



UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Ciencias de la Computación

Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Redes sociales basadas en geolocalización en la Ciudad Autónoma de Buenos Aires (2009-2015)

*Tesis presentada para optar al título de Magíster en Explotación de Datos y
Descubrimiento del Conocimiento*

Leonardo Salvador Rocco

Director de tesis: Dr. Marcelo A. Soria

Buenos Aires, 2021

Resumen

La localización, en tanto ubicación en un espacio geográfico dado, es uno de los componentes principales del contexto de las personas a partir del cual se pueden estudiar cuestiones tales como intereses individuales y sociales, comportamiento en términos espaciales y temporales así como relaciones con otros individuos. El tipo de redes sociales que se basan en la localización de sus usuarios recibe el nombre de redes sociales basadas en ubicación (LBSN). Utilizando datos registrados en Foursquare (<https://foursquare.com>), aplicación para dispositivos móviles que provee a sus usuarios búsquedas personalizadas y localizadas, se analizaron las actividades espacio-temporales y redes sociales que registraron los usuarios dentro de la Ciudad Autónoma de Buenos Aires entre 2009 y 2015.

Entre los hallazgos más importantes, se encuentra que las personas que tienen un lazo de amistad en la plataforma sociabilizan más cerca, en términos geográficos, que aquellas personas que no son amigas en la misma. Se encuentra evidencia de que los usuarios no hispano parlantes frecuentan sectores más concentrados geográficamente de la ciudad y tienen diferentes características que los usuarios hispano parlantes.

Palabras Claves: LBSN, Foursquare, redes sociales, geolocalización, Buenos Aires, análisis de redes sociales

Abstract

Location is one of the main components of people's context from which subjects such as individual and social interests, behavior in spatial and temporal terms as well as relationships with others can be studied. The type of social networks that are based on the location of their users is called location-based social networks (LBSN). Using data registered in Foursquare (<http://foursquare.com>), an application for mobile devices that provides its users with personalized and localized searches, the spatio-temporal activities and social networks registered by users between 2009 and 2015 within the City of Buenos Aires were analyzed.

Among the most relevant findings, we could see that people who have a friendship bond on the platform socialize geographically closer than those who do not have a friendship bond. Evidence shows that non-hispanic users tend to frequent more concentrated geographic areas in the city and have different characteristics than hispanic users.

Keywords: LBSN, Foursquare, social networks, geolocation, Buenos Aires, social network analysis

Agradecimientos

A mi director de tesis Dr. Marcelo Soria por su paciencia, aliento y guía en el desarrollo de esta tesis.

A la universidad pública, libre y gratuita, en donde he decidido caer.

A mi familia, por siempre incentivarne, y en especial a mi vieja que ya no está.

A Rolo, por su afecto y la paciencia con la tesis.

A mis amigas/os formados en la cursada de la maestría, y en especial a Jorgito que ya no está.

Finalmente a las organizaciones de tecnología, RLab y P.V.M., que me ayudan a pensar la informática como un espacio de libertad, colaborativo y político.

Aclaración sobre el uso del género gramatical:

Salvo que el género sea explícitamente definido (“personas de género femenino”, “usuarios identificados como género masculino”, etc), se utilizará indistintamente los pronombres y el género gramatical femenino y masculino, intercalando los mismos a lo largo de este trabajo, para referirse a la totalidad de personas.

Índice

Introducción

1 Introducción	8
1.1 Sociabilidad en línea y redes sociales	8
1.2 Extensión de las redes sociales localizadas	9
1.3 Redes sociales localizadas: el caso de Foursquare	10
1.4 Objetivos	12

Materiales y métodos

2 Elementos e instrumentos empleados	13
2.1 Descripción de métodos para la recolección de datos	13
2.2 Información geográfica	17
2.3 Diseño de base de datos	19
2.4 Procesos y componentes de software	20
2.5 Procedimiento de recolección de datos	22
3 Procesamiento de datos	23
3.1 Detección y aislación de valores extremos	23
3.2 Cálculo de variables adicionales	25
3.3 Clasificación de Idioma	25
4 Teoría y modelado de red social	28
4.1 Teoría de Grafos	28
4.2 La red social de comentadores de CABA en Foursquare	29
4.3 Análisis de Redes Sociales	31
4.4 Métricas y propiedades de las Redes Sociales	31
4.5 Medidas de Centralidad	32
4.4.1 Centralidad de grado (Degree)	32
4.4.2 Grado promedio (Average Degree):	32
4.4.3 Camino más corto (Shortest Path):	32
4.4.4 Excentricidad (eccentricity):	33
4.4.5 Centralidad de cercanía (closeness centrality):	33
4.4.6 Grado de Intermediación (betweenness):	34
4.4.7 Centralidad de Autovector (Eigenvector Centrality):	34
4.4.8 Cierre Triádico (Triadic Closure):	35
4.4.9 Coeficiente de clustering local:	35
4.5 Medidas de estructura de red	36
4.5.1 Componentes	36
4.5.2 Distribución de grado (degree distribution)	36
4.5.3 Distancia Promedio (Average Distance)	37

4.5.4 Densidad	37
4.5.5 Coeficiente de afinidad selectiva (Assortativity)	37
4.5.6 Coeficiente de clustering global y transitividad (transitivity)	38
4.5.7 Promedio de caminos más cortos (average shortest path length)	39
4.5.8 Diámetro y Radio	39
4.5.9 Modularidad (Modularity) y heurística de Louvain para la detección de comunidades	39

Resultados y discusión

5 Análisis de usuarios, comentarios y lugares	41
5.1 Descripción de usuarios y sus amigos	41
5.2 Comentarios	42
5.2.1 Comentarios según día y hora	48
5.3 Indicadores de uso de la plataforma	50
5.4 Lugares (venues)	52
5.5 Características de las comunas según lugares y comentarios	56
6 Análisis de la red social de Foursquare de comentadores de lugares en CABA	59
6.1 Comparación grafo extendido y grafo base	59
6.2 Descripción de grafo original (GO)	60
6.3 Resiliencia	64
6.4 Afinidad selectiva	65
6.5 Redes de escala libre y ley de potencias	69
6.6 Un mundo pequeño	73
6.7 Estructura comunitaria	77
7 Vínculos y espacialidad	83
7.1 Caracterización geoespacial de la red social	83
7.2 Calculo de centroides	85
7.3 Cálculo de distancia intra centroide y entre centroides	86
7.3.1 Distancia intra centroides	87
7.3.2 Distancia entre centroides	88
7.3.3 Dispersión intra centroide y entre centroides	90
7.4 Distancia geográfica y social	91
7.4.1 Distancia geográfica entre amigas y no amigas	91
7.4.2 Distancia Social y geográfica entre amigas	93
7.5 Probabilidad de Amistad	96
7.6 Resultados y discusión sobre distancia social y geográfica entre amigos	97

Conclusiones y trabajos futuros

Conclusión 100

Trabajos futuros 101

Bibliografía 104

Introducción

1 Introducción

1.1 Sociabilidad en línea y redes sociales

En las últimas décadas ha habido una sostenida expansión del acceso a Internet en toda América Latina, sin embargo, la sociabilidad¹ a través de medios virtuales y redes sociales en plataformas en línea han sido escasamente estudiadas en el país. Las plataformas y medios virtuales comprenden una parte de las comunicaciones que impacta en forma concreta en el comportamiento e interacciones de las personas entre sí.

La conceptualización de los sistemas sociales como grafos ha facilitado el estudio de los patrones de interacción social. Si bien diversos estudios indican que las comunicaciones en línea funcionan socialmente como otros tipos de interacciones tales como conversaciones telefónicas o en persona [1], existen también características propias de las redes sociales en Internet.

El uso de Internet y la constitución de interacciones y redes online tienen implicancia y continuidad en las experiencias y vida social de los individuos. El ámbito virtual es percibido por sus usuarios como un lugar real, donde se desenvuelven nuevas formas de interacción social y las personas de una red social en línea pueden transformar su estructura social a raíz de su localización al momento de hacer uso de las redes sociales.

Mientras en la década de 1990 el uso de Internet estaba restringido a una porción reducida de la población, en la actualidad se presenta como un medio de uso extendido para diferentes propósitos y en diferentes estratos socioeconómicos. Si bien en Argentina existe una brecha digital² persistente, en el año 2018 la penetración de Internet fue del 93,1% de la población, siendo la tasa más alta de Sudamérica [2]. Según datos del INDEC [3], considerando el último trimestre del 2019, se registró que el 60,9% de los hogares urbanos tenía acceso a una computadora y el 82,9%, a internet. Además, los datos muestran que, en la Argentina, 84 de cada 100 personas emplean teléfono celular y 80 de cada 100 utilizan internet.

En relación con la creciente importancia de Internet y las TICs (Tecnologías de la

¹ Según Simmel [5] se entiende por sociabilidad la forma lúdica de socialización que "...no tiene una finalidad material, no tiene contenido ni resultado que estuviera, por así decir, fuera del momento sociable como tal, se apoya por completo en las personalidades; no se persigue nada más que el estar satisfecho de este momento...". Se utilizará el concepto de "sociabilidad" sobre "socialidad" para hacer énfasis en la acepción de ser sociable y el trato con personas.

² Brecha digital o "digital divide" es una expresión que hace referencia a la diferencia socioeconómica entre aquellas personas y comunidades que tienen acceso a Internet y aquellas que no. Este término también hace referencia a las diferencias que hay entre grupos según su capacidad para utilizar las TIC de forma eficaz, debido a los distintos niveles de alfabetización y capacidad tecnológica.

Información y Comunicación) en todo aspecto de la vida social, el Estado Argentino en los últimos años ha desarrollado políticas para la alfabetización digital orientados a personas de diferentes rangos etarios, así como planes oficiales para facilitar el acceso a computadoras³.

La expansión de Internet en diversos campos de prácticas sociales posibilitó la utilización de este medio como forma de recolección de datos y análisis de distintos fenómenos sociales emergentes, tanto del medio online como offline.

Inicialmente, en las denominadas primera y segunda época de los estudios sociales sobre Internet, las investigaciones se centraron en describir los usos del medio, las características sociodemográficas de sus usuarios y las interacciones sociales online. Actualmente, en la llamada tercera época de este campo de estudios, el foco de atención se ha desplazado principalmente hacia los aspectos propios de Internet en la vida cotidiana de los usuarios, los vínculos y lazos de las comunidades online con las relaciones offline y la exploración de las modalidades de relaciones sociales que estaría promoviendo o habilitando Internet [4].

En este marco, las redes sociales basadas en ubicación (LSBN, derivado del término en inglés *Location-Based Social Networks*) surgen como un medio propicio para la investigación de la sociabilidad en el cruce de lo on-line y el contexto off-line de sus usuarios, así como la movilidad y patrones espacio temporales de los usuarios de las aplicaciones.

Desde este punto de vista, realizar un primer abordaje en Argentina del estudio de redes sociales localizadas con una perspectiva de minería de datos, en tanto metodologías y herramientas, tiene un carácter reciente en el ámbito local.

1.2 Extensión de las redes sociales localizadas

La localización es uno de los componentes principales del contexto de las personas y a partir de la cual se pueden estudiar cuestiones tales como intereses individuales, comportamiento, trayectoria espacio temporal y relaciones con otras personas.

Las LBSN se basan en la localización de sus usuarios y funcionan como un modo de mediación entre lo social, lo virtual y las características físicas de las vidas diarias de sus usuarios.

Investigadoras dan cuenta de una nueva concepción del espacio geográfico que involucra las dimensiones de lo virtual, lo social y lo físico donde la representación virtual del espacio físico sirve como un puente de mediación entre el mundo físico, la persona, el universo externo de información y la esfera social [6].

³ Mientras que hace años el PAMI lanzó el "Programa de acceso a computadoras para la tercera edad", también se crea el "Programa Conectar Igualdad com.ar" mediante el decreto 459/10 con el fin de proporcionar una computadora a cada alumno y docente de educación secundaria pública y formación docente. Éste último plan contempla que terminado el ciclo escolar completo cada chico pasa a ser propietario de una netbook.

Es con el advenimiento de los teléfonos celulares y dispositivos móviles así como la incorporación de internet en los mismos que se profundiza la popularización de los servicios basados en la localización (LBS) [7] como también las LBSN. El desarrollo de las tecnologías de comunicación móviles es una base necesaria para el uso extendido de las mencionadas LBSN.

Para el periodo 2009-2015 considerados en este trabajo, se puede dar cuenta de algunos indicadores clave para entender el alcance de la telefonía móvil y el acceso a Internet a través de la misma. A nivel internacional, para el 2011 más de la mitad de los usuarios de teléfonos móviles de Estados Unidos habían usado servicios basados en localización de algún tipo, con usuarios principalmente del rango etario de 18 a 29 años. Este número baja al 12% si consideramos aquellos que utilizan su teléfono para LBSN tales como Foursquare o Gowalla [9]. A nivel local, tomando en cuenta datos del Banco Mundial [10] podemos observar que la proporción de hogares que poseían telefonía móvil en Argentina y que hicieron uso de Internet desde el mismo fue del 48% para el año 2010. En cuanto a producción estadística de parte del estado nacional, podemos encontrar los datos obtenidos en el tercer trimestre de 2011 en el estudio ENTIC [11] cuya estimación se extiende al total de la población residente en hogares particulares urbanos en localidades de 2.000 y más habitantes. El objetivo de esta fue estimar el acceso a las TICs en el hogar y su utilización para el análisis de la inclusión digital en los hogares del país. En la misma se advierte que el 85,6% de los hogares disponen de un celular, mientras que el 52,8% de los hogares urbanos dispone de computadora y solo el 43,8% dispone de Internet. A nivel nacional se advierten notables diferencias regionales como por ejemplo la Ciudad Autónoma de Buenos Aires que se encuentra por encima del promedio total.

Si bien no existen datos respecto a la penetración de las redes sociales en la Argentina ni de las características sociodemográficas de sus usuarios para el periodo analizado en el trabajo, si existen indicadores para el 2020. Para junio de ese año, la cantidad de usuarios de Facebook en Argentina ascendió al número de 33.400.000 usuarios, lo que representa un 73.9% de la población. Si bien la red social mencionada no se cataloga como una LBSN, la misma da cuenta de la expansión de las redes sociales en línea en Argentina.

1.3 Redes sociales localizadas: el caso de Foursquare

Aunque no existen datos estadísticos sobre la popularidad de las LBSN en la República Argentina ni trabajos académicos al respecto, se reconoce a Foursquare como la plataforma más extendida a nivel mundial⁴ durante el periodo analizado en

⁴ En su página (<https://foursquare.com/about>) se especificaba que más de 50 millones de personas han hecho checkin en 6 billones de *venues*.

este trabajo.

La propuesta conceptual de la plataforma es la de recomendar *venues*⁵ (lugares) al usuario de acuerdo con sus gustos, hábitos, necesidades y localización geográfica en el momento de la búsqueda, ranking de los *venues* y comportamiento similar de los amigos. El lema de la plataforma es “*Hemos reinventado la búsqueda local, te ayudamos a encontrar el lugar perfecto, en cualquier lado que estés*” [13].

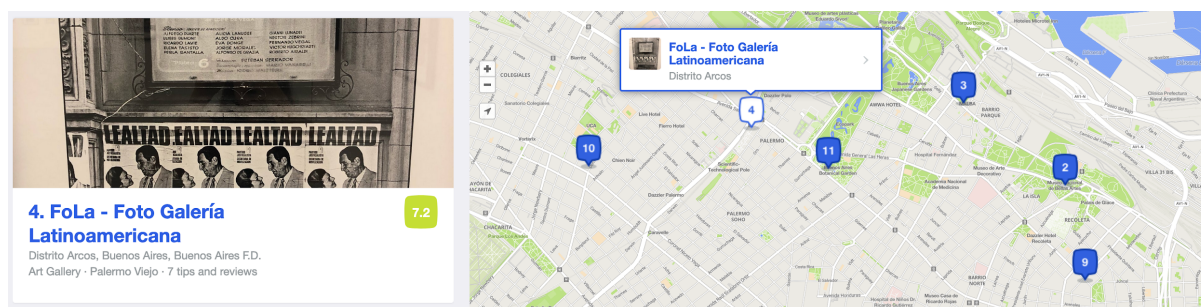


Figura 1.1: Ejemplo de lugares recomendados por la aplicación en base a una búsqueda y ubicación del usuario. En azul lugares recomendados y su ubicación en el mapa, en blanco la selección de una de las recomendaciones.

En la mencionada LBSN los usuarios pueden registrarse y compartir su localización geográfica marcando su presencia (*check-in*) en un sitio específico (*venue*) que esté localizado físicamente cerca de donde se encuentra el usuario en ese momento. Además de realizar *check-in* en *venues*, los usuarios pueden leer y publicar comentarios de sus experiencias, que serán vistos por futuros visitantes y los representantes de la localización. Los comentarios, como por ejemplo sugerir el mejor plato de un restaurante o quejas del servicio recibido, sirven como recomendaciones para los demás usuarios a la hora de elegir lugares a visitar. Cada usuario posee en su perfil público la historia de todos los comentarios que ha publicado, con la información del lugar asociado y la fecha y horario del mismo. Además de los comentarios, los usuarios pueden compartir fotos sobre los lugares, listas de lugares a visitar y calificar numéricamente las mismas, entre otras características complementarias.

Es importante aclarar que, a mediados del 2014, Foursquare lanza una aplicación dedicada a registrar *check-ins* denominada Swarm, y en agosto del mismo año la opción de *check-in* es removida de la aplicación Foursquare. Si bien Swarm está focalizada en dejar *check-ins* en los lugares que un usuario visita, y por ende posee mucha más información geolocalizada sobre el comportamiento de los usuarios que

⁵ Un *venue* representa cualquier localización física, tal como un restaurante, una universidad, una plaza o un aeropuerto en donde un usuario puede registrarse, hacer “*check in*” y publicitar su ubicación en una red social on-line.

Foursquare, no es considerada en esta tesis debido a que la información no es accesible públicamente.

1.4 Objetivos

En este trabajo se describen las características y dinámicas de los usuarios de la plataforma Foursquare dentro de la Ciudad Autónoma de Buenos Aires (CABA), así como las redes sociales geolocalizadas que conforman y el espacio en el que toman lugar. Se toma como marco temporal para el análisis al período comprendido entre diciembre del 2009 y abril del 2015.

Un objetivo derivado es crear un corpus de conocimiento lo suficientemente rico sobre redes sociales localizadas en Argentina que sirva de base de conocimiento, disparador de futuros trabajos de investigación y su utilización para el diseño de políticas públicas o aplicaciones comerciales.

Materiales y métodos

2 Elementos e instrumentos empleados

2.1 Descripción de métodos para la recolección de datos

Para la tesis se trabajó con datos recolectados de la plataforma de Foursquare, en un período que comenzó en febrero del 2014 y terminó en marzo del 2015.

La obtención de datos se realizó a través de la API⁶ de la plataforma, ya que la misma provee métodos para acceder a distintos aspectos tales como consultar información sobre *venues* (lugares), *tips* (comentarios) o usuarios. A través de una interfaz REST⁷ y autenticación mediante OAuth⁸ fue posible obtener en formato JSON⁹ el resultado de las consultas que se le realizaron a Foursquare, las cuales luego se extrajeron partes específicas del contenido a través JSONPATH¹⁰.

Si bien la mayor parte de la información registrada es de acceso público en la plataforma, la API del sistema tiene diversas restricciones en relación con la cantidad de consultas, límites de datos devueltos por consulta y disponibilidad del sistema¹¹. Esta particularidad hizo el proceso de recolección de datos sujeto a distintos inconvenientes técnicos que fueron resueltos *ad hoc*. Una de las principales dificultades fue lidiar con los límites diarios, por hora y por minuto para la recolección de información, por lo que el proceso fue a lo largo de más de un año. Este proceso de recolección extensa en el tiempo también introdujo diferencias en los datos, por lo que hubo un trabajo muy extenso de limpieza y reconciliación de información¹².

Un aspecto importante en la recolección de los datos fue la de secuenciar las diferentes consultas realizadas a la API de Foursquare, debido a la interdependencia entre los objetos con los que se trabajó. La construcción del conjunto de datos comprendió el uso de distintos objetos interrelacionados entre sí,

⁶ Application Programming Interface (API) es un conjunto de rutinas, protocolos y herramientas para construir aplicaciones de software. La API de Foursquare se encuentra documentada en:

<https://developer.foursquare.com/>

⁷ Representational state transfer (REST) consiste en un conjunto coordinado de componentes, conectores y elementos de datos dentro de un sistema distribuido de hypermedia.

⁸ OAuth es un protocolo abierto para permitir autorizaciones seguras en una manera simple y estandarizada desde aplicaciones web, móviles y de escritorio.

⁹ JSON es un estándar abierto alternativo al XML que usa texto como medio de transmisión de objetos de datos. El mismo consiste de pares valor-atributo y es usado principalmente para transmitir información entre un servidor y una aplicación web.

¹⁰ No es un estándar pero ofrece maneras de consultar estructuras JSON directamente sin necesidad de convertir a XML.

¹¹ Al momento de recolección de datos la API era parcialmente abierta, mientras que a este momento (Octubre del 2020) la misma ha mutado en sus características y ha sido considerablemente restringida.

¹² Un ejemplo del tipo de diferencias encontradas, fue al momento de recolección de los usuarios que habían dejado un comentario. La cantidad de amigos y comentarios de cada uno de los usuarios, iba cambiando en el tiempo a medida que se consultaron diferentes objetos de la plataforma.

tanto a nivel analítico como por características de la API, tales como *venues*, *tips*, usuarios comentadores, amigos de usuarios, detalles de los *venues*, detalles de usuarios, etc.

Para tal tarea se eligió una herramienta de Extracción, Transformación y carga de datos (ETL) de código abierto, denominada Talend [15]. La misma facilitó la conexión de distintas tecnologías utilizadas en el proyecto posibilitando el uso de distintos lenguajes y cierto grado de abstracción para el control de errores y orquestación de los procesos, debido a la complejidad de articular distintos pasos en la recolección de los datos.

Se elaboraron 12 procesos para extraer datos de distintos objetos de Foursquare, transformarlos y cargarlos a una base de datos diseñada en PostgreSQL¹³ a tal efecto, la cual será más detallada en la sección 2.3.

A modo de ejemplo, en la figura 2.1 se puede observar el diseño a alto nivel de uno de los procedimientos que fue ejecutado durante el periodo de recolección de datos para la extracción de comentarios de *venues* de la plataforma de Foursquare.

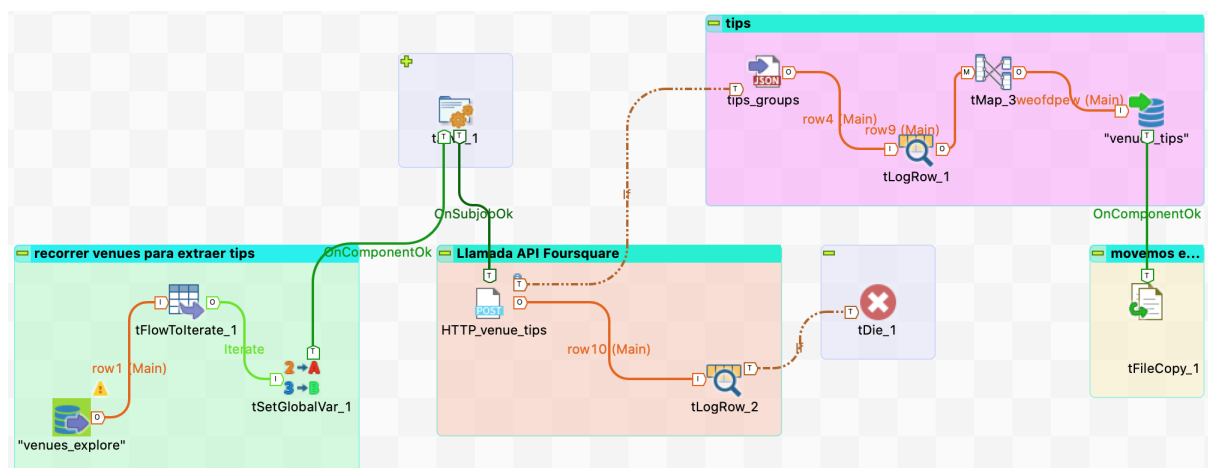


Figura 2.1: Proceso con acciones a alto nivel para la extracción de datos de un objeto de la plataforma de Foursquare

En la figura 2.1, observando de izquierda a derecha en la primer sección en color verde, el proceso itera sobre el resultado de la siguiente consulta a la base de datos:

```
select distinct(vee.id) as venue_id, random() rnd
```

¹³ PostgreSQL (<https://www.postgresql.org/>) es un sistema de manejo de base de datos relacional con énfasis en la extensibilidad y cumplimiento de estándares. Actualmente se considera a la misma como la base de datos de código abierto con más reputación en confiabilidad e integridad.

```

from fs.public.venues_explore_cf vee
inner join
    (select id, count(1) as cant
     from fs.public.venues_tips
     group by 1) vetip
on vee.id = vetip.id
where vee.location_cc = 'AR'
    and cast(vee."\"stats_tipCount\"" as int) > 19
    and vetip.cant < cast(vee."\"stats_tipCount\"" as int)
order by rnd

```

La consulta a la base de datos devuelve aquellos *venues* que habían sido previamente recolectados y que tuvieran registrados al menos 20 comentarios. En la segunda sección de la figura 2.1, en naranja claro, se hace una llamada a la API para consultar los comentarios para cada uno de estos *venues*. Esta llamada es diferente a la utilizada previamente en otro proceso de recolección de comentarios, ya que este método en particular permite retornar más de 19 comentarios por *venue*. La llamada a la API de Foursquare en este paso consiste en:

```

https://api.foursquare.com/v2/venues/" +(String)globalMap.get("globalid")+"/tips?sort=
recent&oauth_token=X&v=20140924&limit=500

```

En esa cadena de texto la variable “*globalid*”, contiene el valor de “*venue_id*” el cual representa el identificador único de un lugar, mientras que “*oauth_token*” es el código autenticado de OAuth (necesario para interactuar con la API) y “*v*” representa la versión de la API que está siendo usada en la consulta.

En la siguiente sección de la figura 2.1, en lila, en caso de éxito se descarga el JSON con la respuesta de Foursquare, o caso contrario se aborta el proceso. La consulta devuelve un resultado similar a la siguiente cadena JSON¹⁴ detallada en la figura 2.2.

¹⁴ La misma es a modo ilustrativo y fue considerablemente reducida para que sea legible, no responde al resultado de la consulta ejemplificada. El nombre del lugar fue ofuscado.

```

{
  "meta": {
    "code": 200
  },
  "response": {
    "venue": {
      "id": "4dcela6745ddbe15f881522f",
      "name": "Indu XYZ",
      "location": {
        "lat": -34.535725,
        "lng": -58.468645,
        "cc": "AR",
        "city": "Baires",
        "formattedAddress": [
          "Baires",
          "Buenos Aires F.D.",
          "Argentina"
        ]
      },
      "categories": [
        {
          "id": "4bf58dd8d48988d1le941735",
          "name": "Cocktail Bar",
          "pluralName": "Cocktail Bars"
        }
      ],
      "stats": {
        "checkinsCount": 2,
        "usersCount": 2,
        "tipCount": 0
      },
      "likes": {
        "count": 0
      },
      "createdAt": 1305352807,
      "mayor": {
        "count": 1
      },
      "tips": {
        "count": 2,
        "groups": [
          {
            "type": "others",
            "name": "Tips from others",
            "count": 0
          }
        ]
      },
      "timeZone": "America\\Argentina\\Buenos_Aires"
    }
  }
}

```

Figura 2.2: Ejemplo de respuesta simplificada de una consulta a la API de Foursquare

El resultado es a su vez consultado mediante JSONPath y se carga a una tabla de la base de datos identificando los atributos de destino con los de la fuente. Por ejemplo se puede averiguar la ciudad del lugar mediante la sentencia “\$.response.venue.location.city”, que en el caso de este ejemplo de la figura 2.2 se obtiene el valor “Baires”. Es importante acotar que el formato JSON permite guardar más de una instancia por variable, por lo que en el caso de lugares que tuvieran más de un comentario se puede obtener una lista de comentarios o varias instancias de otra variable en el mismo resultado.

