks; Beauty, Sports, and Wellness; and Fashion and Styles. Dominant interest in Beauty, Sports, and Wellness; and Fashion and Styles, is mostly detected during communities' establishment in January, the first half of the year, and the second half of the year. Dominant interest in Food and Drinks is commonly shown in the second half of the year and, significantly, in the communities established in December. Thus, Food and Drinks is the most influential interest topic which initiates communities' establishment, especially during holidays and festive seasons toward the end of the year.

Māori Vocabulary (as found)	Word Frequency
AE	3
AOTEAROA	26
EHOA	1
HORI	1
KAHA	1
KAIMOANA	1
KIA	6
KIAORA	1
MANUKAU	4
MAORI	6
MATAMATA	1
MOANA	2
MOKO	1
OI	6
ONEHUNGA	1
ORA	2
OREWA	2
PAKI	1
PAPAMOA	1
POHUTUKAWA	1
POI	1
TĀMAKIMAKAURAU	1
TEKOHA	3
TIKI	2
TIMU	1
TUI	1
WAIHI	1

WAIRANGIIIII	1
WAKA	2
WHANAU	3
WHĀNAU	1