La interfaz entre la base de datos y Foursquare a través de los procesos ETL facilitó la conversión de estructuras multidimensionales JSON a una base de datos relacional y que los datos posteriormente pudieran ser consultados y analizados fácilmente mediante consultas SQL.

Así mismo el hecho de guardar los datos obtenidos de Foursquare en una base de datos permitió articular las consultas realizadas a la API de la plataforma utilizando los valores de objetos alojados en la base de datos que fueron guardados en procesos anteriores. Por ejemplo, a la hora de querer saber las características de algún usuario que había dejado *tips* en los *venues* de Capital Federal fue necesario invocar un método específico de la API pasándole como argumento el identificador de usuario. De manera iterativa entonces se le consultaba a Foursquare para cada uno de los usuarios que habían comentado en cada uno de los *venues* ya registrados anteriormente en la base de datos.

2.2 Información geográfica

Se recorrió exhaustivamente el espacio de la Ciudad de Buenos Aires, recolectando todos los lugares registrados en la plataforma. La API de Foursquare posee un método para consultar el listado de lugares (*venues*) alrededor de un punto geográfico dado (expresado en coordenadas de latitud y longitud). Por cada localización consultada se obtiene un máximo de 50 lugares por consulta, los cuales están ubicados en un radio variable del punto consultado (en base a la densidad de lugares en el área).

Para recorrer todo el espacio de CABA fue necesario recurrir a información geográfica oficial que provea la localización de cada una de las cuadras de la ciudad. A tal efecto se trabajó con la base de datos *Manzanero* [15] provista por el área de datos abiertos del gobierno de la ciudad de Buenos Aires y desarrollado por el Ministerio de Desarrollo Urbano- Secretaría de Planeamiento- Dirección General de Registro de Obras y Catastro.

La misma está codificada en coordenadas Gauss-Krüger, por lo que se escogió QGIS¹⁵ como aplicación de código abierto capaz de manejar información geográfica

¹⁵ QGIS (<https://qgis.org/>) es una aplicación multi plataforma de código abierto para el manejo de información geográfica que permite visualizar, editar y analizar los datos.

en diferentes formatos y facilitar su transformación e interacción con bases de datos. El uso del mismo fue también útil para visualizar el resultado del proceso de recolección de datos sobre el territorio de la capital federal a modo de controlar que se cubra la recolección de *venues* en todas las áreas de la ciudad, entre otras cosas.

En la figura 2.3 se puede visualizar la composición de datos de Manzanero, una representación gráfica del mapa, y un detalle de cada una de las unidades (en general responden a regiones que representan “manzanas”) que fueron tomadas en cuenta para realizar la recolección de *venues* de Foursquare.

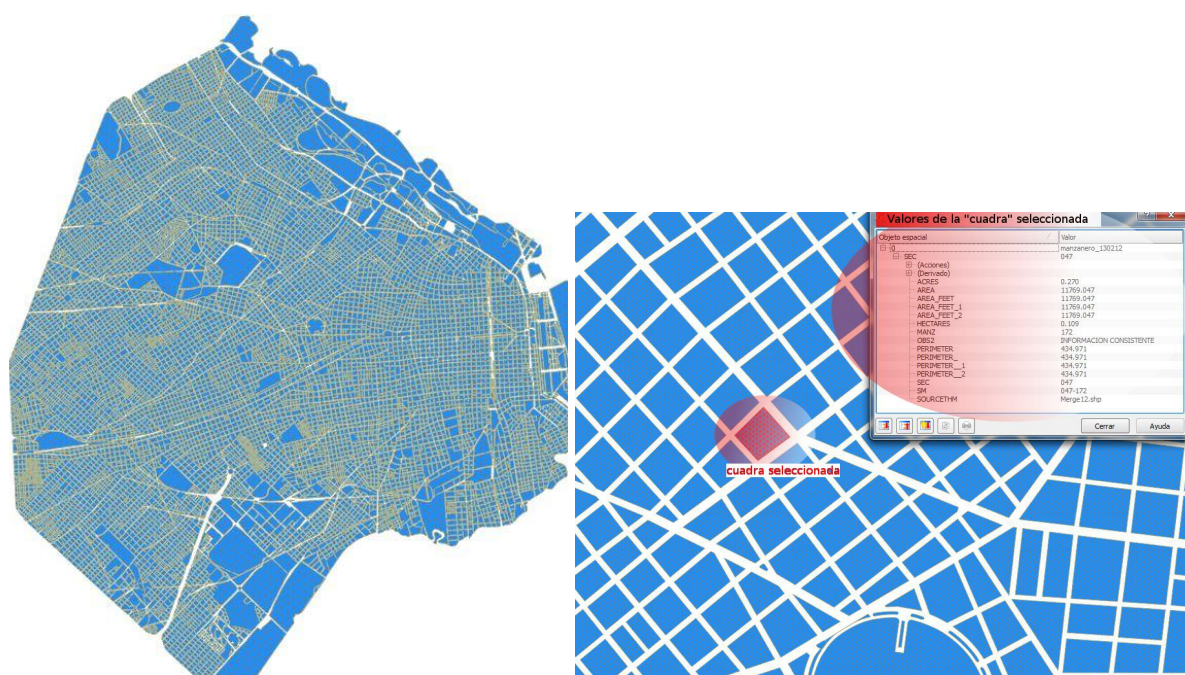


Figura 2.3: Izquierda: Mapa de CABA utilizado a nivel de manzanas. Derecha: Detalle de una cuadra del mapa

El mapa se compone de 12373 unidades (manzanas, cuadras o áreas geográficas en que el conjunto de datos subdivide el espacio de la ciudad de Buenos Aires). Se decidió calcular el centroide geográfico¹⁶ de cada una de estas unidades para luego poder consultar a Foursquare los lugares aledaños a cada uno de estos puntos del espacio. Es decir, en el trabajo se recorrió cada manzana de la ciudad para las cuales se obtuvieron como máximo 50 lugares cercanos a estas. Cabe aclarar que los 50 lugares que devolvieron cada una de las consultas contenían lugares duplicados en algún número, por lo que se considera que fueron recolectados todos los lugares registrados en la plataforma. Debido a este mecanismo de muestreo de 50 localizaciones cercanas a cada cuadra de CABA, se obtuvieron 633 *venues* que fueron incluidos en el análisis, pero

¹⁶ Es el punto que es el centro geométrico de masa de una geometría.

no fue posible luego asignarles una comuna ya que se encontraban fuera del área de CABA. Estos espacios estaban ubicados levemente fuera de los límites, como el siguiente ejemplo (figura 2.4), que representa una clínica de salud ubicada en la localidad inmediata a la ciudad.

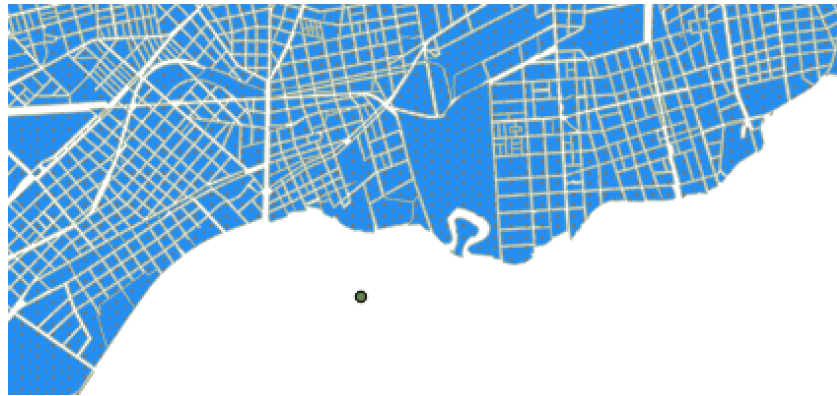


Figura 2.4: Detalle de lugar fuera del ámbito de CABA

En cuanto al manejo de datos geográficos se hizo necesario el uso de una extensión para la base de datos denominada PostGIS¹⁷.

Esta extensión permite realizar cálculos espaciales y además fue utilizado para relacionar los datos geolocalizados de Foursquare con la base de datos del Gobierno de la ciudad, la posterior clasificación en barrios y comunas de los lugares en base a su localización geográfica, la medición de distancia entre comentarios, etc.

2.3 Diseño de base de datos

En cuanto al diseño de la base de datos, se ha tenido en cuenta que cada objeto “teórico” de la plataforma (como por ejemplo *venue*, usuario o comentario) era conveniente traducirlo a múltiples objetos en la base de datos relacional.

Así por ejemplo la información sobre los lugares se guardó en 7 tablas distintas, ya que los datos devueltos por la plataforma para los lugares contienen múltiples dimensiones y jerarquías, dándose casos donde puede haber más de una categoría o atributo para un mismo lugar (*venue*). A su vez, existen más de un método para acceder a los lugares según se quisiera consultar información básica, atributos, comentarios, etc.

En la figura 2.5 se observa un diagrama simplificado de los principales objetos teóricos con los que se trabajará a lo largo de la tesis y algunos de los atributos de

¹⁷ PostGIS funciona como una extensión de base de datos espacial para PostgreSQL facilitando realizar consultas de SQL sobre información localizada geográficamente.

las personas, venues y su interrelación.

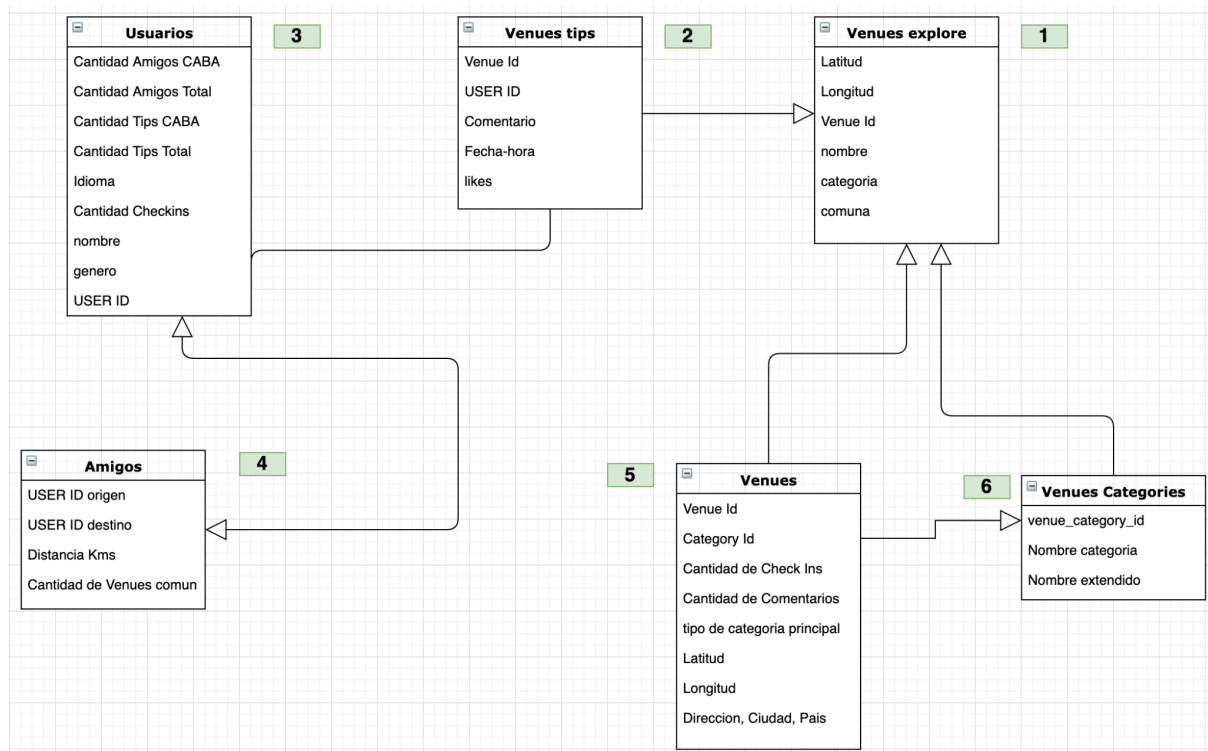


Figura 2.5: Esquema de base de datos.

Versión ilustrativa de las principales tablas y campos

2.4 Procesos y componentes de software

En la figura 2.5 podemos observar un resumen de los procesos a alto nivel y soportes de datos utilizados para este trabajo.

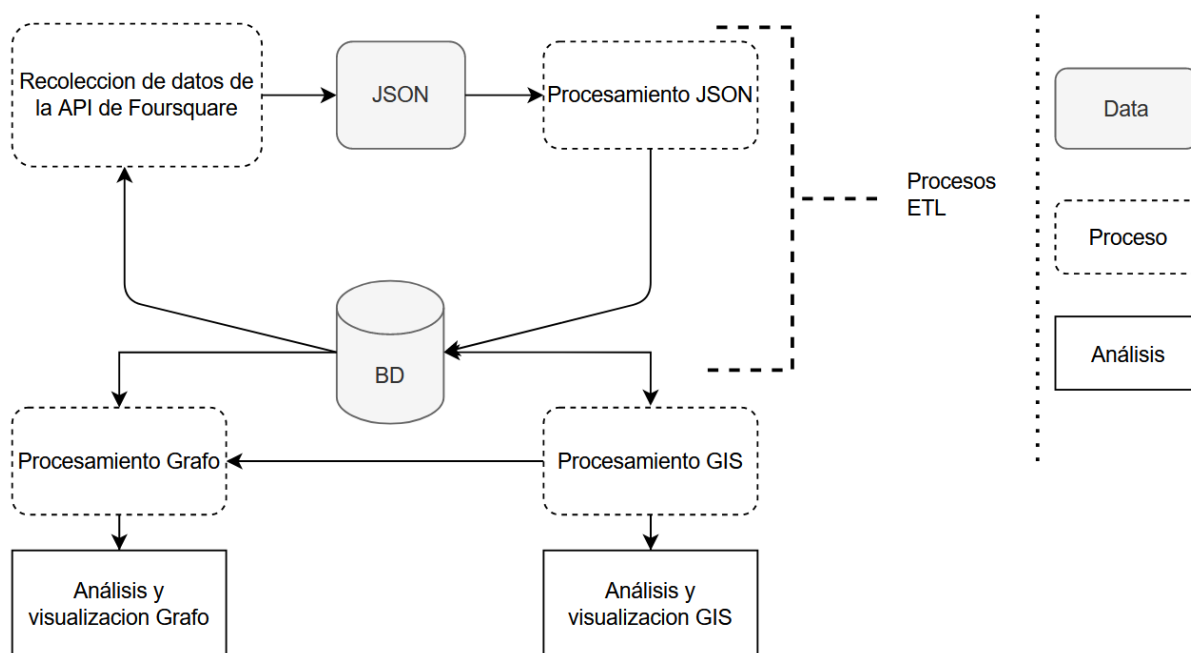


Figura 2.5: Diagrama de procesos, información y análisis realizados a alto nivel

La conjunción de QGIS y PostGIS permitieron manipular la base de datos de manzanas de la CABA, convertirla a coordenadas latitud-longitud y luego transformarla a una base de datos en PostgreSQL para que pueda hacerse cálculos espaciales.

A su vez, su disponibilidad en la base de datos permitió que se usaran los valores de los centroides geográficos de las unidades de Manzanero como variables a ser consultada a Foursquare. Las respuestas JSON para las distintas consultas fueron guardadas en la base de datos nuevamente y desde donde posteriormente se analizaron, procesaron, y validaron los datos.

Respecto al software utilizado para el análisis de redes sociales se escogió Python para el procesamiento (debido a su versatilidad en términos de lenguaje y gran disponibilidad de paquetes para computación científica), SQL y el software de exploración y visualización de redes denominado GEPHI¹⁸.

En términos de paquetes para trabajo con grafos y análisis de redes sociales, se utilizó principalmente NetworkX¹⁹. En ambos casos los softwares fueron vinculados a la base de datos desarrollada en PostgreSQL para facilitar el procesamiento de los datos.

¹⁸ Gephi (<https://gephi.org>) es una plataforma interactiva de visualización y exploración para todo tipo de redes, sistemas complejos y grafos dinámicos y jerárquicos.

¹⁹ NetworkX (<https://networkx.github.io>) es un paquete de Python para la creación, manipulación y estudio de estructura, dinámicas y funciones de redes complejas.

2.5 Procedimiento de recolección de datos

A modo de resumir el procedimiento lógico utilizado para armar el conjunto de datos de Foursquare y alojarlo en la base de datos, se articula el siguiente orden en los procesos de recolección de datos:

1. Por cada punto correspondiente al centroide de cada área delimitada en el conjunto de datos Manzanero: Obtener mediante la API de Foursquare hasta 50 *venues* existentes alrededor de ese punto geográfico (en coordenadas de latitud y longitud). Recorrer las 12373 cuadras.
2. Una vez obtenidos los identificadores de los *venues* correspondientes a CABA, deduplicando y resolviendo inconsistencias, obtener las características de cada uno de estos *venues* (nombre, dirección, categoría de *venue*, coordenadas geográficas, cantidad de check-ins, cantidad de *tips*, etc). Esto requirió de la implementación de otros métodos distintos de la API.
3. En base a los *venues* identificados en el punto 2 que tengan registrados un comentario o más, proceder a recolectar los comentarios registrados en los mismos (texto del comentario, id de usuario que dejó el comentario, dispositivo usado, etc). Esto requiere la implementación de varios métodos de consultas distintos de la API de Foursquare.
4. Tomando en cuenta los usuarios que hayan dejado tips en los *venues* recolectados, proceder a obtener sus características básicas (Identificador de usuario, género, dirección de twitter, cantidad de amigos, cantidad de comentarios, etc). Esto requerirá de la implementación de otros métodos distintos de la API.
5. En base a los usuarios obtenidos en el punto 4, obtener la lista de amigos por cada uno de ellos (nombre del amigo/a, género, Identificador, cantidad de amigos de cada uno de sus amigos, etc). También para este punto, esto requirió de la implementación de varios métodos distintos de la API.

3 Procesamiento de datos

3.1 Detección y aislación de valores extremos

Sobre el total de personas que dejaron comentarios en CABA, se procedió a excluir aquellos usuarios que tenían valores extremos en sus métricas tales como cantidad de comentarios, ya que representan anomalías u *outliers*. La necesidad de excluir estos valores atípicos surgió durante el análisis inicial del grafo, ya que se detectaron personas que evidentemente eran *bots*²⁰ o promociones comerciales que no corresponden a usuarios reales, distorsionando muchas de las métricas que se usaron para el análisis. Los bots y agentes de publicidad en Foursquare, así como otras redes sociales, han sido detectados y analizados en diferentes estudios [16], y se encuentran ampliamente extendidos [17].

Dentro del siguiente cuadro podemos observar la distribución para algunas métricas en el grafo original antes de remover usuarios con valores extremos.

Medida	Comentarios Totales	Comentarios en CABA	Comentarios en CABA únicos por lugar	Amigos en CABA	Amigos Totales
Media	14,44	3,53	3,26	15,53	77,86
Desvío estándar	69,05	7,27	6,09	29,85	118,50
Mínimo	1	1	1	1	1
1er Cuartil	2	1	1	3	15
Mediana	5	2	1	7	38
3er Cuartil	12	3	3	16	97
Máximo	5441	402	293	639	1221

Tabla 3.1: Medidas de tendencia central para comentarios y amigos

Como podemos observar en la tabla 3.1, se encuentran valores máximos muy alejados de la media y la mediana para cada una de las variables.

Debido a la complejidad del conjunto de datos, se fueron probando los resultados de distintas técnicas tales como KNN (K vecinos más cercanos) o *Isolation forest*²¹ para evaluar métodos de aislamiento de *outliers* según múltiples variables y observando cuidadosamente los resultados.

²⁰ Aplicaciones que corren tareas automatizadas en Internet.

²¹ *Isolation forest* [18] fue el primer algoritmo en aislar anomalías usando árboles binarios, enfocado en detectar anomalías sin tener que construir un perfil de lo que son las instancias “normales”. Para aislar anomalías, el algoritmo genera recursivamente particiones del conjunto de datos seleccionando aleatoriamente un atributo y luego un valor de corte aleatorio del mismo.

Finalmente se decidió excluir los outliers univariadamente en base a su puntuación Z, descartando el 0,5% de los casos más alejado de la media para cada una de las variables. Esta decisión fue debido a que los resultados del método eran más fáciles de evaluar y el impacto del umbral era fácilmente interpretable y auditable debido al conocimiento del dataset.

Las variables y valores considerados para la exclusión de los valores extremos fueron aquellas personas que tuvieran más de 40 comentarios en CABA, más de 234 comentarios en total y más de 504 amigos en CABA, con lo cual se excluyeron 582 personas²², debido a que una misma persona podía ser clasificado como un valor extremo en más de una variable.

Dentro de los casos excluidos, se pudieron encontrar mayormente bots y publicidades, pero también algunos usuarios que tenían características inusuales, como por ejemplo pocos o nada de amigos y muchos comentarios, o viceversa.

A modo de ejemplo, los valores extremos encontrados tienen características tales como (Tabla 3.2):

Nombre de usuario	Cantidad de Amigos	Cantidad de Comentarios	Comentarios
restaurantex.com.ar	53	169	20% DE REEMBOLSO, ACUMULABLE, SIN TOPE, LOS JUEVES Y VIERNES, en xxxx & zzzz y todos los restaurantes del país abonando con tarjeta XXX.COM.AR YYYYCARD. VTO. 30/06/2013 ver condiciones
nombre Restaurante	0	79	BUEN VIERNES PARA TODOS!!! Es 26 y ya estamos pensando que vamos a comer todo el fin de semana!! Nosotros ya sabemos que vas a comer el lunes! Es 29 y en RestauranteX te esperamos con los ñoquis más ricos!
NombreX ApellidoY	3	72	Estos pibes son unos chorros 50 pesos este platito con queso salame y jamón no existe, XXXXX fue!!! En los alrededores está lleno de lugares 10 veces mejor

Tabla 3.2: Características de muestra de usuarios excluidos

Luego de la poda de usuarias con valores extremos del componente principal del grafo, quedaron 31385 usuarias, con las cuales se procederá el resto del análisis.

²² La determinación del 0,5% extremo sobre cada una de las variables fue realizada sobre el grafo total de usuarios que dejaron comentarios en CABA.

3.2 Cálculo de variables adicionales

Una vez identificado el grafo final, se procedió a la creación de variables calculadas que fueron utilizadas posteriormente para el análisis. Se mencionan a continuación algunas de estas variables calculadas:

- Cantidad de comentarios en CABA por usuaria
- Cantidad de comentarios sobre lugares únicos en CABA por usuaria
- Cantidad de amigos en CABA por usuaria
- Porcentaje de comentarios y amigos de capital federal respecto al total
- Centroides geográficos de la usuaria en base a sus comentarios
- Distancia entre el centroide geográfico de la usuaria y sus amigos
- Clasificación del lenguaje principal de los usuarios en base a sus comentarios
- Clasificación de los lugares en comunas y barrios según información oficial de CABA

Muchas de estas variables serán consideradas en distintas partes del análisis y explicadas en mayor profundidad.

Se detalla a continuación la clasificación de idiomas de los usuarios.

3.3 Clasificación de Idioma

La ubicación geográfica declarada por los usuarios en su perfil en Foursquare se trata de un campo abierto que puede contener cualquier o ningún valor (ej: "NY", "mi casa", "Argentina", "CABA", "Distrito Federal", "123", "bares", "Cuauhtémoc, DF", etc), por lo que no fue factible atribuir una ubicación donde vive o sociabiliza cada persona utilizando esta información.

Si bien no fue posible atribuirle una ubicación geográfica en base a su perfil, se pudo utilizar el contenido textual de los comentarios que estas personas dejaron en CABA para ser clasificados entonces según su idioma utilizado.

Para la clasificación de idiomas se utilizó un modelo [19] de clasificación entrenado sobre wikipedia, en el que demostró un 99% de fiabilidad en la clasificación de los idiomas de los artículos.

Cabe aclarar que se observó cierta distorsión en la clasificación de comentarios con escasa cantidad de palabras, utilización de palabras de diferentes lenguajes en el mismo comentario o uso de símbolos y emoticones. En estos casos el modelo produjo resultados ambiguos en una porción menor de comentarios hechos presuntamente en castellano, pero clasificados en otros lenguajes como catalán, inglés u otros.

A continuación podemos observar algunos ejemplos de comentarios de personas presuntamente locales o de habla castellana que fueron clasificados como otros

lenguajes:

"Wifi password: e9f99999"
"Salieron los drinks"!!"
"Hermosooooo"
"2x1 Club LaXYZ"
"Proxy.campus.frba.xyz.edu.ar (proxy)"
"Tel 45555555"
"Los falocutus son im-per-di-bles!"
"Pringlesssssss"
"Abierta 24hs! Open 24hs!"
"Ginger Martini 🍷"

Adicionalmente a la ambigüedad descrita, 118 comentarios fueron inclasificables, como podemos observar estos ejemplos a continuación:

"👍👍👍👍👍👍👍👍👍👍👍"
"http://ar.linkedin.com/in/xxxxxxxxxxx"

Finalmente se identificó una decena de idiomas distintos, sin embargo se encontraron 3 grandes grupos: castellano, inglés y portugués. A continuación se observa la distribución de comentarios según los idiomas identificados, así como la cantidad de caracteres promedio que contienen los comentarios clasificados dentro de cada categoría.

Lenguaje clasificado	Porcentaje	Cantidad de comentarios	Cantidad de caracteres promedio
Castellano	79,45%	83514	71,77
Otros	9,09%	9557	27,02
Portugués	6,13%	6445	72,22
Inglés	5,21%	5478	63,02
Error	0,11%	118	30,63

Tabla 3.3: Tipificación de lenguajes en base a comentarios de los usuarios

Las categorías identificadas como Castellano, Portugués e Inglés, poseen una cantidad promedio de caracteres mucho mayores que las clasificaciones con error u otros idiomas, ya que el modelo de clasificación fue más preciso en comentarios extensos.

Los usuarios con múltiples comentarios pueden contener idiomas diferentes, por lo que se realizó código ad-hoc para escoger el idioma de más frecuencia para cada uno de los usuarios.

En la distribución de idiomas a la granularidad del usuario, se puede contemplar una reducción porcentual de los errores²³, la categoría de “otros” y “Castellano”, mientras que un aumento porcentual de “Inglés” y “Portugués”.

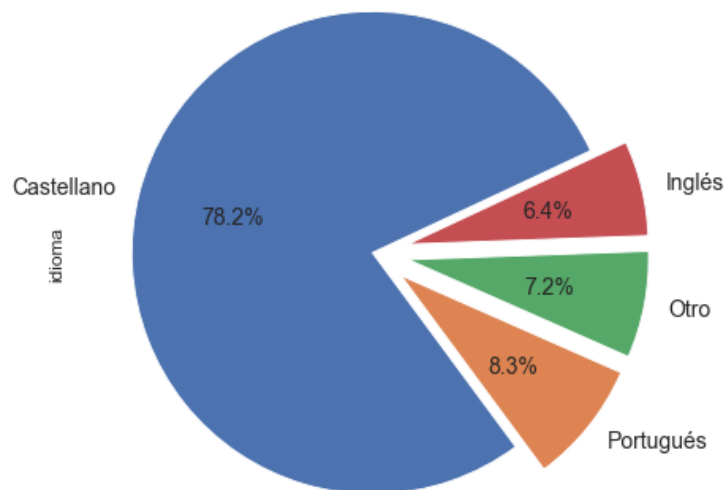


Figura 3.1: Distribución porcentual del lenguaje principal de los usuarios

A efectos prácticos de la tesis y facilitar el análisis se decidió dividir a los usuarios solamente en dos grupos: hispanos y no hispanos. En el primero se agrupó a todos los usuarios donde la mayor parte de sus comentarios fueran clasificados como castellano, el cual compone el 78,2%, mientras que en el segundo se incluye el 21,8% restante de los usuarios.

²³ La categoría fue ocultada del gráfico debido a que tiene un valor de 0,1%

4 Teoría y modelado de red social

4.1 Teoría de Grafos

Los grafos son estructuras matemáticas utilizadas para modelar y representar relaciones y conexiones entre objetos, procesos o sistemas. Un grafo simple G consiste de un conjunto finito no vacío $V(G)$ de elementos denominados vértices (también llamados nodos), y un conjunto finito $E(G)$ de pares no ordenados de elementos distintos de $V(G)$, denominados aristas [20].

Denominamos $V(G)$ al conjunto de vértices y $E(G)$ al conjunto de aristas de G . Una arista $\{v, w\}$ une los nodos v y w , y es usualmente abreviada como vw .

Por ejemplo, la siguiente figura (4.1) representa un grafo simple el cual tiene como conjunto de nodos $V(G)$ a $\{A,B,C\}$ y como conjunto de aristas $E(G)$ a las aristas AB , CB , AC .

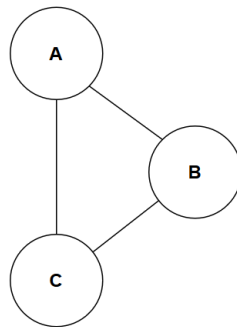


Figura 4.1: representación de grafo básico

Usando notación de conjuntos, la figura 4.1 es una representación del siguiente grafo:

- $V: = \{A, B, C\}$
- $E: = \{\{A, B\}, \{B, C\}, \{A, C\}\}$

Donde E consta de pares no ordenados de vértices, tales como $\{A, B\} \in E$, entonces se dice que A y B son adyacentes; y en el grafo se representa mediante un segmento que une dichos vértices.

Los grafos pueden ser usados para modelar variadas formas de relaciones y procesos en campos tan diversos como la biología, las ciencias de la computación, la sociología o la física, entre otros.

Si bien existen distintos tipos de grafos como dirigidos, multigrafos y mixtos, en esta tesis nos enfocaremos en los grafos no dirigidos²⁴ y sin ponderación para modelar la

²⁴ Se denomina grafo no dirigido a aquel que tiene sus aristas sin orientar. Según el problema a analizar, la representación de las relaciones entre los componentes de la red social puede ser

red social de comentadores de lugares de la plataforma Foursquare.

4.2 La red social de comentadores de CABA en Foursquare

Debido a las características de Foursquare²⁵ se utilizó un grafo no dirigido y sin peso en sus aristas, para modelar la red social entre los usuarios de dicha plataforma.

A objetivos de la tesis, se han recolectado las personas comentadoras en venues en el espacio de CABA y las amigas directas de estas.

Desde el punto de vista de un usuario en particular que ha dejado un comentario, la estructura de grafo es como la que se describe a continuación.

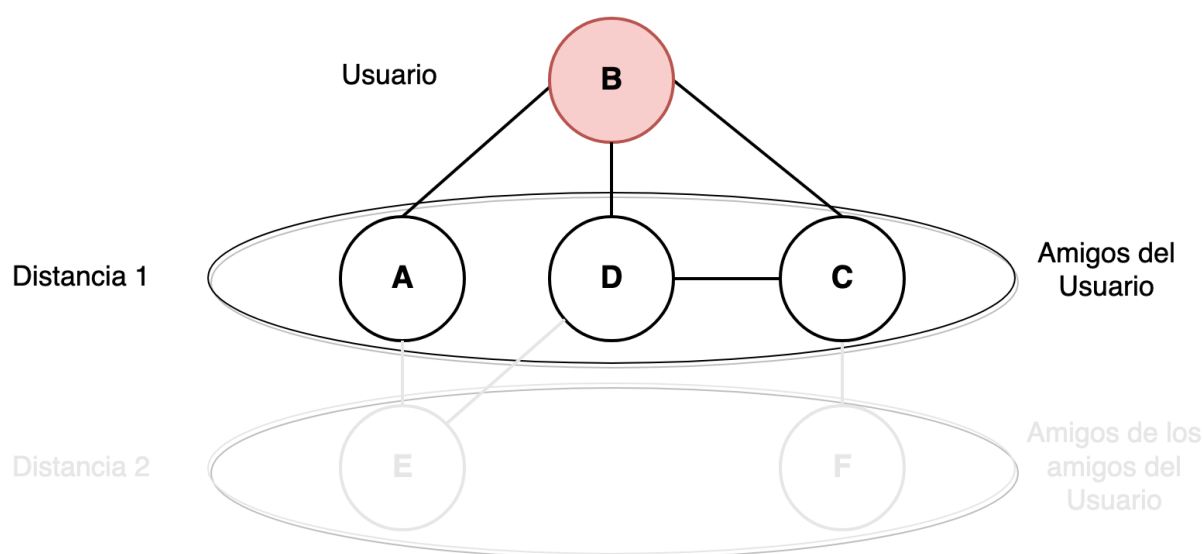


Figura 4.2: representación de grafo de Foursquare desde el punto de vista de un usuario en particular (B)

Si bien se ha recolectado solamente un nivel de distancia respecto a los usuarios comentadores, como se observa en la figura 4.2, las amigas de la comentadora a su vez pueden ser amigas entre sí. Este vínculo entre amigas comentadoras solamente se ha tomado en cuenta para los casos en que al menos una de las dos amigas directas también fue comentadora de un venue de CABA.

A modo de ejemplo, en la figura 4.2, B es un usuario comentador, C es su amiga directa que también es comentadora de un venue de CABA, mientras que D no es

realizada con grafos bipartitos, dirigidos, etc. Las relaciones de “amistad” entre los usuarios de Foursquare se enmarcan en la de grafos sin orientación entre sus nodos, ya que la propiedad de “amistad” que representa una arista está presente en los dos nodos que forman parte.

²⁵ Inicialmente y al momento de la recolección de datos la plataforma era una red social no dirigida ya que el vínculo de amistad era bi direccional. La herramienta ha cambiado en el tiempo y a agosto del 2020, independientemente de un vínculo de amistad, se permite ser un seguidor o tener seguidores uni direccionalmente, un modelo similar a la red social Twitter.

comentadora de CABA pero es amiga directa de B y C. En cambio, A no es comentadora pero es amiga directa de B, aunque no lo es de C. El escenario posible en que A y D sean amigas no fue tenido en cuenta en la fase de recolección de datos ya que implicaba un segundo grado de distancia respecto a B.

Tomando en cuenta esta relación entre amigas de amigas, se han construido dos grafos sobre los cuales se guiará todo el análisis:

- **Grafo Original (GO):** Contiene a las personas que han dejado un comentario en algún venue de CABA con todas sus relaciones (amigas de primer grado) que también hayan dejado algún comentario en CABA.

- **Grafo Extendido (GE):** Contiene a GO y además todas las amigas de primer grado de las personas comentadoras, así no hayan dejado comentarios en CABA.

A continuación se puede observar un esquema del grafo de usuarios junto con sus relación a los lugares (*venues*), representando cuales usuarios y relaciones son contemplados en cada uno de los grafos GO y GE.

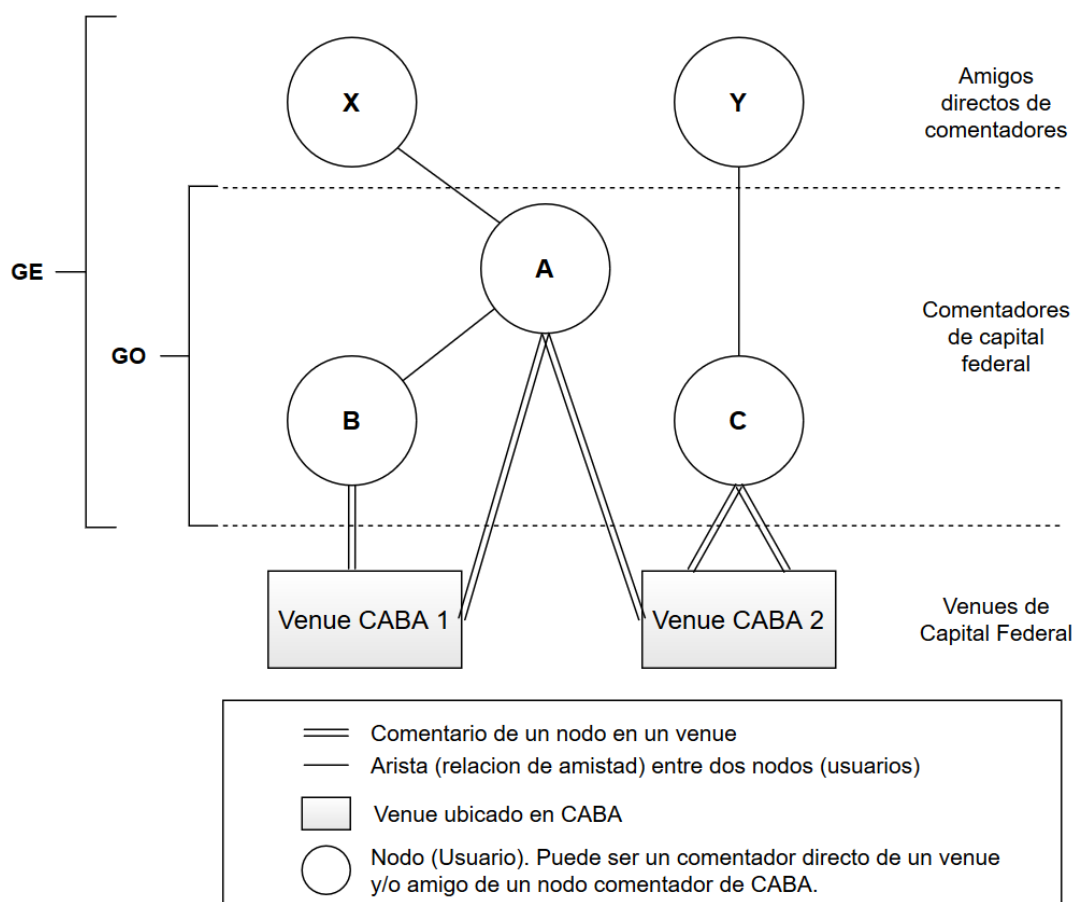


Figura 4.3: Esquema de Grafo extendido (GE) y Grafo Original (GO)

Podemos observar que una misma persona puede tener comentarios en diferentes lugares, así como una amiga de un comentador puede ser: 1) comentadora de un mismo lugar, 2) comentadora de un diferente venue o 3) ser amiga directa de primer grado de una persona comentadora pero no ha comentado en ningún venue de CABA.

Vale remarcar que esta relación entre lugares y usuarios podría ser modelada alternativamente con un grafo del tipo bipartito, en donde los venues serían un tipo distinto de nodos, pero dicho modelo no es el más apropiado para los objetivos de este trabajo en donde el foco está puesto sobre los usuarios que comentan y las relaciones que constituyen entre sí. A motivos de esta tesis, en los grafos los venues solo son considerados para identificar los usuarios analizados y las propiedades de sus amigos.

Finalmente, salvo explicitado, siempre se estará analizando el grafo GO, el cual contiene el objeto de análisis de este estudio: las personas comentadoras de venues de CABA y sus interrelaciones entre sí.

4.3 Análisis de Redes Sociales

La utilización de la teoría de grafos para estudiar las estructuras sociales se denomina comúnmente como Análisis de Redes Sociales (ARS) o Social Network Analysis (SNA), en donde se representa a los actores sociales como nodos y sus interconexiones como aristas.

Si bien el análisis de redes sociales tiene origen en teóricos fundadores de la sociología, como Durkheim y Simmel [21], los primeros métodos analíticos son recién introducidos en 1930 por Jacob Moreno para el análisis formal de las relaciones sociales [22].

Actualmente estas técnicas se han expandido a diversas disciplinas académicas y aplicaciones comerciales, con un especial crecimiento a partir del surgimiento de los sitios de redes sociales en línea (Facebook, Twitter, etc).

Usualmente el foco de análisis de redes sociales son las relaciones de los actores sociales, estudiando los patrones de una manera estructural o individual.

Utilizando las técnicas propias de SNA se pueden identificar la importancia de cada uno de los actores, comunidades e interacciones existentes entre los mismos [23].

4.4 Métricas y propiedades de las Redes Sociales

Existen diferentes niveles en que se pueden analizar las redes sociales, desde los actores y sus características a las relaciones diádicas, triádicas entre los actores, o al nivel de comunidades dentro de la red, así como la red en su totalidad.

Revisaremos a continuación algunas de estas medidas que serán usadas para describir los distintos grafos de redes sociales en este trabajo, sin necesidad de representarlas gráficamente. Las mismas pueden ser subdivididas de acuerdo con

el nivel de análisis, desde lo más molecular, como los actores sociales, hasta medidas descriptivas de toda la red en su conjunto.

4.5 Medidas de Centralidad

Las medidas de centralidad se usan comúnmente para identificar la importancia relativa de los nodos, calificados con valores dados por una función, como por ejemplo los actores más influyentes dentro de una red [24].

Las funciones de centralidad producen un ranking de cada uno de los nodos, de manera de poder ordenar los mismos en base a la medida evaluada. Es importante remarcar que, de querer usarse comparativamente entre grafos distintos, estas medidas deben ser primero normalizadas para tales motivos.

A continuación las medidas de centralidad más comunes:

4.4.1 Centralidad de grado (Degree)

Es el número de aristas o conexiones incidentes a un nodo dado. Una medida de centralidad a nivel individual, utilizada para identificar la cantidad de relaciones de cada individuo.

Para un grafo $G=(V,E)$, es expresado como $C(v) = \deg(v)$, donde $\deg(v)$ representa la cantidad de aristas que inciden en dicho vértice.

Más allá de su simplicidad, y que no toma en consideración la estructura global de la red, esta medida es central para determinar la influencia de un actor.

En este trabajo se obviaron aquellos nodos que tenían un grado (degree) = 0 ya que no fueron considerados en el componente principal analizado, ni tampoco existen los denominados *loops* (en donde un nodo se referencia con una arista a sí mismo).

4.4.2 Grado promedio (Average Degree):

Se usa para entender la conectividad global del grafo y nos da una idea de si hay muchos nodos muy conectados o si el nodo típico del grafo tiene una pequeña cantidad de aristas. Para calcular el grado promedio, se calcula con la suma de centralidad de grado para toda la red dividido por la cantidad de nodos totales.

4.4.3 Camino más corto (Shortest Path):

Un camino es una secuencia de nodos en donde pares consecutivos de nodos no repetidos son unidos por aristas. El camino más corto, también denominado camino geodésico (*geodesic path*), es el camino mínimo entre dos nodos.

A continuación (figura 4.4) vemos un ejemplo en donde se grafica el camino más

corto entre los nodos A y D.

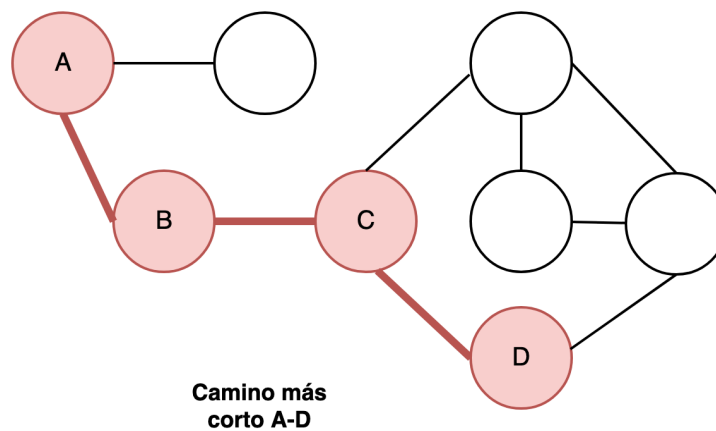


Figura 4.4: Camino más corto entre nodos A y D

4.4.4 Excentricidad (eccentricity):

Para un nodo dado v , la excentricidad está dada por la distancia máxima entre todos los caminos más cortos encontrados entre ese nodo v y cualquier otro nodo i integrante del grafo.

Esta medida será utilizada también para calcular el radio y diámetro (explicados en la próxima sección).

$$e_v = \max_{i \in V(G) \setminus v} d(v, i).$$

4.4.5 Centralidad de cercanía (closeness centrality):

Representa el grado de cercanía entre un nodo dado y el resto de nodos de la red. Es la inversa de la suma del promedio de distancia más corta entre un nodo dado con cada otro nodo en la red.

La centralidad de grado de cercanía en su versión normalizada para un nodo v dado de un grafo G con n nodos, y donde $d(u, v)$ representa la distancia más corta entre el nodo v y u , puede ser expresado como:

$$Cl_v = \frac{n - 1}{\sum_{u \in V(G) \setminus v} d(u, v)}.$$

La versión normalizada de la fórmula permite comparar grafos de diferentes tamaños y habitualmente el numerador para grafos grandes consiste solamente en n , sin restar 1.

A nivel analítico, el grado de cercanía es un buen indicador de la posición del actor en la red social, indicando cuán cerca está un actor de contactar al resto de la red.

4.4.6 Grado de Intermediación (betweenness):

Esta medida cuantifica el grado de intermediación de un nodo dado entre los caminos más cortos que conectan otros dos nodos.

Esta medida ayuda a identificar aquellos actores sociales que tienen un rol central en la difusión de información y cohesión en una red.

Habitualmente también son vitales en conectar diferentes regiones de una red o comunidades.

La intermediación de un nodo v , puede ser expresado como:

$$b_v = \sum_{s,t \in V(G) \setminus v} \frac{\sigma_{st}(v)}{\sigma_{st}},$$

donde σ_{st} denota la cantidad de caminos más cortos entre los vértices s y t , y $\sigma_{st}(v)$ expresa la cantidad de caminos más cortos pasando a través de v .

4.4.7 Centralidad de Autovector (Eigenvector Centrality):

Esta métrica se hizo especialmente conocida luego de ser usada como base para el algoritmo Pagerank de Google. La misma mide la influencia de un nodo en la red social, y le asigna al mismo un puntaje relativo no sólo en base a la importancia del nodo en particular sino también en base a qué tan bien conectados están sus nodos adyacentes.

La idea de esta métrica es que la importancia de un actor está recursivamente definida por sus relaciones, o de otra manera, que la centralidad de un actor depende de la sumatoria de centralidades de sus conexiones.

La centralidad de autovector C_{EV} del nodo v es proporcional a la suma de centralidades de sus vecinos. Para el nodo v , donde $N(v)$ es un conjunto de vecinos de v , y la constante λ es el mayor autovalor de la matriz de adyacencia, la centralidad de autovector es entonces definido como:

$$C_{EV}(v) = \frac{1}{\lambda} \sum_{u \in N(v)} C_{EV}(u)$$

4.4.8 Cierre Triádico (Triadic Closure):

Un concepto sugerido inicialmente en 1908 por el sociólogo Georg Simmel [8], en una red social dada es la tendencia a ser un contacto entre aquellas personas que comparten contactos en común.

Se puede ejemplificar entre tres nodos A, B, y C, de manera que si existe un vínculo entre AB y AC, muy probablemente también lo exista entre BC.

La cantidad de cierres triádicos, o relaciones en forma de triángulos, que hay en una red puede ser usado para encontrar comunidades de individuos que se conocen bien entre sí.

A modo de ejemplo, podemos definir un triángulo entre tres nodos i,j,k como un conjunto representado por tres aristas:

$$\{(i,j), (j,k), (i,k)\}.$$

A continuación se verá otro índice que usa el concepto de cierre triádico.

4.4.9 Coeficiente de clustering local:

El mismo mide la proporción de relaciones triádicas de un nodo sobre el número posible de relaciones triádicas para el mismo nodo.

Las redes sociales son usualmente transitivas, lo que significa que los amigos de un actor dado tienden a ser amigos entre sí también.

Esta propiedad de transitividad es medida por el coeficiente de clustering, el cual puede ser a nivel global de la red o para cada uno de los nodos. En la próxima sección se ampliarán los conceptos de cierre triádico, y coeficiente de clustering a nivel global.

A modo de ejemplo, a continuación (figura 4.5) se observan en un grafo algunas de las medidas descritas anteriormente.

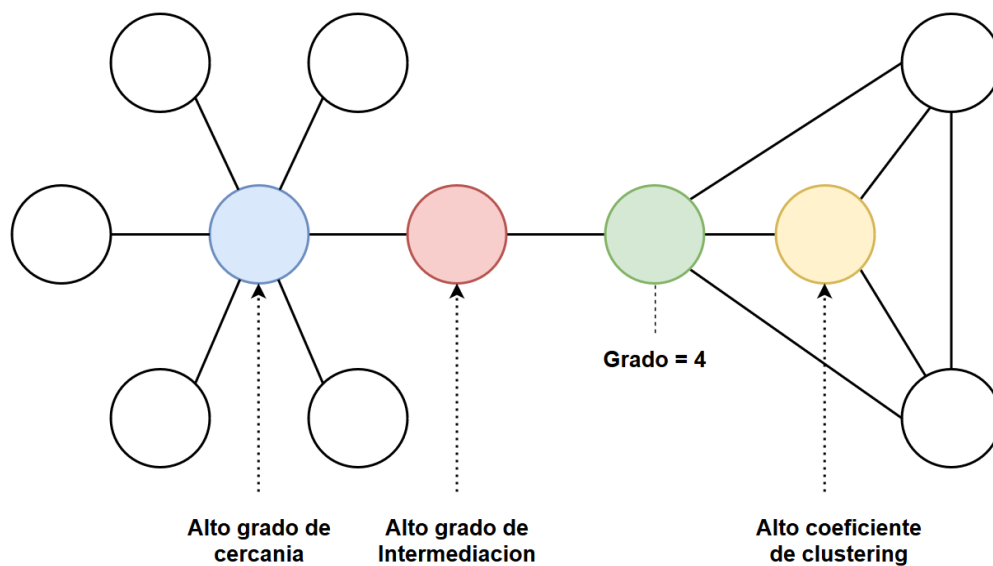


Figura 4.5: Representación de medidas de cercanía

4.5 Medidas de estructura de red

Las redes sociales pueden ser densas, fragmentadas, descentralizadas y/o extensas, entre otras características de su estructura. En SNA, una serie de medidas nos ayudan a entender la topología de la red, su forma y propiedades básicas.

A continuación se enumeran algunas de estas medidas utilizadas en este trabajo.

4.5.1 Componentes

Si un grafo no está conectado, significa que se separa en distintos conjuntos de grupos de nodos conectados entre sí y desconectados de otros grupos. A esos grupos conectados entre sí se les denomina componentes.

Un componente conectado, es aquel subgrupo de nodos donde:

- Todos los nodos del subgrupo tienen un camino (considerando aristas no dirigidas) a cualquier otro nodo del subgrupo
- Ningún otro nodo fuera del subgrupo tiene un camino a algún nodo del subgrupo.

4.5.2 Distribución de grado (*degree distribution*)

Así como el grado de un nodo es el número de aristas incidentes al mismo, la distribución de grado es la probabilidad de ocurrencia de estos grados a lo largo de

toda la red.

Si hay n nodos en total en la red y n_k de ellos tienen un grado k , entonces tenemos que la distribución de grado es:

$$P(k) = n_k/n.$$

Estudiar la distribución de grado es importante ya que la mayoría de las redes sociales tienen nodos con un grado bajo, pero una cantidad pequeña de nodos tienen un grado alto y están conectados a muchos otros nodos, siguiendo la ley de potencia. Estas redes son llamadas de escala libre y serán vistas en otro capítulo.

4.5.3 Distancia Promedio (Average Distance)

Es la distancia promedio de los caminos más cortos entre cada par de nodos de la red social. Esta medida frecuentemente es usada complementariamente al diámetro de la red social.

4.5.4 Densidad

La densidad de un grafo cuantifica el número de conexiones entre los actores de una red.

Un grafo es considerado denso si el número de aristas alcanza un número cercano al máximo que puede alcanzar en un grafo determinado. La densidad de un grafo no dirigido puede ser expresada como:

$$\rho(G) = \frac{m(G)}{m_{\max}(G)}.$$

Donde m es la cantidad de nodos en la red, y $m_{\max}(G)$ denota la cantidad máxima de nodos posibles. El valor puede ir en un rango entre 0 y 1.

4.5.5 Coeficiente de afinidad selectiva (Assortativity)

El coeficiente de afinidad selectiva es el coeficiente de correlación de Pearson del grado entre pares de nodos conectados y representa en qué intensidad los nodos de la red se asocian con otros parecidos u opuestos.

Por ejemplo, en las redes sociales en línea tienden a ser afines los nodos de alto grado. Esta medida tiene valores desde +1 (afín), 0 (neutral), a -1 (no afín).

La afinidad selectiva entre los nodos j y k puede ser definida como:

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{\sigma_q^2}.$$

Donde el término q_k es la distribución del grado remanente, y e_{jk} es la distribución de probabilidad conjunta del grado remanente de ambos vértices.

4.5.6 Coeficiente de clustering global y transitividad (transitivity)

El coeficiente de clustering de Watts-Strogatz y la transitividad de la red son métricas usadas comúnmente para caracterizar la transitividad en redes del mundo real [25].

La transitividad es una propiedad que mide la relación entre la cantidad de triángulos existentes en la red respecto a todos los posibles triángulos, o la tendencia de los nodos a formar relaciones triádicas.

$$C = \frac{\text{cantidad de triángulos cerrados}}{\text{cantidad de triángulos totales (abiertos y cerrados)}}$$

Otra manera de medir si existen muchas relaciones triádicas en la red social, o la prevalencia de cierres triádicos en una red estática como la analizada en este trabajo, es usando un coeficiente de clustering global.

Si bien hay muchas maneras de calcular este coeficiente, se adopta en este trabajo la propuesta por Watts y Strogatz, en la cual se obtiene el coeficiente para toda la red a través del cálculo del promedio de todos los coeficientes de clustering local (definido anteriormente).

$$\bar{C} = \frac{1}{n} \sum_{i=1}^n C_i.$$

Las redes sociales de mundo pequeño se caracterizan por coeficientes de clustering global alto, debido a que la propiedad de transitividad entre nodos es frecuente [26]. Se analizará más en detalle en otro capítulo.

Ambas medidas, así como la densidad, se expresan con un valor entre 0 y 1 para indicar cuan interconectado está un grafo, con un valor bajo indicando baja transitividad y un valor alto indicando gran proporción de relaciones triádicas cerradas. Estos indicadores, nos permiten entender el grado de cohesión de la red y

nos permiten también pensar en el potencial de relaciones que podría tener un grafo respecto al valor máximo.

Cabe aclarar que si bien la transitividad y el coeficiente de clustering global, ambos miden las relaciones triádicas, el primero le da más importancia relativa a los nodos de alta centralidad de grado. Es esperable encontrar divergencia entre ambas medidas según los tipos de grafos que se estén midiendo [27].

4.5.7 Promedio de caminos más cortos (average shortest path length)

El camino más corto (shortest path) es utilizado a nivel individual en la métrica de grado de intermediación, pero también se puede calcular a nivel global para la red social.

El promedio de todos los caminos más cortos posibles entre todos los pares de nodos de la red social nos da una idea complementaria a la de diámetro y radio de una red.

4.5.8 Diámetro y Radio

El diámetro está dado por el camino más largo de todos los caminos más cortos calculados entre todos los pares de nodos de la red, o bien la máxima excentricidad. Esta medida nos da una idea aproximada de la distancia entre un extremo y el otro de la red social. El radio, por el contrario, es la mínima excentricidad de la red.

$$D = \max\{e_v : v \in V\},$$

$$R = \min\{e_v : v \in V\}.$$

4.5.9 Modularidad (Modularity) y heurística de Louvain para la detección de comunidades

La modularidad es una medida similar a la densidad, pero utilizada para identificar qué tan fraccionado está el grafo. Los módulos, o clases de módulos, tienen una alta densidad entre los nodos que lo componen, pero una densidad baja por fuera de los nodos que integran el componente o módulo.

La modularidad va en un valor entre -1 (no modular) a +1 (totalmente modular). Es comúnmente aceptado que un valor mayor a 0,3 indica un grado significativo de

estructura de comunidades.

La modularidad es definida como:

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \gamma \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

Donde m es el número de aristas, A es la matriz de adyacencia del grafo, k_i es el grado de i , γ es el parámetro de resolución, y $\delta(c_i, c_j)$ es 1 si i y j están en la misma comunidad, sino es 0.

La modularidad es habitualmente utilizada en métodos de optimización para detectar estructura de comunidades en una red. El procedimiento de computar las particiones óptimas de nodos que maximizan la modularidad es NP-difícil por lo que se implementan heurísticas.

Uno de los módulos²⁶ para la detección de comunidades para NetworkX utiliza la heurística de Louvain [28], la cual optimiza la modularidad, y es utilizada en esta tesis.

²⁶ Community detection for NetworkX (<https://python-louvain.readthedocs.io>)

Resultados y discusión

5 Análisis de usuarios, comentarios y lugares

Una vez que fueron excluidos los usuarios con valores extremos, explicado en el capítulo de procesamiento de datos, se describen a continuación las principales características de las 31385 usuarias consideradas en el análisis, así como sus amigas.

5.1 Descripción de usuarios y sus amigos

El 60,99% de los usuarios analizados en el grafo principal se declaran de género masculino en el perfil de la plataforma, mientras que el 35,9% es femenino y 3,01% no es declarado. Estas proporciones se mantienen muy similares cuando consideramos el total de usuarios y también sus amigos directos²⁷ que componen el grafo extendido (GE). Estos valores son muy similares a diversos estudios [29][30]. Según afirmaciones de Foursquare, en el año 2011, el 40% de sus usuarios eran mujeres, por lo que las características de las personas que realizan comentarios en CABA se adecuan a estas características a nivel global.

A continuación podemos observar la cantidad de amigos de CABA y amigos totales, para los usuarios considerados en el análisis.

Medida	Cantidad de amigos en CABA (GO)	Cantidad de amigos en total (GE)	Porcentaje de amigos en CABA
Media	14,60	79,10	27,74%
Desvío Estándar	33,12	115,27	23,06%
Mínimo	1,00	1,00	0,10%
1er cuartil	3,00	17,00	7,69%
Mediana	6,00	40,00	24,14%
3er cuartil	14,00	100,00	41,18%
Valor Máximo	509,00	1.221,00	100,00%

Tabla 5.1: Medidas de tendencia central para métricas de usuarios analizados

²⁷ Los usuarios y amigos que componen el grafo extendido, se caracteriza en un 36,86% de mujeres, 60,64% hombres y 2,49% no declarado.

El grado medio de amigos por usuarios es de 14,6, considerando solo los amigos en GO (grafo de comentadores y amigos también comentadores en CABA), y el mismo es de 79,10 si se consideran todos los amigos directos y que constituyen GE. A pesar de que el grado es relativamente bajo para GO, se trata de una red social de mundo pequeño que es analizado en la sección correspondiente.

Si bien se ha encontrado que solo entre el 27,74% (media) de los amigos de los usuarios de GO también son integrantes de la red de comentadores de CABA, estas proporciones varían de manera considerablemente al subdividirse por comentadores hispanos y no hispanos, teniendo una media de 30,8% para los primeros y 16,79% para los segundos. Así mismo la media de amigos totales es mayor para los no hispanos (117,23) frente a los hispanos (70,23).

No se encuentran diferencias significativas por género al interior de cada grupo idiomático.

Podemos observar a continuación la distribución porcentual de amigos de CABA sobre amigos totales para no hispanos (azul) e hispanos (rojo).

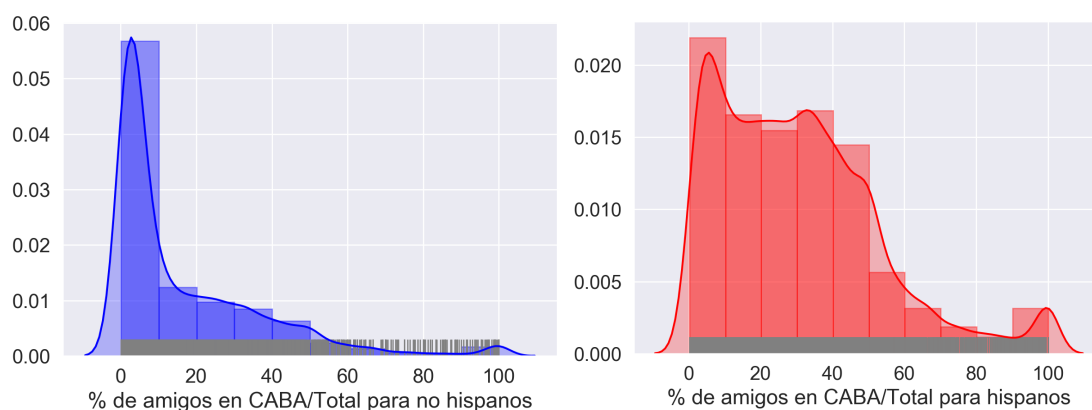


Figura 5.1: Histograma y estimación de densidad Kernel para proporción de amigos en CABA sobre el total de amigos. Usuarios no hispanos en azul y usuarios hispanos en rojo

5.2 Comentarios

Por las limitaciones de la API de la plataforma, no es posible estimar qué porcentaje de usuarias que hayan registrado check-in en lugares de CABA también dejaron comentarios, ya que solamente era posible identificar las usuarias que dejaron comentarios públicos.

En estudios globales se caracteriza a los usuarios de Foursquare como principalmente consumidores de comentarios, con solo 0,89 tips por persona, y en

donde sólo un 17% de los usuarios dejan al menos un comentario, y un porcentaje muy chico deja la mayoría de estos [30].

En este análisis la cantidad de comentarios totales considerados es de 103.410, de los cuales podemos ver a continuación (tabla 5.2) la distribución junto con la cantidad de usuarios comentadores únicos por año.

Año	cantidad de comentarios	cantidad de usuarios distintos
2009	15	9
2010	5283	2323
2011	17414	7290
2012	23198	10108
2013	32548	12863
2014	24109	10525
2015	843	573

Tabla 5.2: Cantidad de comentarios y usuarios únicos, agrupados anualmente

La gran parte de los comentarios recolectados fueron realizados entre el 2011 y 2014, que corresponde con el crecimiento de Foursquare a nivel global. Este patrón se asemeja con otros estudios globales en donde el pico de comentarios se da en 2013 y luego presenta un descenso [30].

Aproximadamente el 40% de las usuarias consideradas dejan comentarios en más de un año distinto, siendo el promedio de 1,39 años distintos por usuaria.

De los 31385 usuarios que han dejado comentarios en CABA, 16545 de ellos, el 52%, lo han hecho en más de un lugar distinto.

En el total de las usuarias analizadas, estas dejan comentarios en 1,83 comunas distintas, mientras que si consideramos a las usuarias que tienen más de un comentario, este promedio asciende a 2,58 comunas.

En cuanto a cantidad de lugares distintos en los que dejan comentarios, en el total de usuarios analizados, han dejado en promedio comentarios en 3,05 lugares distintos, mientras que este valor asciende a 5,05 si consideramos usuarios con más de un comentario.

Estos valores se acercan a la cantidad de comentarios en promedio que han tenido los usuarios, el cual es 3,29. En la figura 5.2 podemos observar la similitud entre ambas distribuciones, las de comentarios únicos por lugar y la de comentarios totales.

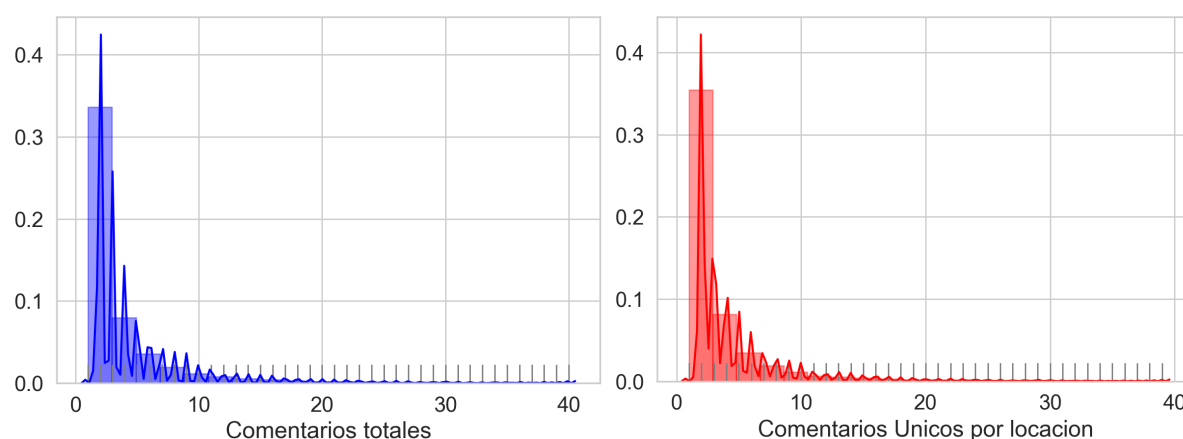


Figura 5.2: Histogramas de distribución de comentarios totales (azul) y comentarios únicos por lugar (rojo)

A continuación se encuentran las medidas de tendencia central para comentarios y check-ins, que describen la utilización de la plataforma tanto en CABA como fuera de ella por parte de los comentaradores de CABA.

Medida	Comentarios en CABA	Comentarios Totales	Check-in Totales
Media	3,29	11,55	650,13
Desvío Estándar	4,48	20,34	1385,37
Mínimo	1	1	1
1er cuartil	1	2	47
Mediana	2	5	208
3er cuartil	3	12	670
Valor Máximo	40	234	38109

Tabla 5.3: Medidas de tendencia central para comentarios y check-ins para comentaristas de lugares en CABA.

Los comentarios en CABA representan entre un 28,4% y 40% de los comentarios totales de las usuarias, según tomemos la media o la mediana para comparar los mismos²⁸. Esto nos indica que los usuarios analizados no se desplazan solamente

²⁸ Cabe aclarar que la cantidad de comentarios en CABA representa un cálculo exacto sobre los comentarios recogidos de los lugares, mientras que la cantidad de comentarios totales, check-ins, fotos y listas responden al valor indicado en el perfil del usuario al momento en que fue recolectado, por lo que es esperable cierta distorsión de estos valores.

por el espacio de CABA, ya que también dejan comentarios por fuera de la misma. A continuación (figura 5.3) podemos observar la distribución porcentual de comentarios en CABA sobre el total de comentarios realizados por cada usuario.

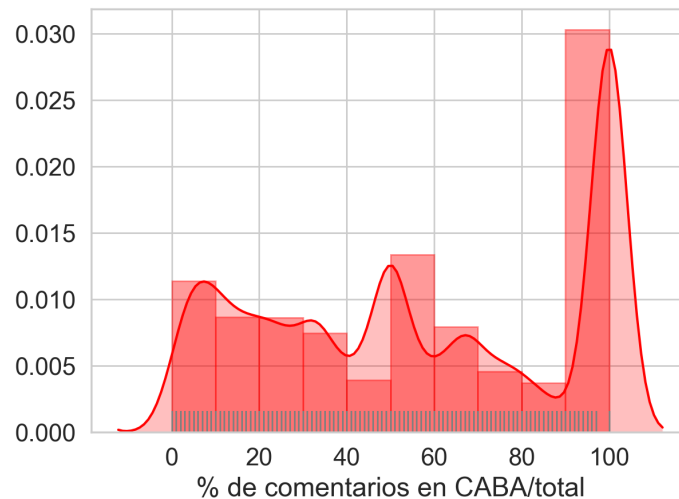


Figura 5.3: Histograma y estimación de densidad Kernel para comentarios en CABA sobre comentarios totales para cada usuaria

Cuando analizamos esta proporción entre comentarios de CABA sobre comentarios totales según se considere usuarios hispanos o no hispanos, podemos observar que para los primeros la mediana representa el 40% de los comentarios totales, mientras que para los no hispanos esta proporción desciende a tan solo el 16%. Esto nos indica que los usuarios no clasificados como hispanos, tienden a tener menor proporción de sus comentarios en CABA que aquellos usuarios tipificados como hispanos, y puede deberse al presupuesto de que se tratan de turistas. Esto será de particular importancia a la hora de estimar las distancias entre amigos y no amigos en el capítulo 7 y será analizado en más detalle.

En otro aspecto, también podemos observar que los comentarios totales de las usuarias representan sólo entre el 1,78% y 2,40% de los check-ins que realizaron. A continuación (figura 5.4) se observa la distribución de comentarios totales y check-ins según categoría de idioma, donde se puede observar que las personas no hispanas tienen una mayor cantidad tanto de comentarios totales como check-ins.

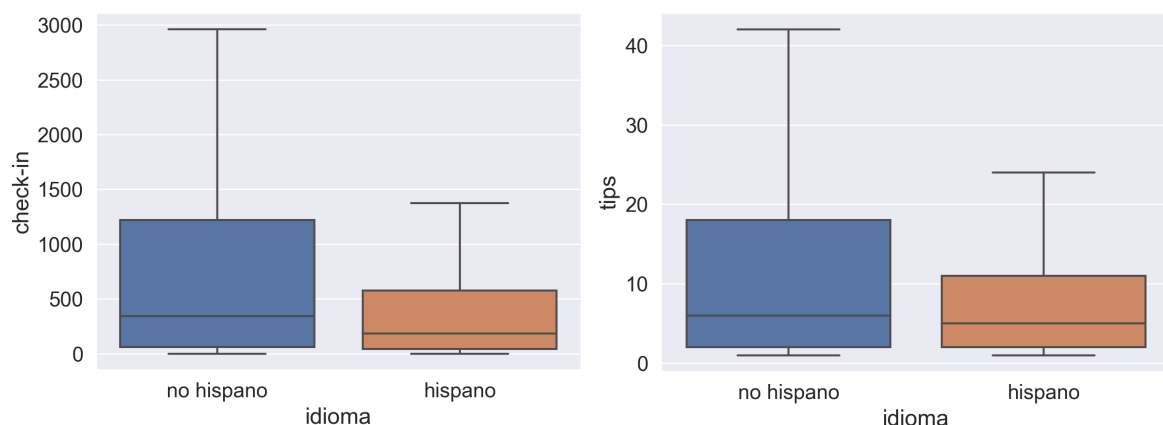


Figura 5.4: Box-plots para check-in (izquierda) y cantidad de comentarios totales (derecha), según idioma. Personas hispanas en naranja, no hispanas en azul.

Podemos observar que la proporción de check-ins por usuaria es mucho mayor a lo de los comentarios, lo que nos habilita a pensar que las usuarias parecen responder a la propuesta principal del producto: la búsqueda de lugares en sus cercanías, lectura de comentarios de los lugares y compartir los lugares que visitan con sus amigas directas a través de check-ins (o a través de otras redes sociales).

En menor medida los usuarios parecen participar activamente en el contenido de la red social, deducible por las métricas de fotos, comentarios y listas, que se analizará en la próxima sección.

En cuanto a la distribución según género, los hombres, los cuales representan el 61,0% del total de usuarios considerados, tienen una mayor proporción en cuanto al total de comentarios en CABA (60,9%), check-ins (69,4%) y comentarios totales (63,4%). En ambas métricas evaluadas en la figura 5.5 podemos observar que la mediana es mayor para hombres que para las mujeres.

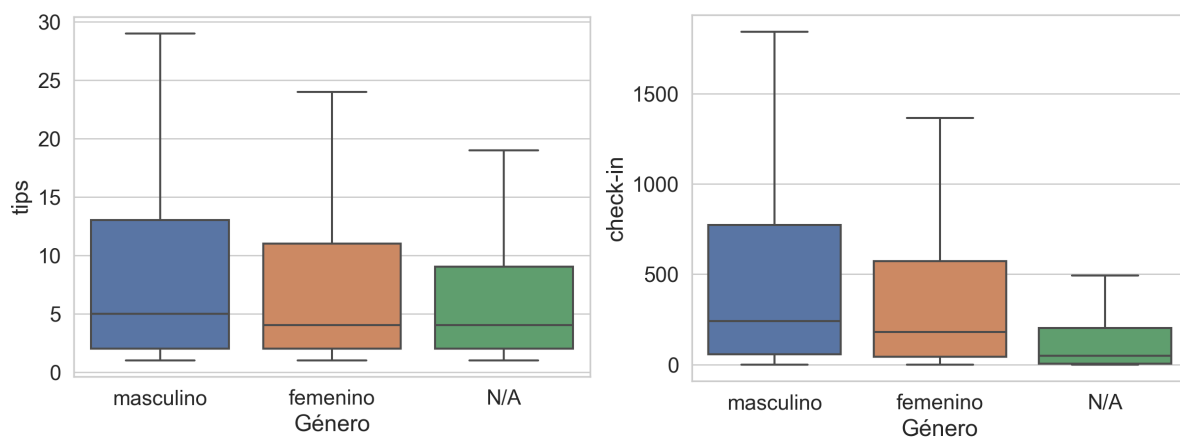


Figura 5.5: Box-plots para comentarios totales (izquierda) y check-ins (derecha), según género. Género masculino en azul, femenino en naranja y no declarado (N/A) en verde.

Sin embargo, podemos encontrar que esta relación según género se invierte si solamente consideramos los comentarios realizados en CABA y analizamos las cantidades por persona. Si bien el 60,9% de los comentarios totales de CABA fueron realizados por hombres, las mujeres publican mayor cantidad de comentarios por persona, lo que va en consonancia con otros estudios [30]. Observamos en la figura 5.6 que el 3er cuartil y máximos de la cantidad de comentarios en CABA es mayor para las mujeres.

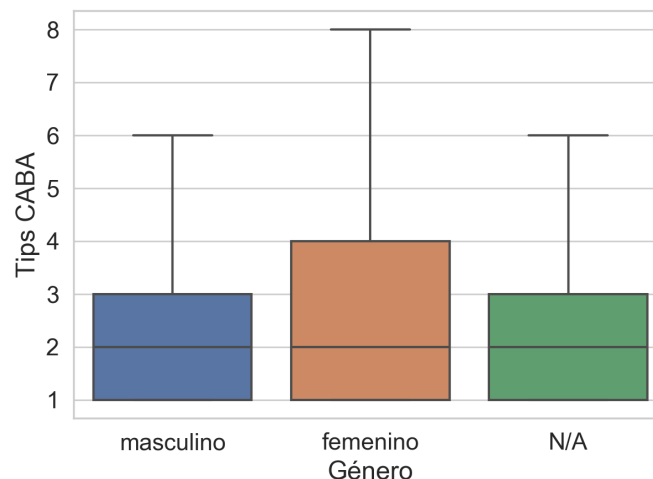


Figura 5.6: Box-plot de cantidad de comentarios en CABA según género. Género masculino en azul, femenino en naranja y no declarado (N/A) en verde.

A su vez podemos encontrar que la proporción de comentarios en CABA es mayor solamente para las usuarias del género femenino categorizadas como hispanas.

5.2.1 Comentarios según día y hora

La distribución de los comentarios varía según la hora²⁹, día de la semana, tipo de categoría o la comuna del lugar considerado. Por ejemplo, aquellos lugares de entretenimiento (tales como bares, restaurantes, etc) presentan comentarios en horarios más nocturnos que lugares comerciales.

Se encontró que la mayoría de los comentarios son hechos durante los días de la semana a lo largo de todas las horas, excepto en la franja horaria de las 3 a 5 AM, ya que en esta franja horaria son realizados preponderantemente durante el fin de semana.

En la figura 5.7 podemos observar que sábados y domingos son los días con mayor cantidad de comentarios si se consideran los días individualmente.

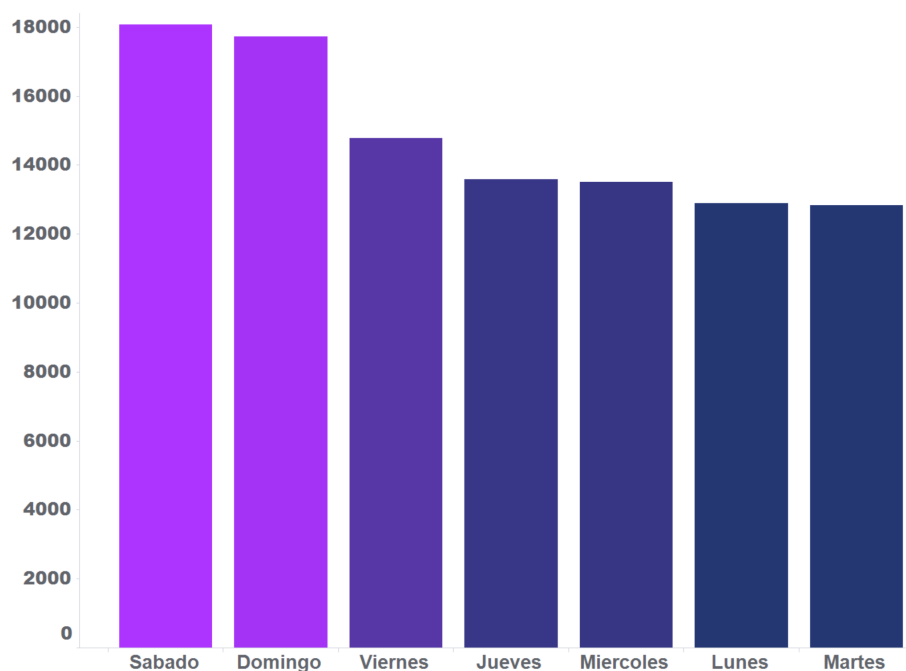


Figura 5.7: Distribución de comentarios en orden descendente según día de la semana (eje x)

En la serie de tiempo de la figura 5.8 podemos observar que, si bien los días de semana y fin de semana tiene un perfil similar de horarios, el fin de semana se

²⁹ Los horarios de los comentarios recolectados se encontraban en epoch, formato UNIX para describir un momento en el tiempo, por lo que fueron convertidos a la zona horaria UTC -3, horario regular en Argentina.

extiende los horarios de trasnoche. Así mismo se encuentra que la mayor cantidad de comentarios son entre el mediodía y la medianoche.

En el eje X se observan las horas del día mientras que en el eje Y en azul los días de semanas, y en rojo los días de fin de semana (sábado y domingo).

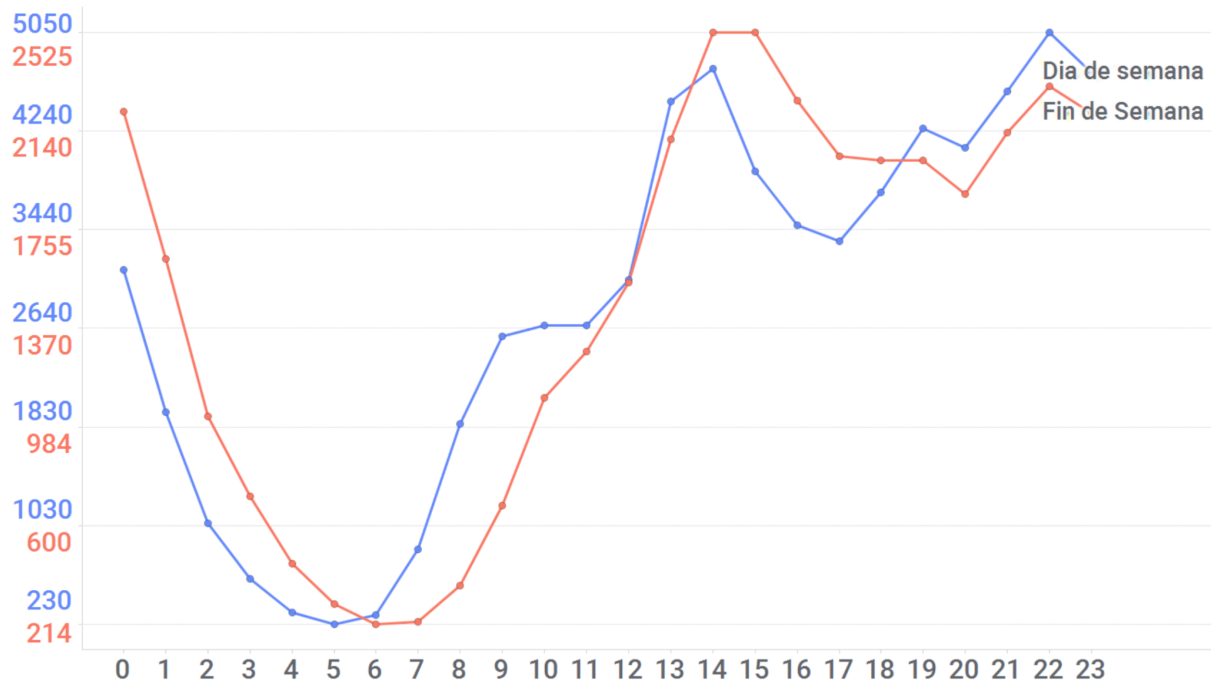


Figura 5.8: Distribución de comentarios por hora (eje x) agrupados según día de la semana (serie azul) o fin de semana (serie roja)

Complementariamente, en la figura 5.9, podemos observar la distribución porcentual entre días de semana y fin de semana para cada una de las horas del día, donde se puede observar el patrón nocturno de los fines de semana.

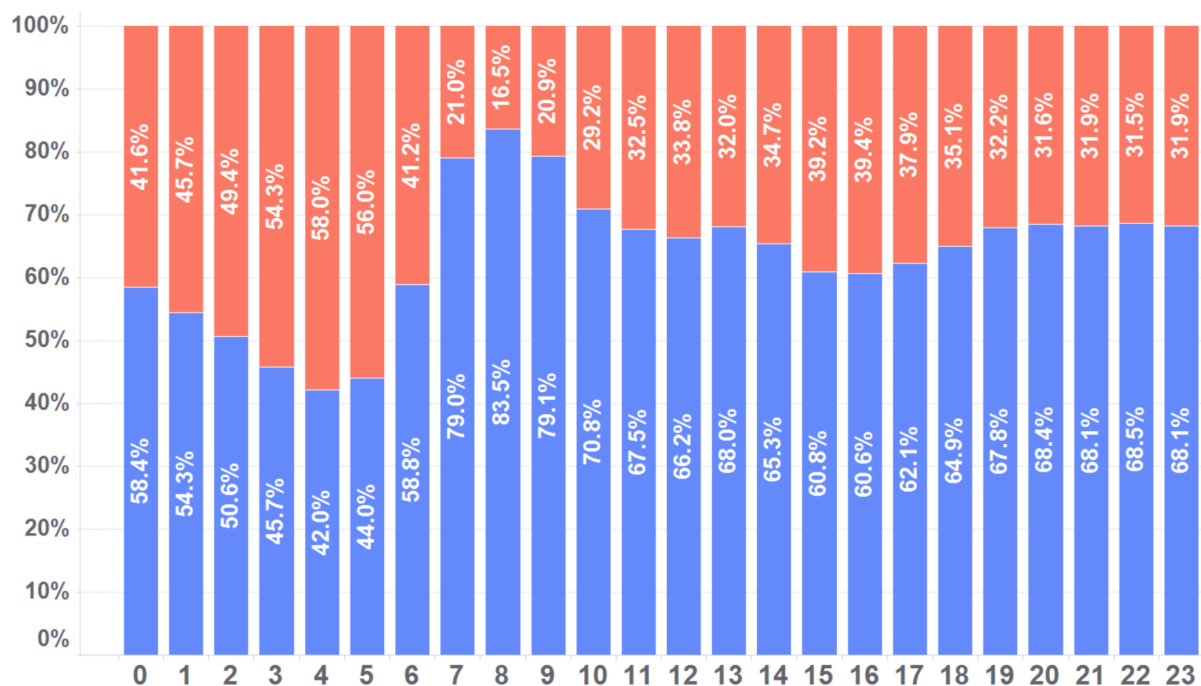


Figura 5.9: Distribución porcentual de comentarios por hora (eje x) agrupados según día de la semana (azul) o fin de semana (rojo)

5.3 Indicadores de uso de la plataforma

Además de amigas, comentarios y check-ins, se obtuvieron otras métricas o atributos de las usuarias³⁰ que, si bien no están restringidas o identificadas necesariamente a su actividad dentro del territorio de la CABA, nos ayudan a entender mejor el uso de la plataforma en su totalidad.

Entre esas métricas podemos encontrar las insignias (*“badges”*), las cuales son otorgadas como reconocimiento por las actividades en la plataforma, mientras que las alcaldías (*“mayorship”*) son obtenidas al hacer check in más que cualquier otra usuaria en los últimos 30 días. Las listas contienen colecciones de lugares, creadas por la misma usuaria o representa una lista seguida por esta, mientras que cuentan como fotos a aquellas subidas públicamente a los venues.

En promedio, los usuarios analizados han subido 32,3 fotos, obtenido 15,36 insignias, han creado o siguen 1,54 listas de lugares y han obtenido 0,23 veces la categoría de alcaldía en algún lugar en toda su historia.

En la tabla 5.4 podemos observar la distribución de tales medidas para usuarias del conjunto GO.

³⁰ Si bien la API de Foursquare fue limitándose en el tiempo debido a la separación del producto mencionada, en la actualidad (Julio del 2020) estas métricas siguen siendo accesibles.

Medida	Insignias	Alcaldías	Fotos	Listas
Media	15,36	0,23	32,3	1,54
Desvío Estándar	15,62	2,76	100,37	9,44
Mínimo	0	0	0	0
1er cuartil	4	0	1	0
Mediana	12	0	7	0
3er cuartil	22	0	24	1
Valor Máximo	436	199	3062	1225

Tabla 5.4: Medidas de tendencia central para indicadores de actividades

También en este caso se encuentran diferencias entre la cantidad de fotos y listas para comentadores hispanos y no hispanos. Esto podría indicar que la intensidad de uso de la plataforma, y/o antigüedad en la misma, es mayor para personas presuntamente no habitantes de CABA.

En tanto las personas de habla hispana subieron en promedio 26,91 fotos por usuario, las personas no hispanas subieron 51,63 fotos. En esta métrica se observa una diferencia significativa ($p < 0,01$) entre usuarios de género masculino (37,36 fotos) y género femenino (24,97 fotos). Esta diferencia es 38,59% mayor en hombres hispanos respecto a mujeres hispanas, pero la brecha entre géneros asciende al 71,66% entre personas no hispanas.

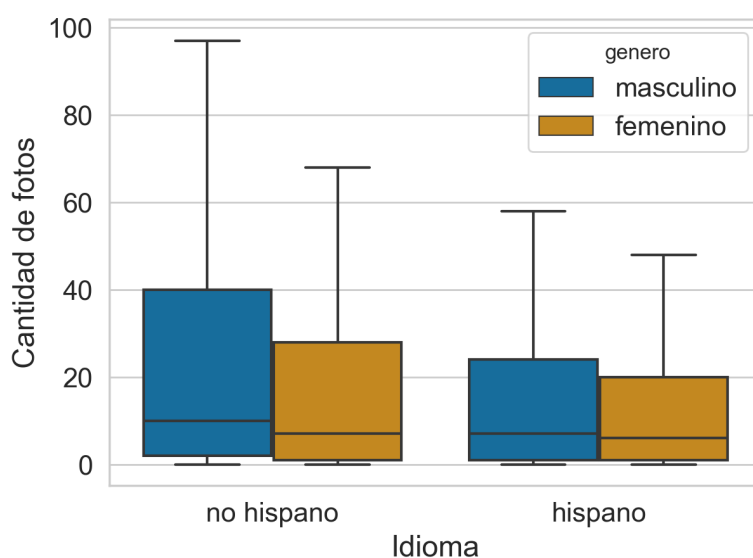


Figura 5.10: Cantidad de fotos según género y grupo idiomático.

Cabe resaltar que ninguna de estas métricas se restringe a CABA, por lo que parece ser una relación también extensible fuera del territorio analizado.

5.4 Lugares (venues)

La cantidad de lugares recolectados para la CABA son 69504, donde solamente el 30,5% de ellos tienen comentarios registrados, y solo el 26,9% corresponden a lugares con comentarios de usuarios considerados en el análisis, lo que restringe el universo analizable a un total de 18709³¹ lugares.

En la figura 5.11 podemos observar la distribución de comunas en CABA, mientras que en la figura 5.12 podemos observar la distribución geográfica de estos venues en las quince comunas de CABA³².



Figura 5.11: Nombres y números de las comunas de CABA y barrios dentro de cada comuna

³¹ 145 de estos venues se encuentran por fuera de CABA. En el proceso de construcción de la base de datos, se obtuvieron los 50 venues más próximos a cada una de las manzanas de CABA, encontrándose algunos de estos venues en las inmediaciones de Capital Federal. Se dejaron los mismos para el análisis bajo la categoría de “Gran Buenos Aires”.

³² Debido a la poca fiabilidad de la ubicación descritos en los perfiles de los lugares en Foursquare, los lugares tuvieron que ser clasificados a las comunas que pertenecen en base a su latitud y longitud en el mapa. El mapa de comunas utilizado para catalogar a las mismas corresponde al publicado el 28 de octubre del 2016 por la Dirección General de Estadística y Censos bajo la etiqueta “división-territorial infraestructura USIG”.

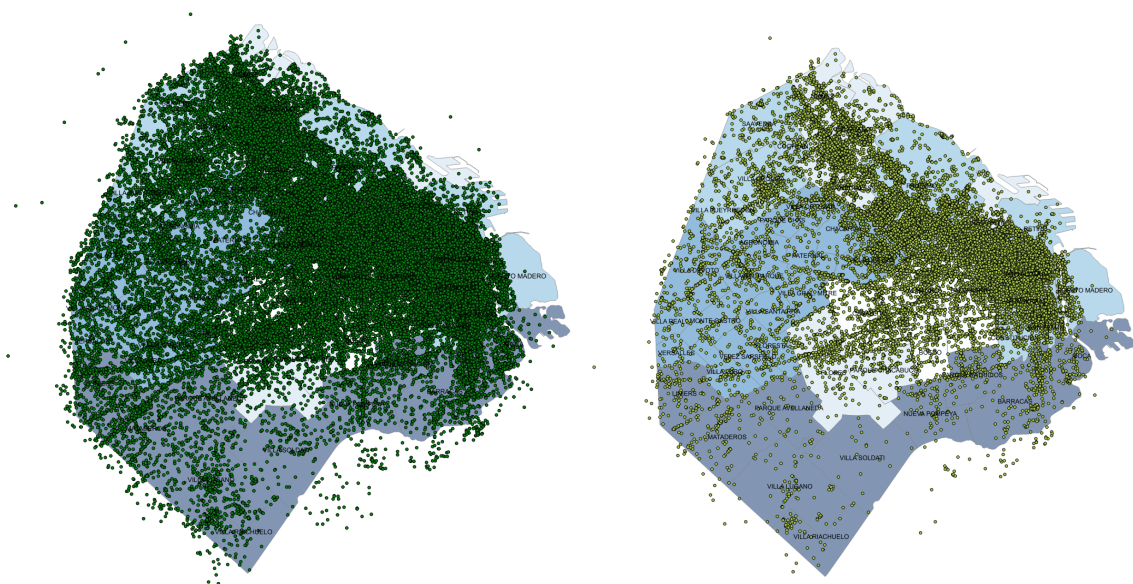


Figura 5.12: Lugares de CABA recolectados inicialmente (izquierda) y distribución de los lugares considerados en el análisis (derecha) sobre el mapa de CABA.

Los puntos oscuros en la figura 5.12 representan lugares recolectados inicialmente pero que no todos tienen comentarios, mientras que los puntos amarillos son lugares que tienen comentarios y analizados en esta tesis. Como podemos ver, ambas distribuciones se asemejan espacialmente.

En la misma figura podemos observar a las comunas de CABA segmentadas en 5 escalas de colores. Los mismos representan el nivel de densidad poblacional de las 15 comunas que conforman CABA, en donde el tono más claro representa las comunas de mayor nivel de densidad poblacional³³.

Se observa que los lugares analizados se distribuyen en mayor medida en aquellos espacios con alta densidad poblacional. También es válido mencionar que se observa que esta distribución de lugares tiene su correlato en la distribución de la riqueza y desigualdades sociales a nivel de la ciudad [33], en donde las zonas con menor nivel socioeconómico también poseen menor cantidad de lugares registrados en Foursquare.

En la figura 5.13 podemos observar la distribución de comentarios según las comunas. En rojo se observa los comentarios realizados en fin de semana mientras que en azul durante el día de la semana (lunes a viernes).

³³ Para la construcción del ranking de densidad poblacional por comuna se utilizó "Población total estimada por sexo, superficie y densidad poblacional según comuna. Ciudad de Buenos Aires. Años 2006/2018" de Dirección General de Estadística y Censos (Ministerio de Hacienda y Finanzas GCBA). Mientras que para el nivel socioeconómico de cada comuna se armó una jerarquía en base a dos fuentes conjuntamente. "Promedio de ingreso total familiar (ITF) según comuna. Ciudad de Buenos Aires. Año 2012." y "Precio promedio del m2 (dólares) de departamentos en venta de 2 ambientes usados por comuna. Ciudad de Buenos Aires. 1er. trimestre 2015/1er. trimestre 2020" de la misma fuente.

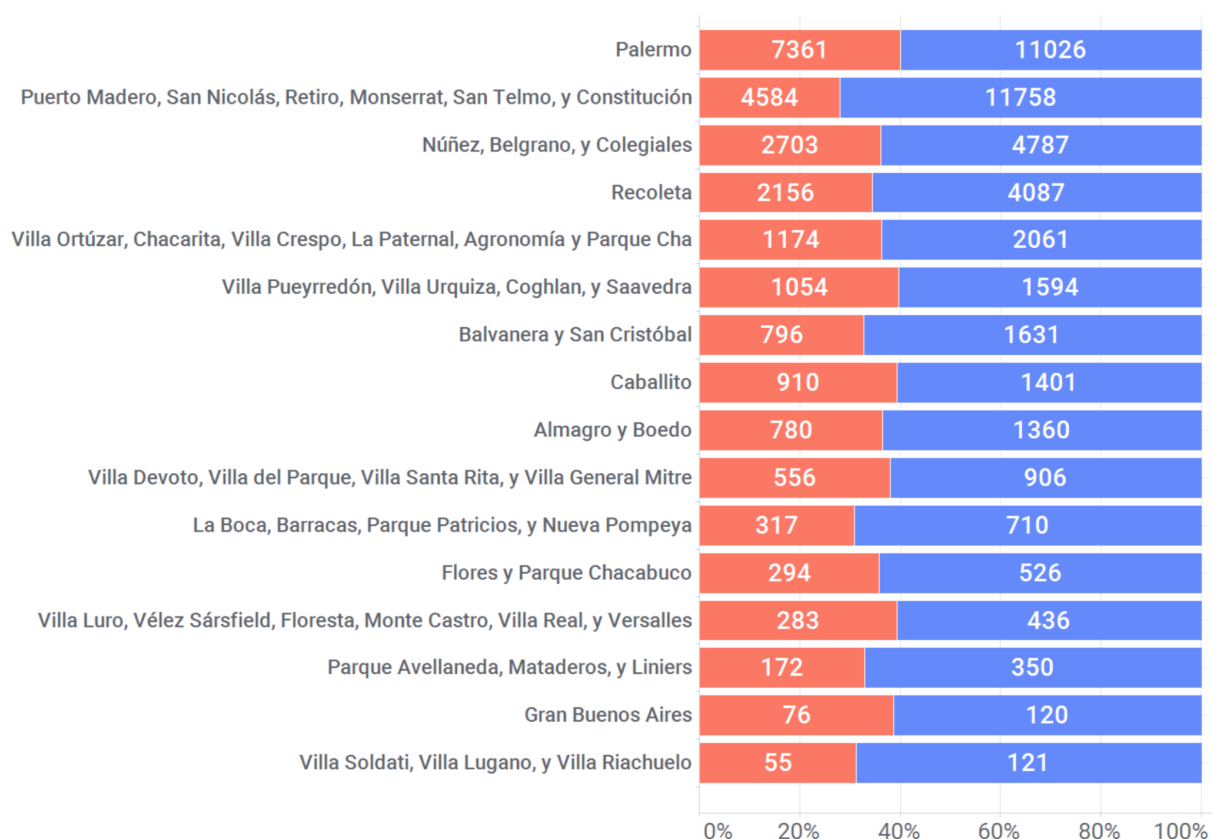


Figura 5.13: Cantidad y porcentaje de comentarios por comuna durante fin de semana (rojo) y día de semana (azul)

Si bien no hay estudios similares de LSBN en CABA para comparar los resultados, existen análisis similares llevados a cabo sobre 4,95 millones de usuarios de teléfonos móviles [34]. Los mismos coinciden en ubicar a la comuna de Palermo como la más importante los fines de semana y Comuna 1 (puerto madero, etc) como la más transitada durante días laborables.

Los lugares considerados tienen un promedio de 347,8 check-ins (es decir, la cantidad de veces que usuarios registraron estar presente en un lugar determinado), mientras que si consideramos la cantidad promedio de usuarios distintos que han registrado check-in, este valor es de 138,81 por lugar y el promedio de comentarios declarados por lugar es de 7,04³⁴.

En la tabla 5.5 podemos observar las medidas de tendencia central para estas

³⁴ Se entiende por “comentarios declarados” a aquellos datos declarados en el objeto de *venues* en la plataforma, mientras que a “comentarios utilizados” a aquellos comentarios que finalmente se consideran dentro del *venue*. Esta diferencia se debe principalmente a excluir los comentarios de los usuarios *outliers*, y en menor medida a las distorsiones producidas durante el proceso de recolección de datos, explicados al inicio de la tesis.

métricas.

Medida	Cantidad de check-ins	Cantidad Usuarios Únicos	Comentarios declarados	Comentarios utilizados
Media	347,80	138,81	7,04	5,52
Desv. Estándar	1.437,88	578,50	18,17	13,28
Mínimo	0	0	0	1
1er cuartil	17	6	1	1
Mediana	65	21	2	2
3er cuartil	244	87	5	4
Valor Máximo	93748	34092	751	241

Tabla 5.5: Medidas de tendencia central para check-ins, usuarios únicos, comentarios recolectados y comentarios utilizados en el análisis de los lugares de CABA.

La cantidad de comentarios que registran los lugares difiere en una proporción que va de 30 a 60 veces (según se utilice la mediana o media) respecto de la cantidad de check-ins, un orden de magnitud similar a la encontrada entre las diferencias de comentarios y check-ins a nivel usuario.

En cuanto a los tipos de lugares, la mayor categoría está relacionada a locales comerciales (pubs, restaurantes, etc) y en mucho menor medida se pueden encontrar lugares “privados” (viviendas particulares, compañías, etc) o lugares sin categoría declarada. El listado de 441 categorías encontradas excede un cuadro resumen, pero a continuación podemos ver un ejemplo de los principales tipos de lugares y la cantidad de comentarios asociados a esos lugares.

Categoría	cantidad de lugares	cantidad de comentarios
Restaurantes Argentinos	969	9289
Oficinas	813	1559
Cafes	681	5319
Pizzeria	626	5040
Restaurantes	430	2926

Tabla 5.6: Categorías, comentarios y cantidades de lugares

Se encontró que la cantidad de lugares por categoría está fuertemente correlacionada linealmente (r de Pearson=0,82) con la cantidad de comentarios. Así mismo se observa una preponderancia de check-ins en categorías de lugares comerciales o turísticos (pubs, bares, restaurantes, etc). Esta distribución de lugares fuertemente orientada a la categoría de “comida” es una característica presente en otros estudios [30][35].

Debido a la gran cantidad de categorías, se procedió a clasificar manualmente una muestra del 68% de las mismas y agruparlas en cuatro tipos³⁵: residencial, laboral, entretenimiento y comercial.

Se comprobó que los lugares no están uniformemente distribuidos en estas categorías, sino que al desagregarse se encuentra que el 50% son de la categoría entretenimiento, mientras el 8% lo es comercial, 5% laboral y 5% residencial.

Considerando solo los lugares que tienen un comentario registrado e incluido en el análisis, todas las comunas presentaron la categoría “entretenimiento” como la más frecuente, lo que confirma la relevancia de esta categoría en el uso de Foursquare por parte de los usuarios.

En cambio, cuando se consideran la totalidad de los 69504 lugares, independientemente de tener algún comentario o no, entonces las categorías de clasificación se comportan de manera diferente según la comuna.

5.5 Características de las comunas según lugares y comentarios

En la figura 5.14 podemos ver un perfil de cada una de las comunas, ordenadas descendientemente por cantidad de lugares, según la distribución porcentual de las tres categorías más importantes de sus lugares.

³⁵ Por “residencial” se considera a categorías de departamentos y hogares, mientras que “entretenimiento” contiene restaurantes, bares, cafés. La categoría “comercial” contiene fundamentalmente tiendas, centros médicos y todo tipo de comercios, y finalmente “laboral” son oficinas, bancos y estaciones de transporte, entre otros.

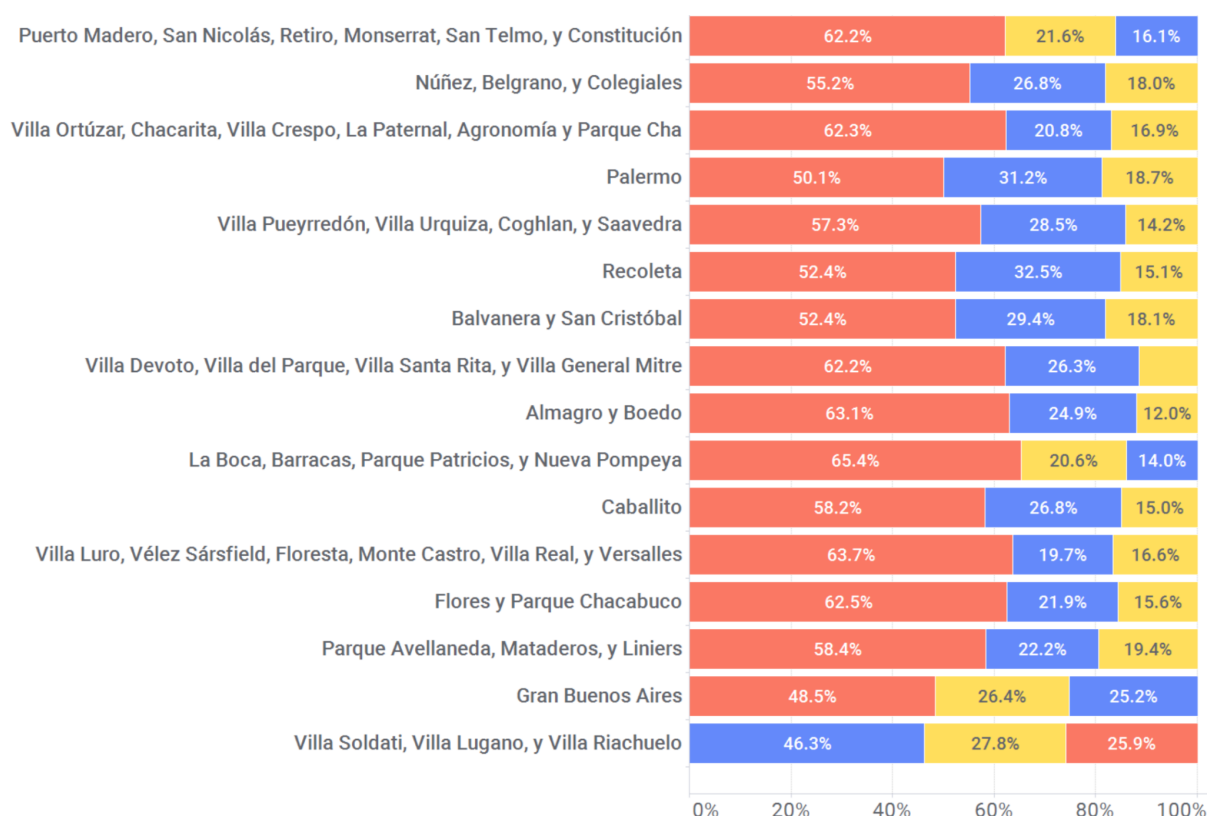


Figura 5.14: Proporción de tipos de lugares recolectados para cada una de las comunas. Categorías: entretenimiento (rojo), comercial (azul) y laborales (amarillo). Las comunas fueron ordenadas en orden descendente según cantidad de comentarios, mientras que las categorías intra comunas fueron ordenadas descendentemente de izquierda a derecha.

En concordancia con el perfil de cada una de las comunas, comerciales o de entretenimiento, también podemos observar que la distribución de comentarios por comuna se da en diferentes momentos del día³⁶. Por ejemplo, Palermo tiene un perfil más nocturno que Puerto Madero y Retiro. La zona sur de la ciudad, comprendida por barrios como La Boca, Liniers, Parque Patricio, Pompeya, Lugano, Soldati y Mataderos, entre otros, presenta por el contrario un perfil más diurno que el resto de la ciudad. En la figura 5.15 podemos ver estas diferencias de la distribución de comentarios en momentos del día según comuna.

³⁶ La clasificación de horarios en categoría fue: 0 AM a 6 AM: Nocturno, 7AM a 6 PM: Diurno, y 7 PM a 11 PM: Tarde-Noche.

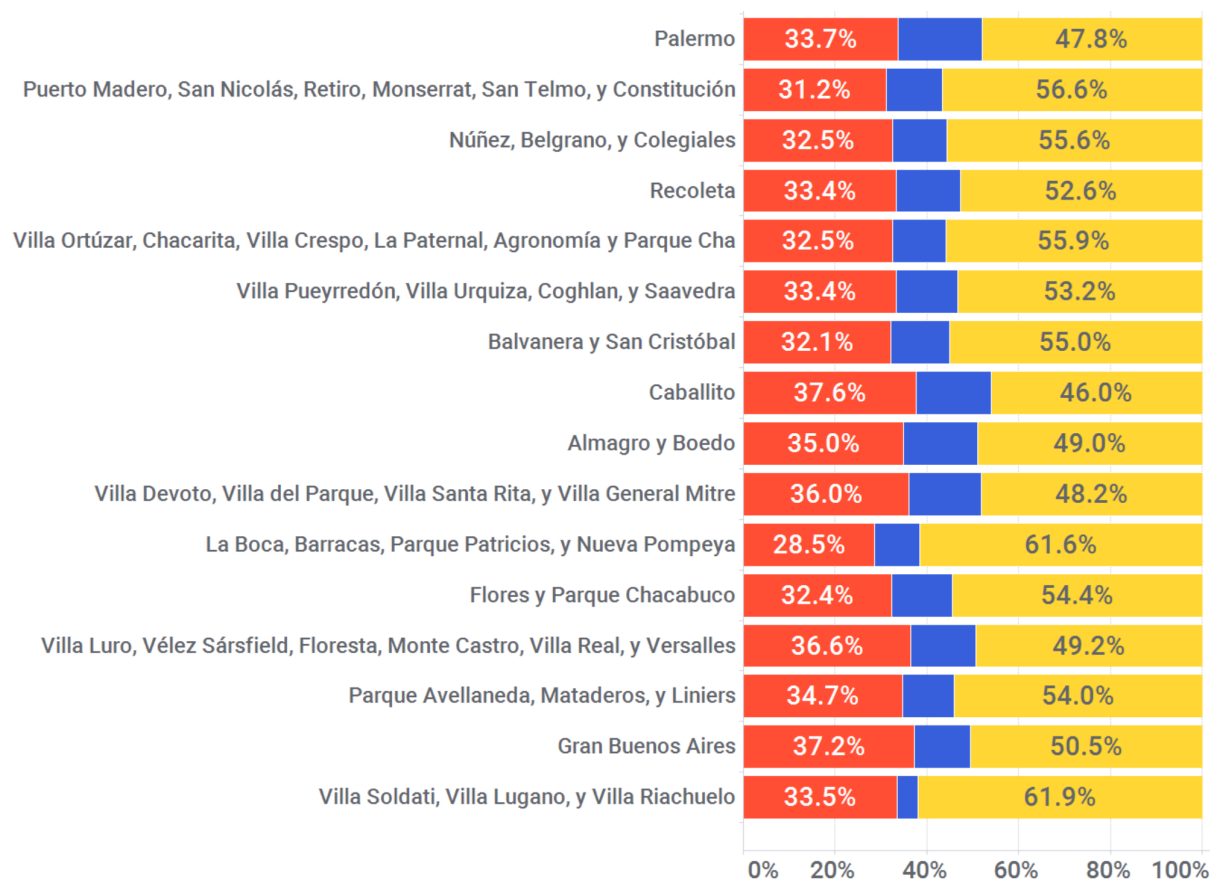


Figura 5.15: Proporción de comentarios según momento del día según comuna. En rojo se observan los comentarios de tarde-noche, en azul los nocturnos y en amarillo los diurnos.

6 Análisis de la red social de Foursquare de comentadores de lugares en CABA

Las redes sociales reales no son aleatorias ni regulares, poseen características únicas donde el orden y la estructura coexiste con el desorden en distintas medidas. En las siguientes secciones se analizan algunas de estas propiedades que nos permitirán identificar la estructura general, comunidades y características de la red social.

6.1 Comparación grafo extendido y grafo base

Si bien el análisis del trabajo se concentra en el Grafo Original (GO), tal como se explicó en la sección 4.2, cabe describir brevemente el Grafo Extendido (GE). El proceso de depuración de ambos grafos implicó excluir outliers (tal como se explicó en la sección 3.1), como también identificar a los usuarios únicos comentadores de los lugares, deduplicar los lazos entre los mismos y finalmente identificar los componentes principales conectados en los mismos.

En el caso de GO, se descartaron 5234 componentes no conectados al grafo principal. De los componentes excluidos, se detectaron dos tipos, los cuales son:

- 218 componentes tienen un tamaño entre 1 y 4 nodos.
- 5016 otros componentes consisten en nodos únicos, sin aristas, los cuales no están conectados con ningún otro nodo de GO.

De igual manera para el Grafo Extendido (GE), se descartaron 892 componentes y se seleccionó solamente el componente conectado principal.

En la tabla 6.1 se puede observar un resumen de ambos grafos y algunas de sus medidas, que serán ampliadas para GO en la siguiente sección.

Métrica	Grafo GO	Grafo GE
Nodos	31385	671805
Aristas	229224	2004852
Componentes conectados	1	1
Grado promedio	14,6	5,9
Densidad	0,000465	0,00000857
Coef. Clustering Promedio	0,1304	0,0613
Transitividad	0,0134	0,0108

Tabla 6.1: Características de grafo original (GO) y grafo extendido (GE)

Como puntos relevantes, se puede encontrar que el grado promedio de los nodos de GE es considerablemente menor, debiéndose principalmente a que se incluyen nodos que no han dejado comentarios en CABA pero no se consideran todas sus aristas excepto aquellas con un usuario que haya dejado un comentario en CABA. Esta última razón también es válida para explicar las diferencias de transitividad, clustering promedio y densidad.

6.2 Descripción de grafo original (GO)

Si bien hay numerosas métricas de centralidades, en este trabajo se describe la red con los índices de centralidad comúnmente utilizados en el campo de SNA que habitualmente son suficientes para caracterizar las redes sociales [36].

Adicionalmente a esto, se encuentra que muchas de estas métricas están correlacionadas entre sí y describen en parte las mismas propiedades, por lo que tan solo algunas de ellas son suficientes a motivos prácticos de análisis de la red social [37].

A modo de ejemplo podemos observar en la figura 6.1 un gráfico de dispersión entre centralidad de auto vector y distribución de grado [38].

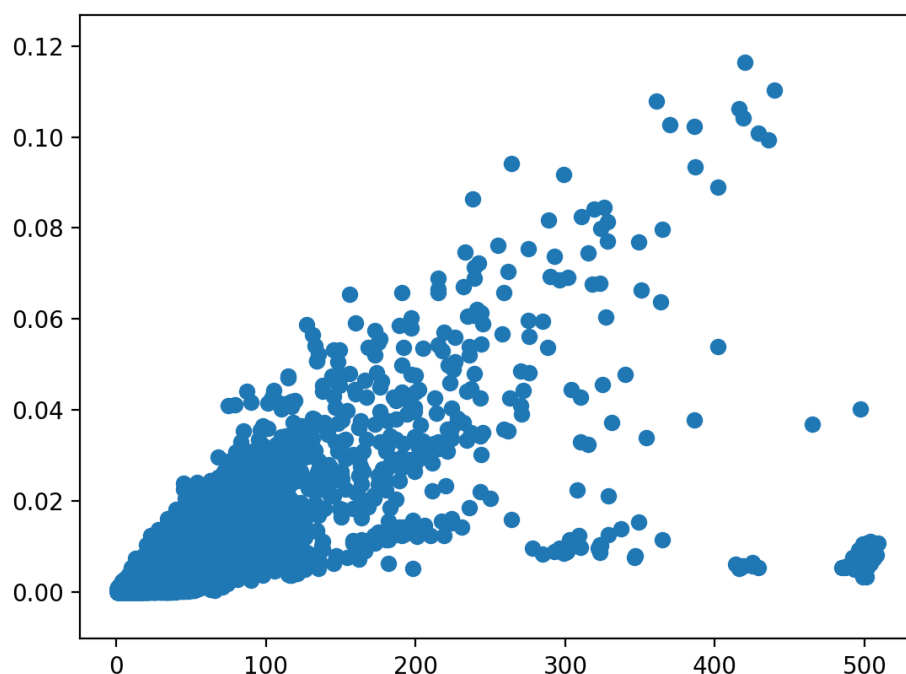


Figura 6.1: Centralidad de autovector (*eigenvector degree*) en el eje Y, por distribución de grado en el eje X

Podemos observar que existe una relación entre el grado y el valor de la importancia del grado de sus amigos. Esta correlación entre métricas se ha observado en algunos estudios [39], y este caso en particular será ampliada en la sección de grado de afinidad selectiva.

En la figura 6.2 se listan las correlaciones más significativas entre una matriz de correlaciones para las medidas más importantes.

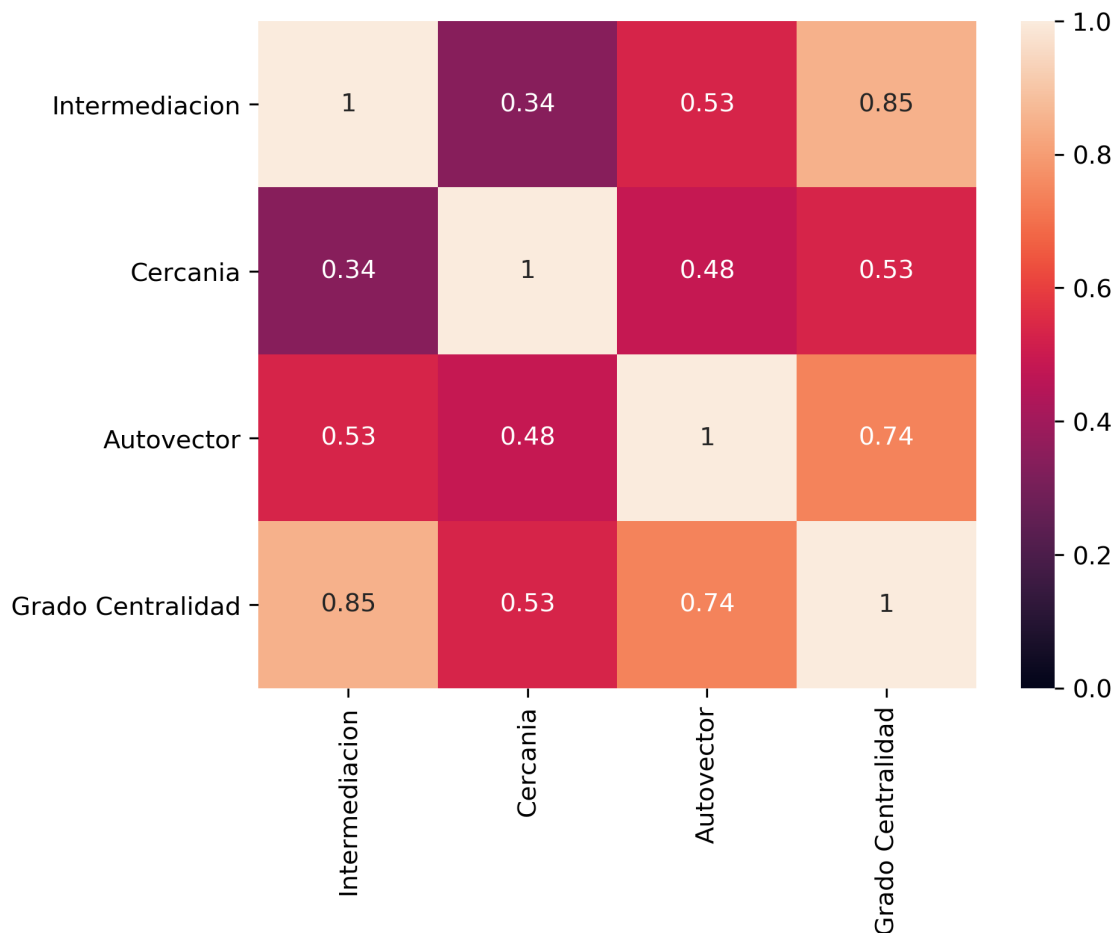


Figura 6.2: Valores de correlación r de Pearson entre medidas de centralidad representadas en una matriz de calor.

Centralidad de grado, Intermediación, Autovector y Cercanía son métricas de centralidad derivadas de la matriz de adyacencia. Aunque existen diferencias teóricas entre todas estas métricas, podemos ver tanto en otros estudios como en esta red, que las mismas están muy correlacionadas.

Comúnmente la recomendación [40] sobre el uso de estas métricas, es usar la centralidad de grado para medir comunicación y transferencia de información, intermediación para control y cercanía para estimar el nivel de eficiencia o conveniencia de un nodo, mientras que autovector para la importancia relativa.

Si bien los objetivos de este trabajo no son describir las relaciones de los actores a escala individual o los detalles de cada una de las comunidades, la importancia de la posición estructural de cada uno de los nodos dentro de la red social ayuda a identificar actores centrales con roles asociados a difusión, estabilidad e intermediación entre personas y comunidades. Las medidas de centralidad son importantes para identificar actores con poder de influencia, mediación, o centralidad en la red.

Como podemos observar en la tabla 6.1, el grado promedio del grafo original es de 14,6 aristas. Esto se analizó en más detalle en la descripción de amigos de la unidad 5.

Podemos identificar que el grado promedio es algo menor que el de otros estudios [41][42] en la misma plataforma, pero las diferencias se pueden deber a la manera de construcción del grafo analizado restringida solamente a usuarios con comentarios en CABA.

Al evaluar la distribución de grado podemos observar en la figura 6.1 que la misma parece seguir una ley de potencia donde existen pocos nodos con mucho grado y muchos nodos con poco grado. Esto se analizará en más profundidad en la sección de Ley de Potencias.

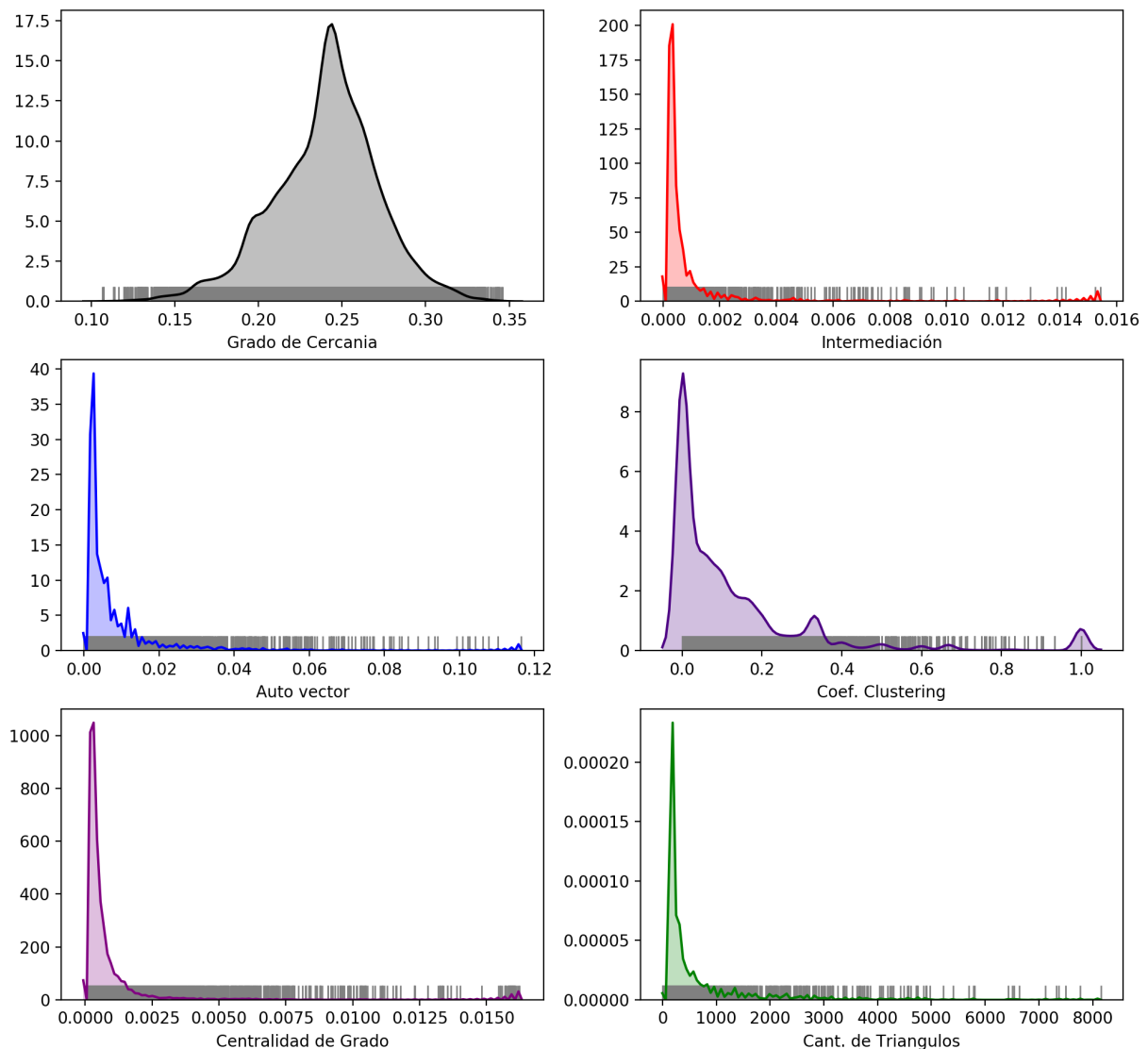


Figura 6.3: Estimación de densidad Kernel para métricas de eigenvector, cercanía, betweenness, grado, coeficiente de clustering y triángulos para grafo GO.

En cuanto a evaluar cuán centralizada está la red social, es decir cuánto de la centralidad está concentrada en pocos nodos, adicionalmente a la distribución de grado ya descrita, también podemos tener en cuenta la distribución del autovector. Esta métrica es una versión más elaborada de la distribución de grado, ya que asume que no todas las conexiones tienen la misma importancia, tomando en cuenta la calidad más que la cantidad de conexiones. Un usuario con pocas conexiones podría tener una centralidad de autovector alta si esas pocas conexiones estuvieran muy bien conectadas a otras.

Como podemos observar en la figura 6.3, esta distribución se asemeja a la de centralidad de grado, y nos indica que la mayoría de los nodos poseen un coeficiente bajo mientras pocos nodos tienen una influencia alta sobre otros nodos de la red social. Esta distribución claramente asimétrica es confirmada al calcular el índice de Gini sobre los valores de autovector. El mismo es de 0,84, sobre un valor

que va de 0 a 1, indicando 1 el máximo nivel de desigualdad entre los nodos, lo que nos está marcando gran disparidad en los valores de autovector.

En tanto a medidas de estructura de la red, como se puede observar en la figura 6.4, se ha encontrado que el coeficiente de clustering disminuye en parte a medida que aumenta el grado de los nodos, indicando que los nodos con mayor grado pueden estar haciendo un uso menos orientado a relacionarse con verdaderas amistades, y en cambio establecen relaciones indiscriminadas.

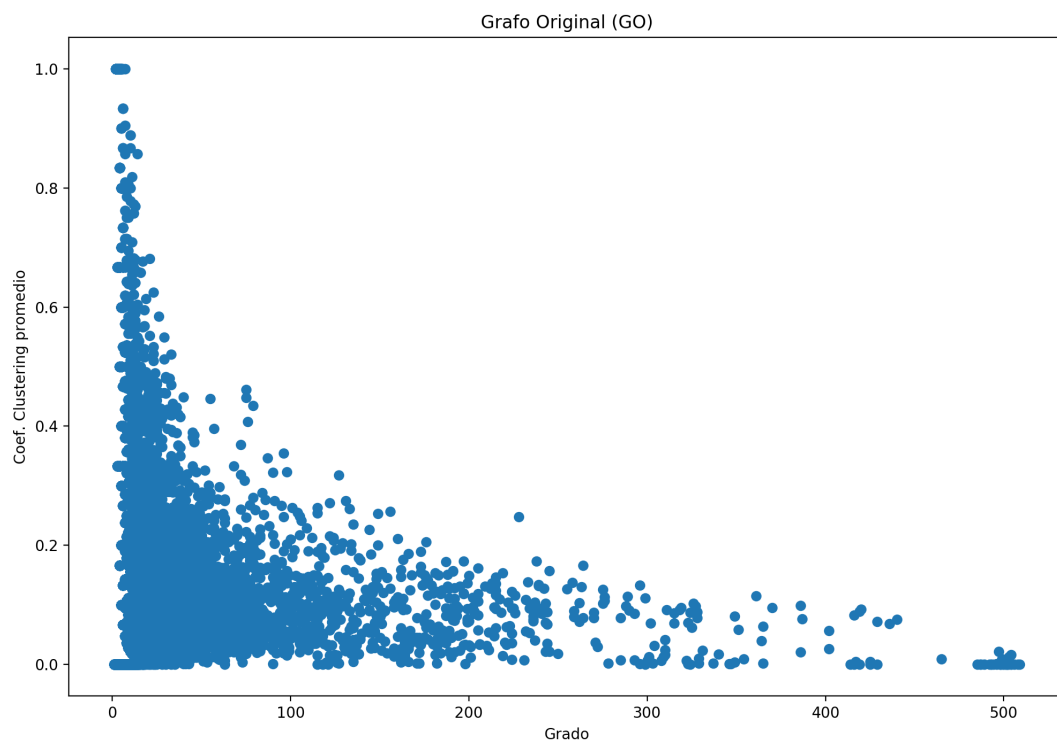


Figura 6.4: Coeficiente de Clustering promedio (eje y) según grado (eje x)

Se seguirá analizando el comportamiento de clustering del grafo en la sección 6.6 de análisis de mundo pequeño.

6.3 Resiliencia

La resiliencia mide el impacto en la conectividad y cohesión de la red cuando uno o más nodos son removidos. La mayoría de las redes sociales son robustas a la remoción de nodos, pero no tanto a la remoción de nodos de alto grado.

La resiliencia es una propiedad importante en tanto mide la redundancia de caminos entre los diferentes nodos y la interconectividad de la red, ya que cuando una arista no está disponible sigue existiendo conectividad en la red.

Debido a la complejidad computacional que era requerida para medir la cantidad de nodos o aristas mínimos para volver el componente conectado en un componente desconectado, se utilizó en cambio el cálculo de densidad, el cual mide la fracción de posibles aristas que existen. A mayor cantidad de aristas que existen en una red, significa más caminos redundantes entre los nodos.

En este caso el cálculo de densidad fue de 0,000465 (sobre un valor máximo de 1, en donde todos los nodos están interconectados entre sí) lo que indica una red muy poco densa.

Este valor de densidad es razonable cuando se tiene en cuenta que el grafo base contiene solamente las personas que han dejado comentario en CABA, y sus vínculos entre estas mismas personas, pero no la totalidad de relaciones que poseen los nodos.

6.4 Afinidad selectiva

Las redes sociales reales suelen presentar características de afinidad selectiva entre sus nodos basado en sus atributos (edad, género, etnia, etc) como también por atributos estructurales o propios de los nodos (grado de cercanía, etc). Esta afinidad selectiva, también suele llamarse homofilia en sociología y mide la tendencia de las personas a relacionarse con personas similares.

El grado de intermediación funciona como complementario de la distribución de grado para caracterizar a las redes complejas, ya que redes con una distribución de grado similar puede diferenciarse significativamente por este coeficiente.

Newman [43] propuso un coeficiente para medir la afinidad selectiva, detallado en la sección 4.5.5, con el cual se evaluó cuan correlacionado linealmente están los grados de los usuarios que tienen un lazo de amistad.

Es decir, se evaluó si aquellos nodos con alto grado están conectados entre sí, mientras que aquellos nodos con bajo grado se conectan con otros de similar grado, o por el contrario se trata de la relación inversa o se trata de relaciones de grado aleatorias. El coeficiente indica un valor entre +1 y -1, con un valor positivo cuando los grados de los nodos están correlacionados entre sí, un valor más cercano a -1 indica que están correlacionados inversamente, mientras que 0 que no están correlacionados.

En primera instancia se analizó la relación entre el promedio de grado de los vecinos de cada nodo y el grado de estos para analizar la afinidad selectiva de grados (figura 6.5).

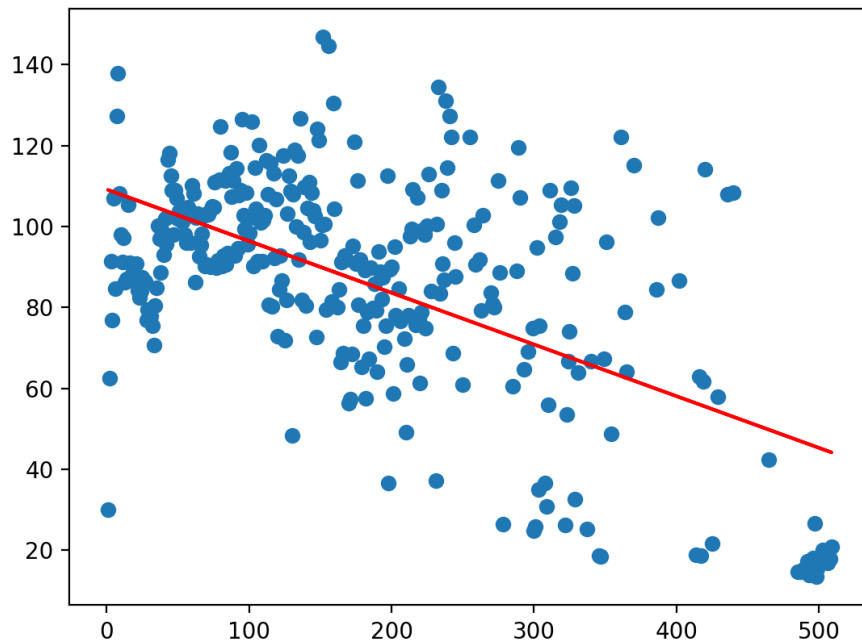


Figura 6.5: Asociación entre el grado del nodo (eje x) y el promedio de grado de vecinos (eje y)

Se puede observar a simple vista una tendencia de afinidad selectiva negativa con un valor de r de Pearson de 0,59, en donde nodos de mayor grado tienen vecinos con un grado menor en promedio.

Un patrón similar emerge cuando graficamos los grados de los pares de nodos conectados. En la figura 6.6 se observa que a medida que crece el grado de un nodo parece relacionarse con nodos de menor grado.

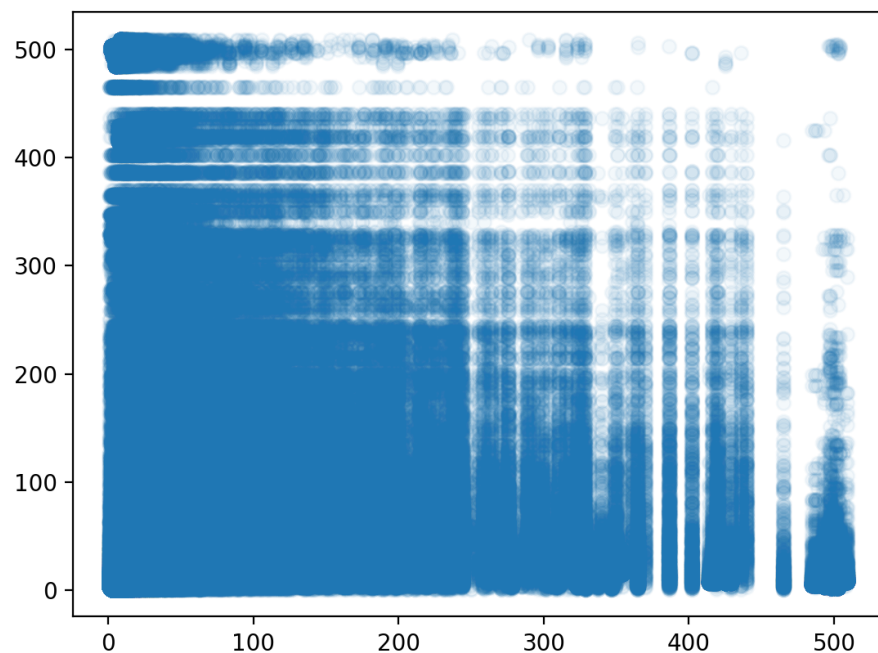


Figura 6.6: Grados de los nodos (ambos) conectados por las aristas.

Finalmente este mismo patrón se comprueba con el coeficiente de Newman, el cual fue $-0,11$, lo que nos indica que los nodos de la red están en un grado intermedio o neutral de afinidad selectiva con una tendencia a la afinidad selectiva inversa.

Este valor está en concordancia con algunos estudios [44] donde proponen que las redes reales tienen un bajo grado de afinidad selectiva de grado y en un 58% de ellas se clasifican en un grado denominado “neutral” (entre $-0,19$ y $+0,19$). Este valor también es consistente con el grado encontrado en otras redes sociales en línea como Youtube [45].

Cabe destacar que en la bibliografía se describe que una red puede no ser preponderantemente de afinidad selectiva, pero algunos nodos en particular serlo, habiendo regiones de homofilia local [46].

Finalmente se estudió la afinidad selectiva no en base al grado sino a dos atributos que resultaron significativos en el análisis de la red: género declarado e idioma de los usuarios clasificado en base a sus comentarios.

La fracción de nodos que se relacionan con el mismo género, esto es la proporción de aristas que conectan entre sí usuarias mujeres (10%) sumado a las aristas que conectan usuarios hombres entre sí (45,9%), suman apenas más que el 55% del total de aristas por lo que se puede decir que la red social no es homofílica desde el punto de vista de género. Esto se comprueba al calcular el coeficiente de afinidad selectiva para el género, el cual es de apenas $-0,0078$.

Si bien a nivel global no parece haber afinidad selectiva, cabe resaltar que redes

sociales como Facebook pueden tener valores globales de afinidad selectiva para género que no corresponden con la alta variación en diferentes regiones del grafo [47].

Como se observa en la tabla 6.2, si se toma en cuenta la afinidad selectiva para categorías hombre-hombre y mujer-mujer independientemente, se encuentra que el 45,9% de las aristas conectan nodos hombre-hombre, mientras solo un 10% lo hacen mujer-mujer.

Considerando relaciones en donde participan hombres, los mismos se relacionan entre sí en un 67,6% del total de sus relaciones. Esta proporción es similar para las mujeres en relación con los hombres, ya que poseen lazos con los mismos en un 68,7% del total de sus relaciones.

% aristas	hombre	mujer	Subtotal
hombre	45,9%	22,0%	67,9%
mujer	22,0%	10,0%	32,1%
Subtotal	67,9%	32,10%	100%

Tabla 6.2: Matriz de afinidad selectiva por género

Para el caso de idioma de los usuarios, observamos en la tabla 6.3 que el 76,9% de los nodos se relacionan con nodos de su mismo lenguaje, lo cual el coeficiente de afinidad confirma con un valor positivo de 0,11. Mientras que el 73% de las aristas conectan hispanos-hispanos, solo el 3,9% lo hace entre no-hispanos.

Sobre la totalidad de aristas que establecen las personas hispanas, los hispanos se relacionan en un 86% entre sí. Mientras que considerando la totalidad de relaciones que establecen los no hispanos, en un 74% lo hacen con hispanos.

% aristas	no-hispano	hispano	Subtotal
no-hispano	3,9%	11,5%	15,4%
hispano	11,5%	73,0%	84,6%
Subtotal	15,4%	84,6%	100%

Tabla 6.3: Matriz de afinidad selectiva por clasificación de idioma

6.5 Redes de escala libre y ley de potencias

Es posible clasificar las redes según su distribución de grados, las cuales muchas veces cumplen la ley de potencias y son denominadas como “redes de escala libre” (*free scale networks*).

La ley de potencias es un tipo especial de relación matemática entre dos magnitudes en donde una de las cantidades decrece según un exponente cuando la otra variable aumenta, tal como por ejemplo $f(x) = ax^{-k}$ donde habitualmente el exponente k es un valor entre 2 y 3.

A continuación podemos observar en la figura 6.7 una distribución que sigue la ley de potencias, en donde se identifica en violeta oscuro el 80% de los casos, y en tono claro el 20% restante.

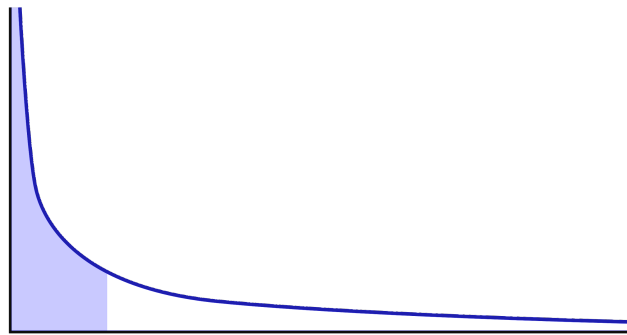


Figura 6.7: Distribución ley de potencias

Esta distribución de grados representa una de las tres propiedades principales que habitualmente se encuentran en las redes sociales, junto con las distancias cortas y el alto grado de clustering (descritas en la sección 6.6 de redes de mundo pequeño). La ley de potencias no es solo la menos intuitiva de estas propiedades sino la más estudiada y debatida desde que se describió a fines del siglo pasado [48]. Cabe resaltar que, si bien era de común entendimiento que las redes sociales presentaban habitualmente esta propiedad, este presupuesto está bajo cuestionamiento en la actualidad [49].

En un reciente estudio sobre una base de 928 redes biológicas, tecnológicas y sociales analizadas, se encontró que 49% de ellas carecen evidencia directa o indirecta de estructuras de red de libre escala. En el caso de las redes sociales, el 50% carece de evidencia y un 41% lo hace en forma débil, y para aquellas redes en donde era posible una distribución de potencias, otros tipos de distribución diferentes se ajustaban mejor a los datos.

Teniendo en consideración esta discusión actual sobre la estructura de escala libre en redes sociales, se siguió la premisa original de Barabási y Albert [50], quienes afirmaron que una gran cantidad de redes reales presentan una distribución de ley

de potencias, con una gran mayoría de nodos con bajo grado y una poca cantidad de ellos con un alto grado [51]. A continuación se analizará si la distribución de grados de la red analizada cumple dicha condición.

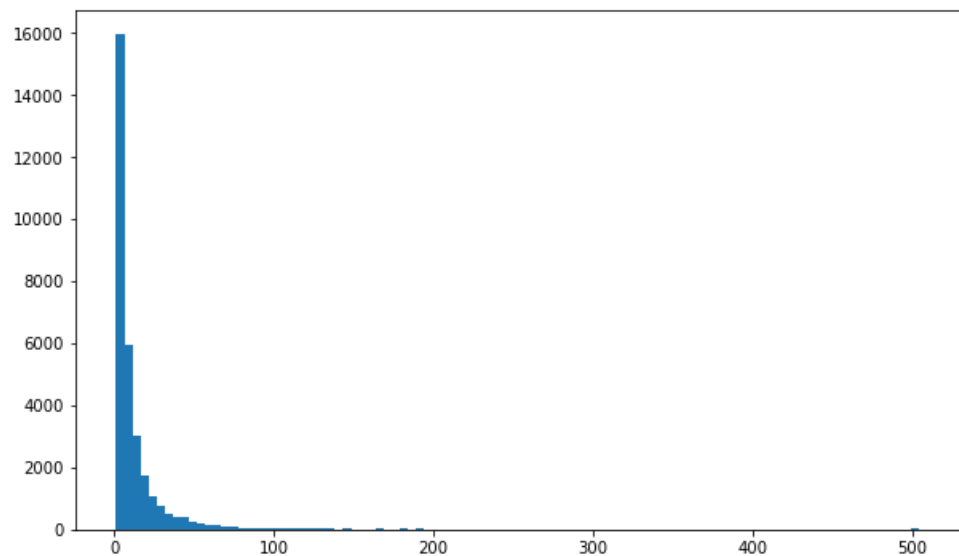


Figura 6.8. Distribución de grados de la red GO. Grado (eje x), cantidad (eje y)

En la figura 6.8, podemos observar que la distribución de grados tiene una similitud parecida a la ley de potencias, lo que indica que si bien la gran mayoría tiene relaciones con una baja cantidad de personas, en cambio una minoría se relaciona con cientos de usuarios.

Un primer paso para comprobar si la misma sigue la ley de potencias [52] consiste en escalar el gráfico aplicando el logaritmo en ambas partes de la ecuación, $\log f(x) = \log a - k \log x$, y al graficarlo deberíamos ver idealmente una línea recta.

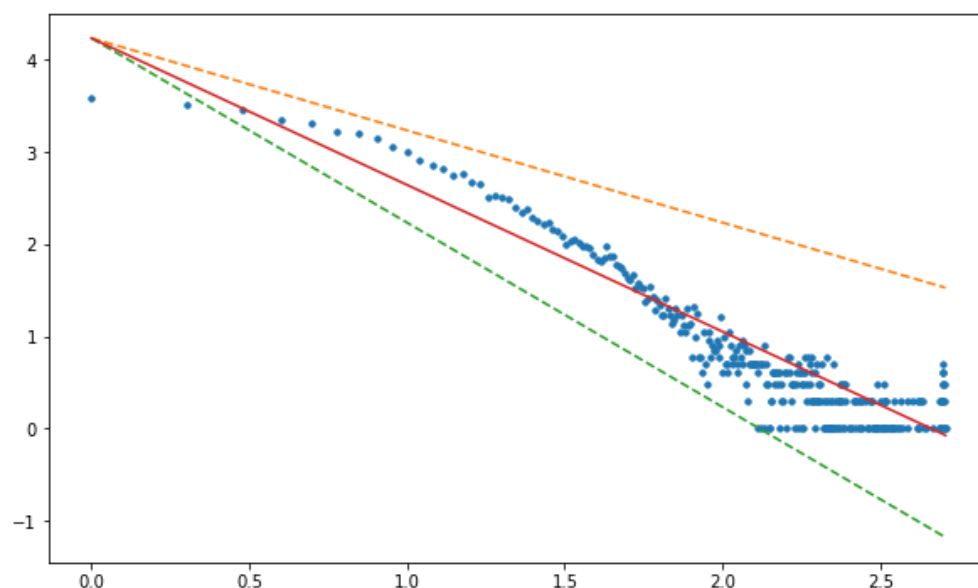


Figura 6.9: Números de nodos (eje y) según centralidad de grado (eje x) en escala logarítmica base 10. Rectas calculadas con diferentes beta de regresión lineal que mejor se ajusten a la distribución para una intersección = 4,22.

Como observamos en la figura 6.9, la distribución para toda la red sigue aproximadamente una recta, por lo que se profundizó el análisis, identificando en primer lugar qué porción de la información se ajusta mejor a la ley de potencias. Se utilizó el método de ajuste del paquete para Python denominado *powerlaw* [53]. El mismo busca el valor mínimo óptimo de la distribución de grado que ajusta a la ley de potencias (este parámetro es también denominado *xmin*), ya que la ley de potencias puede darse en una porción limitada de la distribución. El algoritmo ajusta la distribución a la ley de potencias empezando por cada valor único en el set de datos, y seleccionando el que presenta la distancia mínima Kolmogorov-Smirnov entre los datos y la ley de potencias.

Se encontró que el valor alfa, que es la pendiente de la recta, y habitualmente se encuentra entre 2 y 3, en este caso ajustaba con un valor de 1,59 (la cual debe ser siempre mayor a 1 para la ley de potencias). Este valor es similar al encontrado en otras redes sociales en línea, tales como Youtube, Orkut, y Flickr [54] y estudios donde se analizó la distribución de check-ins [55].

El mejor ajuste a la distribución de la ley de potencias, sin embargo, se obtiene utilizando un parámetro de corte $x_{min} = 32$, para el cual el valor de alfa es 2,49. En la figura 6.10 se puede observar en el gráfico izquierdo el ajuste del total del set de datos, mientras que a la derecha se observa el mejor ajuste con los parámetros descritos.

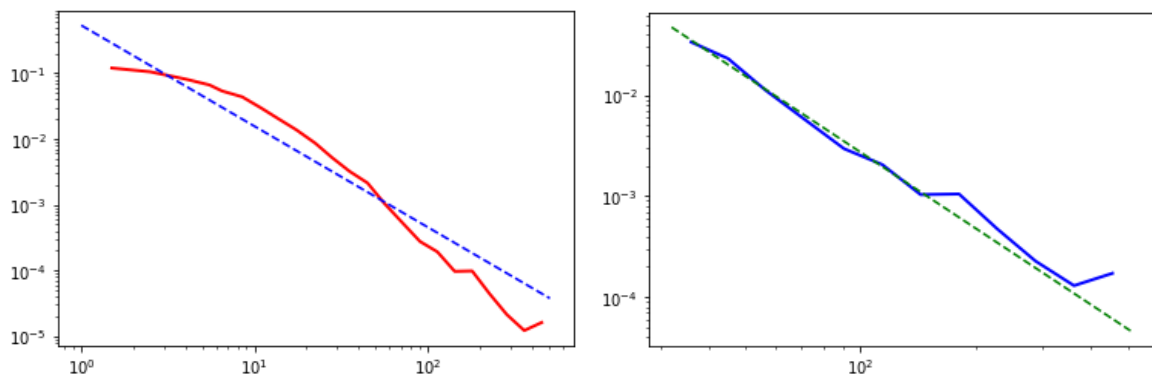


Figura 6.10: Ajuste de la distribución total ($\alpha=1,59$) en la izquierda, y ajuste parcial óptimo a la ley de potencias ($\alpha=2,49$) en la derecha, según distribución de grado.

Si bien la descripción gráfica nos puede dar una rápida idea del ajuste del conjunto de datos, es necesario probar la bondad de ajuste de distribuciones alternativas, como la normal logística (log-normal), o exponencial, entre otras, ya que estas deben ser evaluadas antes de concluir que la ley de potencias es la que mejor describe la distribución.

A tal razón se compararon gráficamente las distribuciones de la ley de potencia contra las log-normal y exponencial, que podemos ver a continuación.

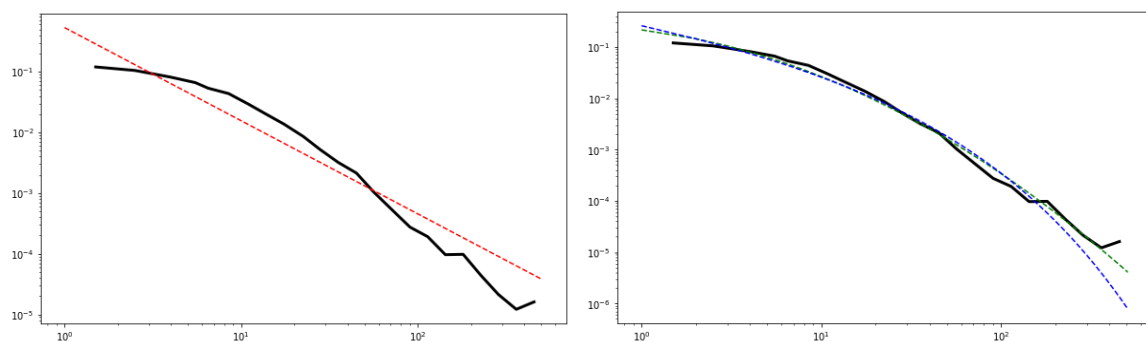


Figura 6.11. Cuadro izquierdo: Ajuste de distribución para ley de potencias. Cuadro derecho: Ajuste a log-lineal (verde) y exponencial estrecha (azul) según distribución de grado (eje x)

Si bien las dificultades en distinguir las distribuciones log-normal de la ley de potencias son comunes [56], se evaluó comparativamente, utilizando la razón de verosimilitud para identificar cuál de las dos distribuciones se ajustaban mejor.

La distribución log-normal (también llamada log-lineal) se ajustó mejor.

El valor de razón de verosimilitud entre la ley de potencia y la distribución log-lineal fue de -48.408. Este número es negativo ya que indica que es más posible que la

distribución se ajuste a la log-lineal. El valor de significancia de p fue $< 0,001$, por lo que se descartó que la ley de potencias sea la mejor distribución teórica que ajuste a la distribución de grados del grafo analizado.

Como ha sido descrito anteriormente, el ajuste a una ley de potencias en redes sociales es un presupuesto actualmente cuestionado dentro de la bibliografía, sumado a que estudios específicos de LBSN han también caracterizado que la distribución de grado ajusta parcialmente a la distribución log-lineal [57] mientras que otros lo describen como ley de potencias [58].

6.6 Un mundo pequeño

El fenómeno de mundo pequeño es la noción de que existe una distancia pequeña entre cualquier nodo de la red. Stanley Milgram fue el primero en demostrar la existencia del fenómeno de mundos pequeños en redes sociales reales a través de una serie de experimentos conocidos como “experimento de mundo pequeño” [59]. La principal hipótesis del estudio fue que pares de individuos aparentemente distantes estaban conectados por caminos cortos. Si bien no fue un término acuñado por Milgram, estos experimentos se conocen comúnmente como “seis grados de separación”, ya que era el valor de la mediana de nodos que separaba una persona de la otra [24].

En diversos otros estudios en redes sociales en línea, se ha encontrado que la cantidad de intermediarios promedio es aún más reducida [60]. La conclusión general ha sido aceptada por muchos años en un sentido amplio: las redes sociales tienden a tener un camino muy corto entre pares arbitrarios de personas [61], incluso comprobándose en redes sociales on-line en donde se tuvo la totalidad del grafo o estudios extensos en Facebook [62].

En el caso del grafo principal considerado en esta tesis, el valor de radio fue de 8, mientras que el diámetro lo fue de 15, pero estas métricas de por sí no nos indica si se trata de un fenómeno de mundo pequeño.

En términos matemáticos, el efecto de mundo pequeño significa que el diámetro de la red, cuantificada por la distancia geodésica promedio entre pares de nodos crece logarítmicamente, o a menor velocidad, con el número de nodos de la red. Entre las características estructurales de este tipo de redes, en un estudio de 1998 Duncan Watts y Steven Strogatz [63] analizaron distintas redes de mundo pequeño (redes de distribución de energía, biológicas y sociales) y concluyeron que una red de mundo pequeño se caracteriza por un promedio de caminos cortos bajo y un coeficiente de clustering alto [63].

El algoritmo de Watts-Strogatz parte de una red en forma de anillo (la cual asegura

un alto valor de coeficiente de clustering), y recablea aleatoriamente algunas de las aristas según una probabilidad p dada. Este recableo se hace entre nodos alejados entre sí para asegurar una creciente disminución del camino más corto en promedio, alcanzando su mayor aleatoriedad en el cableado de nodos cuando $p=1$, el cual se puede observar a continuación.

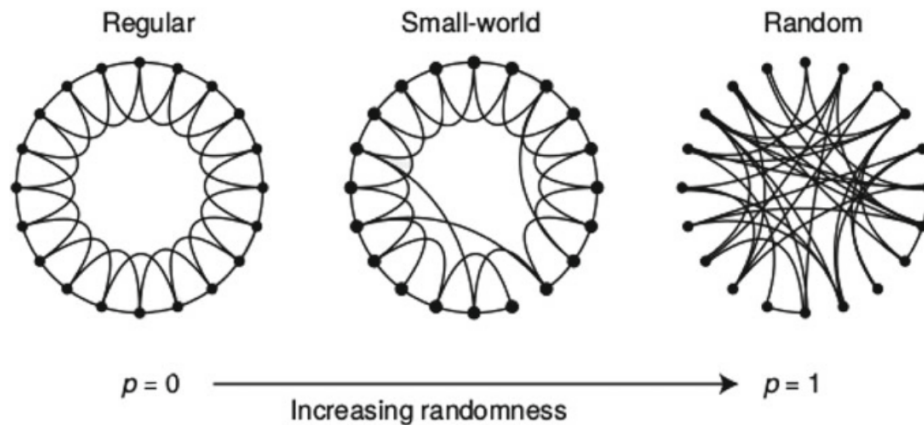


Figura 6.11: Procedimiento de recableado aleatorio para distintas probabilidad p

Este modelo genera un grafo con las características de mundo pequeño, para coeficiente de clustering C y la distancia promedio del camino más corto L , cuando los valores de p son intermedios. A continuación observamos ambos valores para los experimentos publicados por los autores.

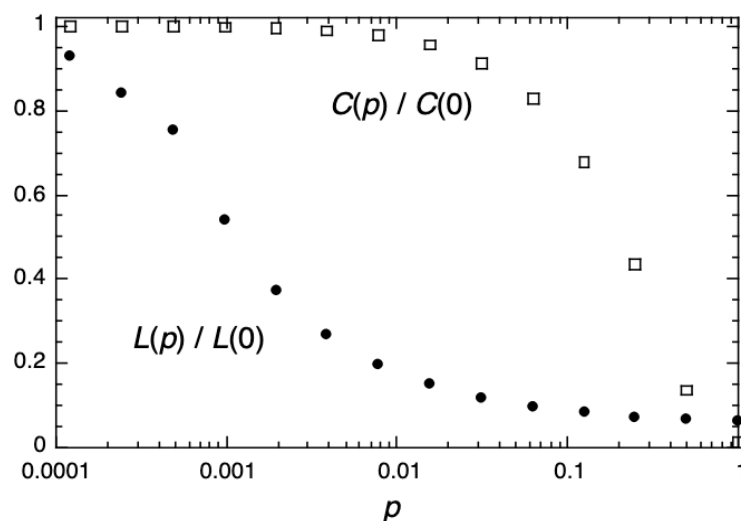


Figura 6.12: Coeficiente de clustering C y camino más corto L para p grafos generados según el proceso Watts-Strogatz.

Solamente los valores intermedios de los grafos generados aleatoriamente presentan características de mundo pequeño, con alto grado de clustering y caminos pequeños de corta distancia. Existen otras propiedades que se asocian habitualmente con este tipo de redes y que se logran reproducir en los grafos generados por el algoritmo descrito. Se encuentra habitualmente sobreabundancia de hubs (nodos de alto grado) que sirven como conexión para mediar y reducir el camino corto entre otros nodos.

Se procedió a realizar un experimento donde se utilizaron 7 probabilidades de recableado (p), utilizando un algoritmo³⁷ generador de grafos basado en Watts-Strogatz y manteniendo siempre las mismas características topológicas en cuanto a cantidad de nodos (31385) y grado promedio para los mismos (14) que presenta el grafo analizado en esta tesis. Los resultados a continuación.

p	Clustering promedio	Camino más corto promedio
0,000001	0,6923	1066,26
0,00001	0,6922	955,34
0,0001	0,6921	210,59
0,001	0,6902	40,03
0,01	0,6721	11,48
0,1	0,5078	5,81
1	0,0004	4,23

Tabla 6.4: Resultado del experimento para grafos aleatorizados usando Watts-Strogatz

Podemos observar en la tabla 6.4 que a medida que se incrementa la probabilidad de recableado p, el coeficiente de clustering disminuye mucho más lentamente que el camino más corto promedio del grafo, siendo que la caída del valor de clustering solo ocurre con grafos con p alto.

Este efecto se debe a que al momento de recablear las aristas, se crean puentes que conectan nodos remotos y reducen así el camino más corto promedio del grafo disminuye, sin alterar el clustering del mismo.

³⁷ Se utilizó el método *connected_watts_strogatz_graph* del paquete NetworkX, el cual devuelve un grafo con un componente único conectado luego de n intentos. Otros algoritmos usualmente pueden devolver múltiples componentes y no es válido para la comparación con el grafo analizado.

En la figura 6.13 se puede observar con más claridad el comportamiento de ambas medidas utilizando una escala logarítmica para p , y su similaridad con la figura 6.12.

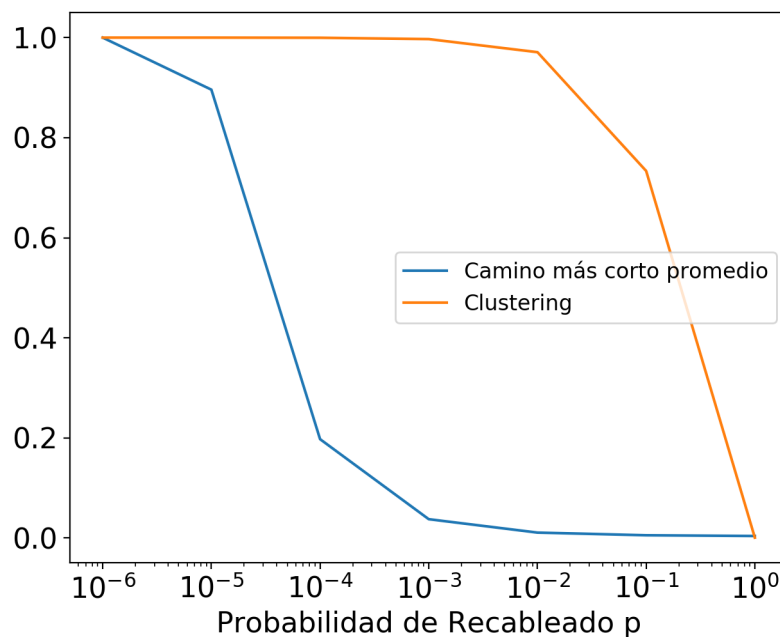


Figura 6.13: Coeficiente de clustering C y camino más corto L según p para experimento de grafos aleatorizados usando Watts-Strogatz

Si tomamos en cuenta que el grafo GO analizado tiene un coeficiente de clustering promedio de 0,148 y un valor de camino más corto promedio de 4,23, podemos concluir que no se encuentra en el extremo en que el recableado es totalmente aleatorio, por lo que podemos clasificar esta red como una estructura espacial de mundo pequeño.

Al mismo tiempo se ha encontrado valores similares para redes sociales como Facebook [62], en donde la distancia promedio es de 4,7 y otras redes sociales en línea [54] o LBSN como Foursquare, en donde la distancia promedio del camino más corto es menor a 6 y el coeficiente de clustering promedio es entre 0,18 y 0,26 [64].

Estas propiedades van también en línea con la naturaleza de mundo pequeño de las LBSN [65] estudiadas.

6.7 Estructura comunitaria



Figura 6.14: Red social GO. Colores según comunidades

La gran mayoría de las redes sociales tienen una estructura comunitaria [66] compuesta por grupos, también llamados módulos o clusters, lo que significa que hay grupos de nodos densamente conectados entre sí y poco conectados con otros grupos de nodos de la red.

Más grande que los nodos, pero a menor escala que la red, las comunidades subyacentes en la estructura de la red ayudan a describir la composición de la red social a un nivel de meso escala.

Podemos observar en la figura 6.14 la estructura comunitaria a muy alto nivel de los usuarios de la plataforma. La misma está graficada en función de realzar la cercanía

entre los nodos y distanciar aquellos que están levemente conectados, en distinto colores para cada una de las comunidades.

Para la identificación y partición de comunidades se utilizó la heurística de Louvain, basado en la optimización de la modularidad del grafo, descrito en la sección 4.5.9. Si bien el resultado de la heurística utilizada de Louvain no es determinista, el mismo presentó resultados similares en diferentes ejecuciones para segmentar en comunidades el grafo.

La modularidad resultado de la partición obtenida del grafo fue de 0,52. Este valor indica una estructura de comunidad alta, en donde las aristas tienden a ser incidentes entre nodos de la misma comunidad.

Sobre un total de 37 comunidades identificadas, las mismas tienen un promedio de 848 nodos, con un máximo de 4229 y un mínimo de 3 nodos. Podemos observar en la figura 6.15 el diagrama de densidad Kernel de la distribución del tamaño de las comunidades.

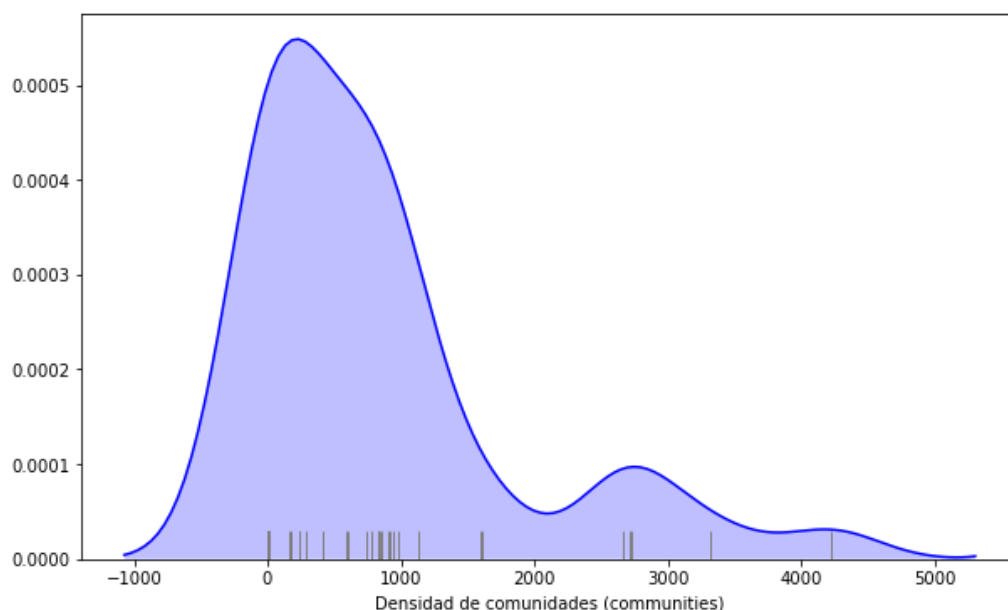


Figura 6.15: Estimación de densidad Kernel tamaño de comunidades

Debido a la gran cantidad de comunidades identificadas, se resume gráficamente en la figura 6.16 las características de estas, identificando en una escala de rojo (mayor valor) a azul (menor valor) comparativamente distintos valores estructurales al interior de las mismas, tales como grado promedio, diámetro, nodos, etc.

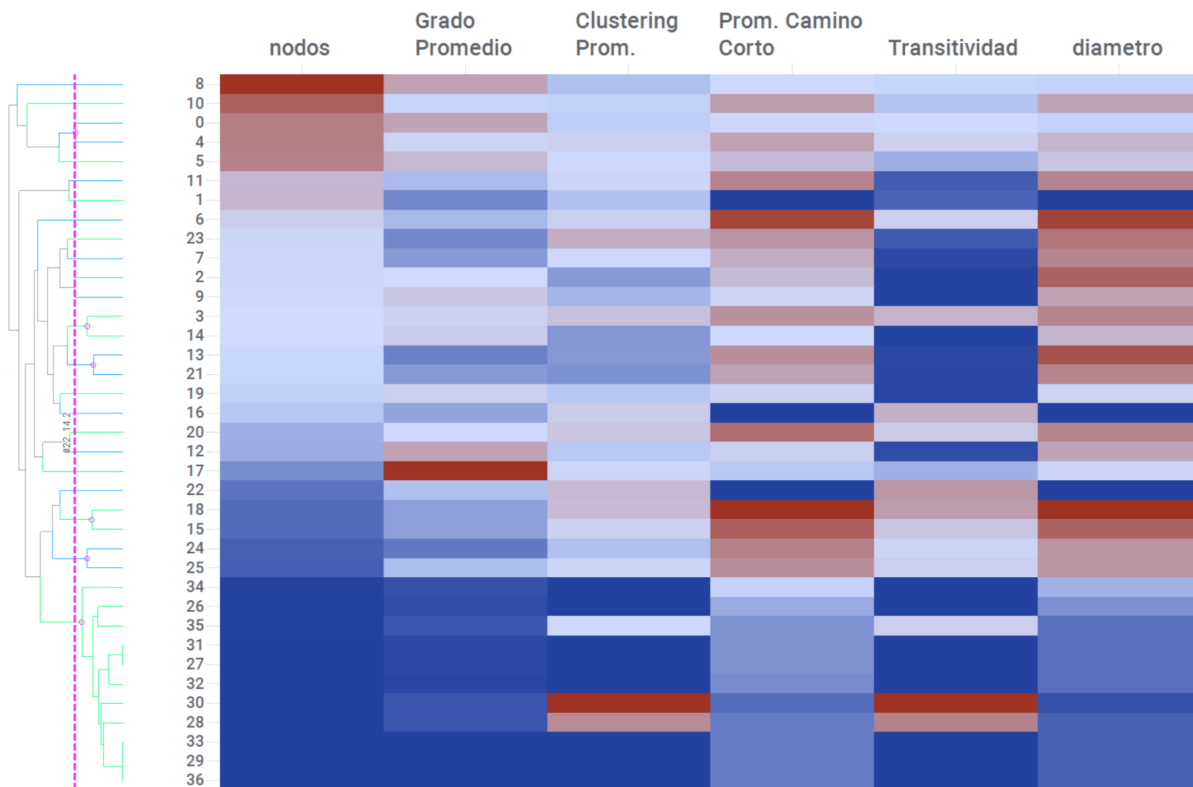


Figura 6.16: Mapa de calor de características y dendograma de las comunidades (método UPGMA). Mayor valor en rojo, menor en azul.

Tal como fue analizado para la totalidad de GO, en la sección 6.2, también hay una correlación entre alguna de las métricas analizadas dentro de cada una de las comunidades, como por ejemplo grado promedio según cantidad de nodos o diámetro.

Sin embargo, algunas de estas comunidades presentan excepciones, como por ejemplo la comunidad 30 o la 22, haciendo cada comunidad un subgrafo analizable en sí mismo. Para el primer caso, se trata de una comunidad de tan solo 3 nodos, mientras que la última tiene un nivel muy elevado de transitividad.

A modo de ejemplo, se observa en la figura 6.17 el grafo de la comunidad 24, de menos de 200 nodos, la cual tiene la característica de tener un 55% de usuarias clasificadas como no hispano parlantes (en color rosa), saliéndose de la distribución promedio de idiomas encontradas en otras comunidades.

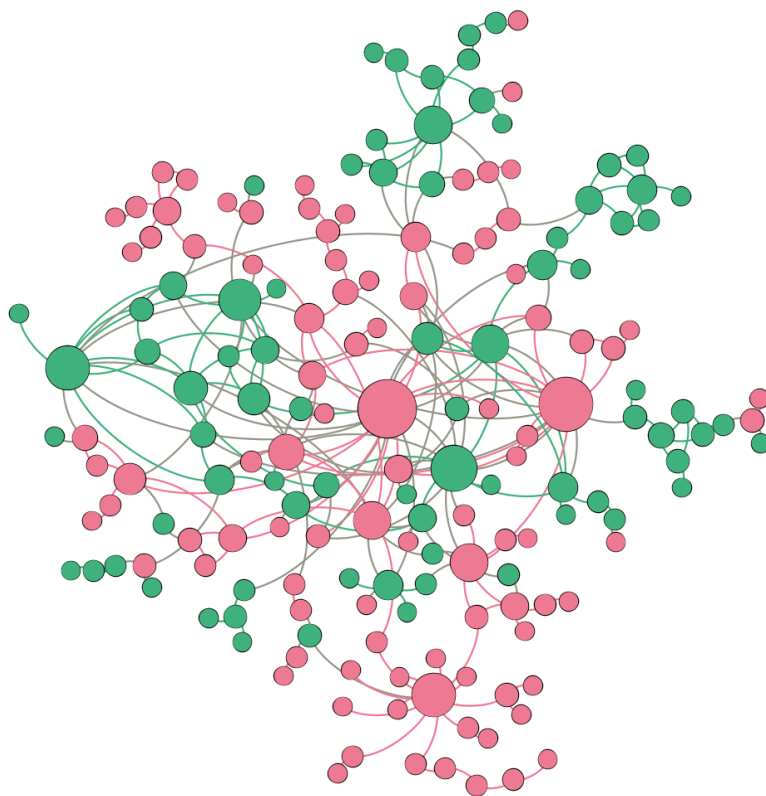


Figura 6.17: Grafo de comunidad 24. Hispano parlantes en verde. Tamaño por centralidad de grado.

Del análisis de la comunidad 24, nos da un indicio de que las personas parecen relacionarse entre sí de acuerdo a su idioma, pero esto no parece ser extensible en gran medida para la totalidad del grafo, y fue abordado en más detalle en la sección 6.4 de afinidad selectiva.

En la figura 6.18 se visualiza en el mapa de CABA esta misma comunidad 24 (rosa) en comparación con la comunidad 25 (azul oscuro), la cual tiene una densidad muy similar de nodos pero por el contrario tan solo un 15% de su composición es de personas no hispano parlantes.

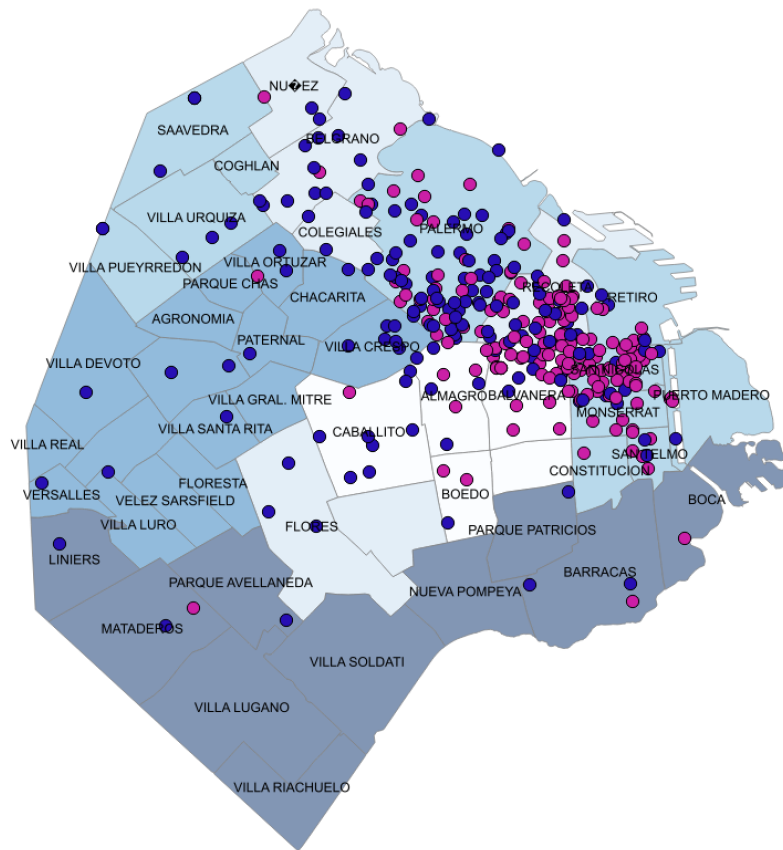


Figura 6.18: Centroides de integrantes de comunidades 24 (azul) y 25 (rosa) con diferente distribución de hispano parlantes.

Se encontró un indicio de que las personas no hispano parlantes frecuentan sectores más concentrados de la ciudad que personas hispano parlantes. Esto será abordado con más profundidad en el capítulo de distancias geográficas.

Finalmente, si bien la identificación de comunidades es útil para descubrir propiedades subyacentes en la red social que puedan ser exclusivas de esa comunidad o luego extenderse al conjunto de la red, también es posible identificar cómo se relaciona cada una de estas comunidades entre sí, y entender la centralidad de cada una de ellas. Para esto, se transformó cada comuna a nodos interconectados por aristas con un peso según la cantidad de usuarios que conectan de una comunidad a la otra.

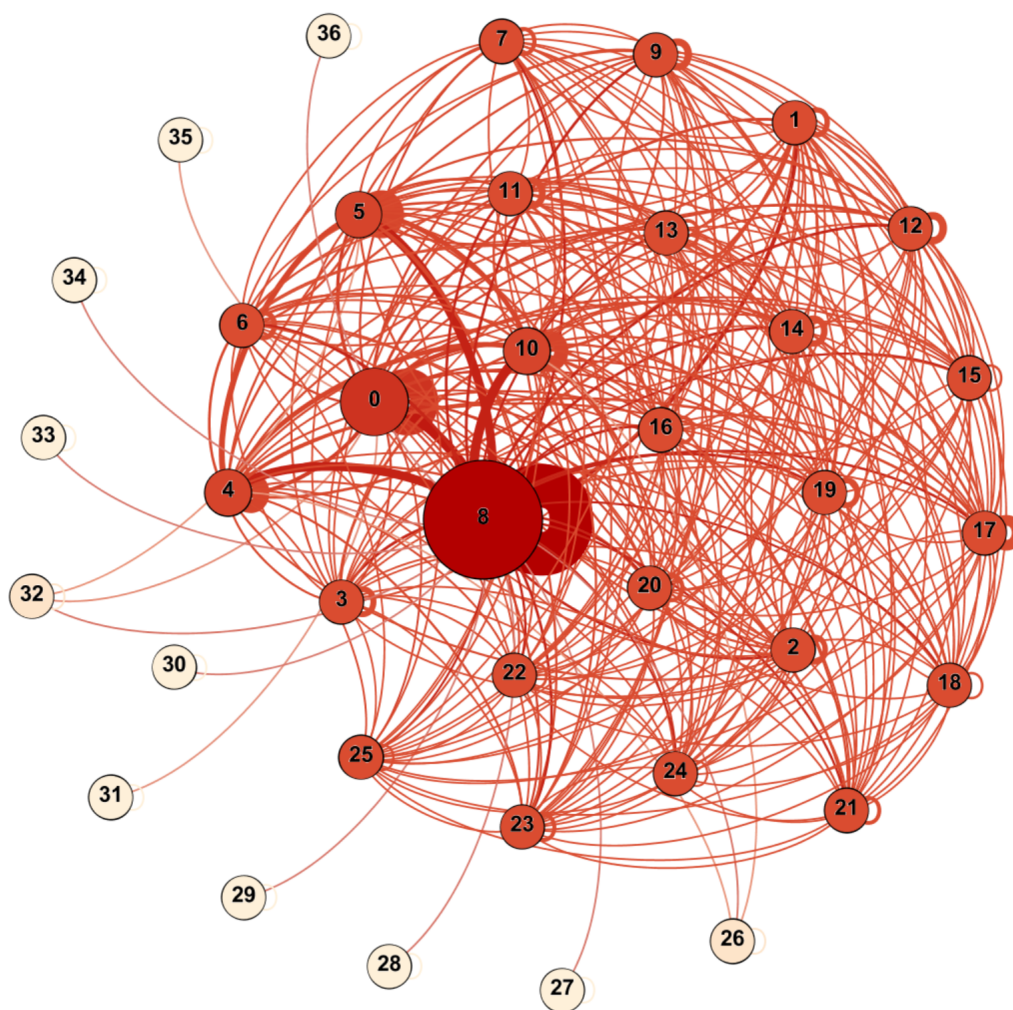


Figura 6.19: Grafo de comunidades. Densidad de conexión entre e intra comunidades representadas en el espesor de las aristas. Tamaño de los nodos según su grado de intermediación.

Podemos observar que la red de comunidades es más densa entre aquellas comunidades con alta centralidad de grado. A su vez, estas comunidades tienen una gran cantidad de aristas con otros nodos de su propia comunidad. En el gráfico se puede observar el tamaño de las comunidades en base a su grado de intermediación, mientras que el ancho de sus aristas está dado por el peso de sus conexiones con otras comunidades y consigo mismo, y el tono del nodo es dado por su grado.

También se puede observar visualmente, como ya ha sido analizada, que la centralidad de grado y el grado de intermediación están relacionados positivamente, principalmente en las comunidades más grandes.

7 Vínculos y espacialidad

La localización es uno de los aspectos cada vez más importantes de los servicios en línea. Mientras las usuarias están más dispuestas a compartir su información geográfica, en general con sus amigas pero también públicamente, los servicios en línea adaptan sus servicios según la localización de estas.

Las redes sociales han sido habitualmente estudiadas en función de su estructura, características de los nodos que los componen, cómo se relacionan y evolucionan en el tiempo, pero no tan extensamente en cuanto al componente espacial de las mismas.

Si bien la perspectiva sociológica ha estudiado la distancia geográfica en relación a los vínculos de las personas [67], la influencia de la misma en los vínculos en comunidades en línea es un campo de estudio que surge más recientemente.

En la bibliografía hay un acuerdo universal sobre que la probabilidad de un vínculo entre dos usuarios decrece a medida que aumenta la distancia geográfica de los mismos, pero las características e importancia de tal relación todavía siguen siendo debatidas según diferentes tipos de redes, plataformas y países. Se ha demostrado en diferentes estudios una influencia clara de la distancia en la probabilidad de contraer vínculos en algunas redes sociales en línea [68][69], junto con la importancia de las características de los nodos en la red social [70] como factor predictivo. Sin embargo, en la plataforma de Foursquare esta propiedad es muy heterogénea y se expresa cuando se consideran grandes rangos de distancia en el orden de los miles de kilómetros [64].

Finalmente cabe aclarar que otras investigaciones presentan la distancia geográfica como un factor explicativo menor de los vínculos entre los usuarios en redes sociales en línea [71] y otros estudios se focalizan en características de los nodos en la red social como manera de inferir los vínculos entre los mismos [72].

Debido a las limitaciones anteriormente descritas en la sección metodológica, no es posible saber la ubicación de los usuarios al usar la plataforma o al hacer check-in en los lugares, pero si es posible atribuir una ubicación a los nodos del componente principal del grafo utilizando la geolocalización de sus comentarios registrados en los lugares de CABA [73]. Esta técnica de inferir la geolocalización de los nodos según el registro de determinadas actividades ha sido utilizada extensamente en estudios variados, como registros notariales de la edad media, el análisis de la plataforma de Foursquare o Twitter [74].

7.1 Caracterización geoespacial de la red social

Para hacer un análisis geoespacial de la red social se analizó primero las características de los comentarios. Como se ha descrito en la sección 5.2, el

promedio de comentarios por persona es de 3,29 mientras que la mediana es de apenas 2 comentarios por persona y el 75% de las personas tiene 3 o menos comentarios. Cabe aclarar que estos valores se ven significativamente afectados cuando se compara con el total de nodos recolectados inicialmente debido a que 9569 nodos fueron excluidos por no pertenecer al grafo principal o caer bajo la categoría de outliers.

Para el cálculo del centroide espacial de los usuarios en base a sus comentarios (el punto medio geográfico que describe a cada usuario) se utilizaron comentarios únicos por persona por cada uno de los lugares en los que hubiera dejado comentarios, descartándose las localizaciones repetidas.

De esa manera, se evita darle un peso mayor a un lugar sobre otro si es que el usuario hubiera dejado múltiples comentarios sobre el mismo lugar. Aplicando esta poda, el universo de 102826 comentarios se ve reducido a 94515 comentarios, impactando levemente a la media y mediana de comentarios por personas como se indica en la tabla 7.1.

Medida	Cant. de comentarios por persona
Media	3,08
Desvío estándar	3,9
Mínimo	1
1er Cuartil	1
Mediana	2
3er Cuartil	3
Máximo	39

Tabla 7.1: Medidas de tendencia central para comentarios únicos por lugar

La poda de comentarios no arrojó una diferencia significativa en la distribución, ya que el 50% de las personas consideradas posee dos o menos comentarios, mientras el 75% tienen 3 o menos comentarios. Esta distribución de pocos comentarios únicos por persona, que podemos ver en más detalle la figura 7.1, será de especial importancia para el cálculo de la distancia geográfica de los nodos, el cual abordaremos en la siguiente sección.

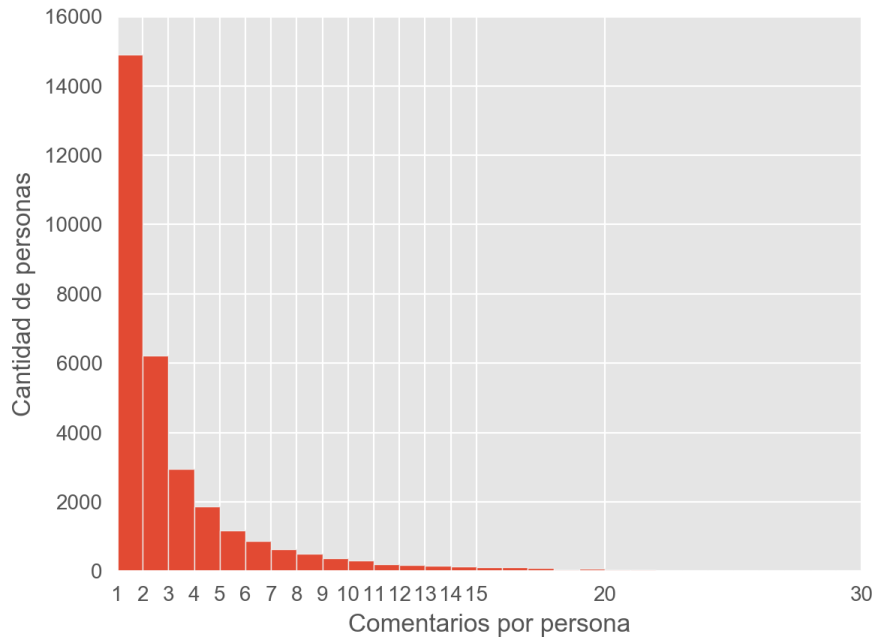


Fig 7.1: Cantidad de comentarios únicos en localizaciones por persona.

7.2 Calculo de centroides

Una vez seleccionado los comentarios, se procedió al cálculo del centroide geográfico de cada usuario en base a la ubicación de sus comentarios. En el caso que el usuario hubiera hecho un solo comentario, el centroide fue tomado de la posición geográfica del venue sobre el que había dejado el comentario, mientras que si el usuario hubiera dejado más de un comentario, entonces fue calculado el centroide espacial entre las múltiples ubicaciones únicas.

Si bien hay estudios que indican que se puede inferir la ciudad de residencia de una persona con información de Foursquare [63], el centroide en este caso no pretende captar donde el usuario habita o trabaja en sentido estricto, sino que representa una media espacial de lugares donde frecuenta y sociabiliza.

Cuando analizamos la distribución espacial de usuarios en relación con los venues donde se recolectaron los comentarios, podemos observar que hay una similitud en la distribución espacial entre los centroides calculados (Fig 7.2) y lugares (Fig. 5.12). Esto es debido a que gran parte de los usuarios tienen pocos comentarios, además de que la distancia entre los comentarios de los usuarios con múltiples venues es muy estrecha.

Sin embargo, en la figura 7.2 también podemos observar que los usuarios catalogados como no hispanos parecen tener patrones distintos a los hispanos, lo cual se analizará en las siguientes secciones.

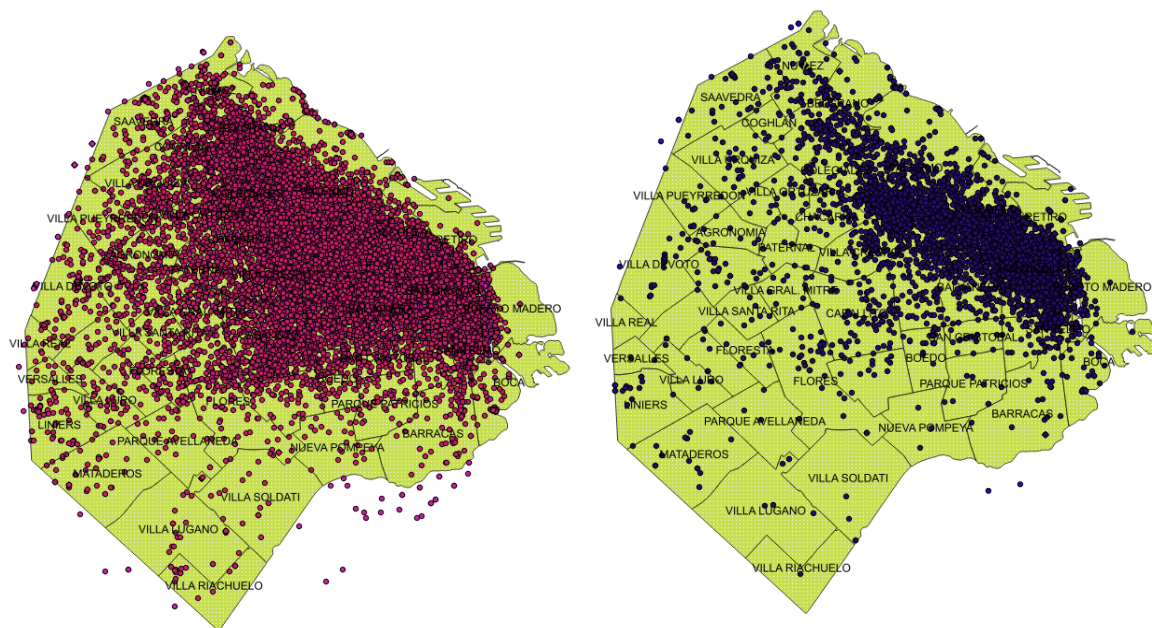


Figura 7.2: Distribución de nodos según su centroide. Izquierda: Hispanos, Derecha: No hispanos

7.3 Cálculo de distancia intra centroide y entre centroides

Se denomina distancia intra centroide a la distancia que existe entre los comentarios de cada uno de los usuarios, utilizados para determinar su media geográfica o centroide. Se denomina distancia entre centroides a la distancia que existe entre dos centroides de usuarios distintos. Podemos observar a continuación (Fig. 7.3) un esquema para clarificar cómo se realizan estos cálculos.

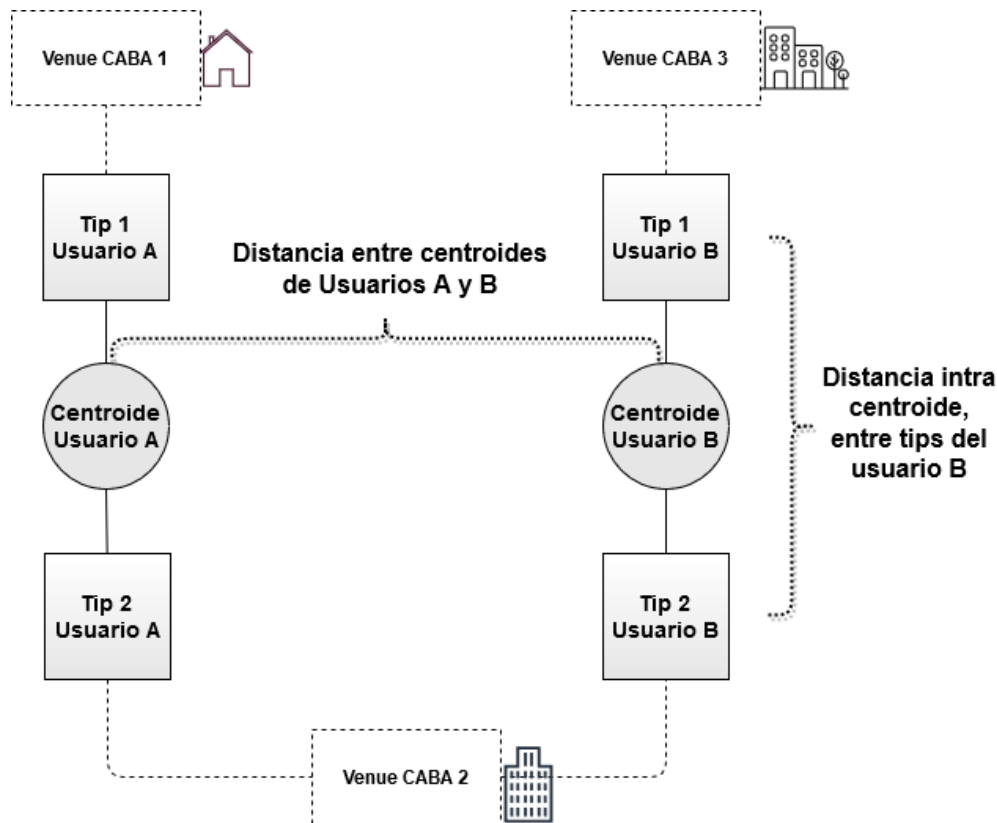


Figura 7.3: Cálculos de centroides según comentarios.

7.3.1 Distancia intra centroides

Cuando calculamos el promedio de la distancia intra centroide para todas aquellas personas que han dejado al menos un comentario observamos que la distancia promedio es de 1,86 KMs, pero estos valores ascienden a un promedio de 3,54km si consideramos solamente las aproximadamente 15000 personas que han dejado más de un comentario en lugares distintos. Es importante remarcar, como ya hemos visto, que los comentarios analizados sólo corresponden entre un 17 y 40% a los comentarios totales de los usuarios, por lo que la distancia entre todos los comentarios por persona es mucho mayor al reportado si se analizara la totalidad de comentarios, incluyendo aquellos hechos fuera de CABA.

En la figura 7.4 se observa en azul la distribución de la distancia intra centroide para todos los usuarios considerados, mientras en rojo solamente aquellos usuarios que tienen más de un comentario. El pico inicial de la distribución azul es debido a que una gran proporción de usuarios solo registran un solo comentario.

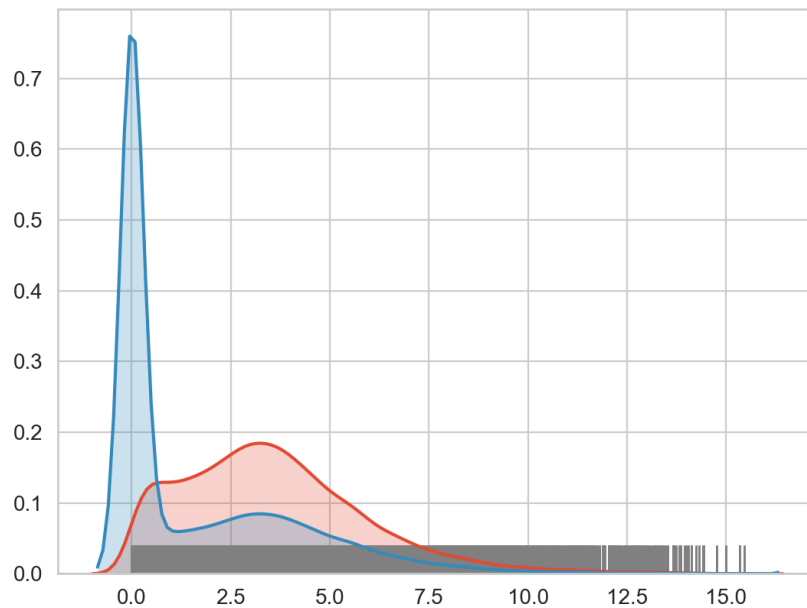


Fig 7.4: Estimación de densidad Kernel de distancias (en kilómetros) entre comentarios para todos los usuarios (azul) y usuarios con más de un comentario (rojo)

Esta última medida nos indica que considerando las dimensiones de CABA, la cual tiene aproximadamente un diámetro irregular de 17 kms, la mayoría de los comentarios son en una distancia mucho más reducida que la extensión de la ciudad y están espacialmente concentrados a poca distancia entre sí.

En otro estudio sobre Foursquare [73] se ha determinado que el 36% de los usuarios tienen en promedio un desplazamiento de menos de 1 kilómetros entre sus comentarios, mientras que el 70% de los usuarios tienen como máximo un desplazamiento de 150 km entre sus comentarios.

7.3.2 Distancia entre centroides

Una vez entendido la distancia entre los comentarios de las personas, analizaremos la distancia promedio que tienen las amigas entre sí.

En la figura 7.4 se observa la distribución de distancia promedio en kilómetros entre centroides de las 229224 relaciones de amistad (aristas) que componen la totalidad del componente principal del grafo reducido.

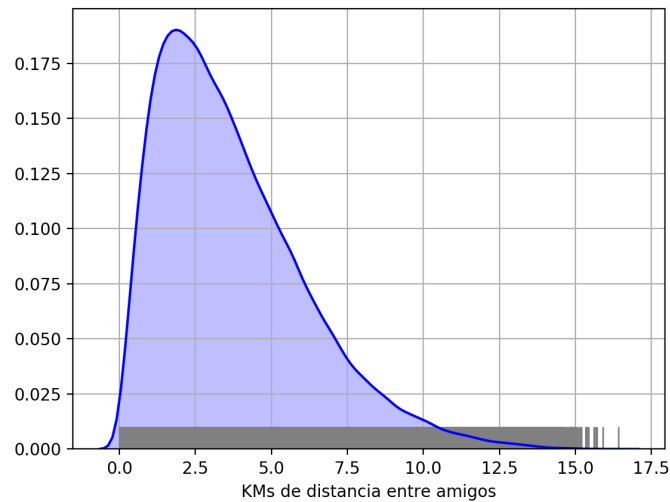


Figura 7.4: Estimación de densidad Kernel. Distancias entre amigos (en kilómetros)

Cuando se comparan la media de distancia promedio entre amigos hispanos y no hispanos utilizando una prueba T de Student de muestras independientes para comparar ambas medias, se rechaza la hipótesis nula con un p menor a 0,01.

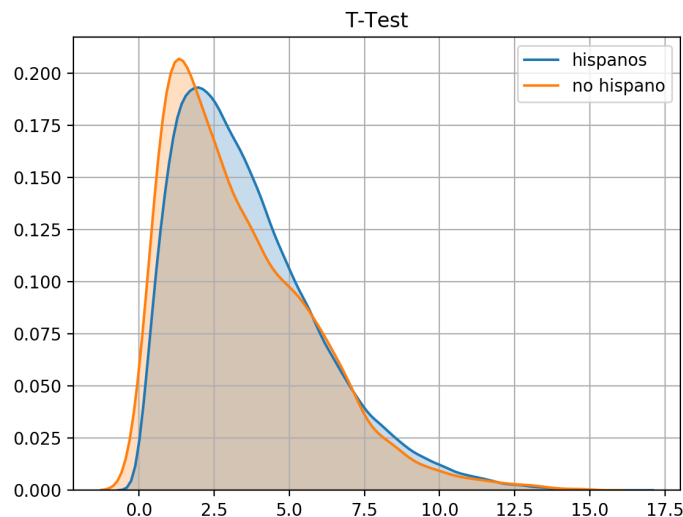


Figura 7.5: Estimación de densidad Kernel de las distancias en km entre amigos según hispanos (azul) y no hispanos (rojo).

La distancia promedio entre no hispanos es de 3,4 km mientras que entre hispanos es de 3,68 km, pero estos valores son más significativos en el caso de considerar personas con más de un comentario solamente, lo cual será abordado posteriormente.

7.3.3 Dispersión intra centroide y entre centroides

En la tabla 7.2 se observan las medidas de tendencia central para intra centroides como entre centroides. A su vez, ambos grupos están segmentados entre quienes tienen al menos un comentario y quienes tienen más de un comentario.

Medida	Distancia en KM intra centroides	Distancia en KM intra centroides para usuarios con más de un comentario	Distancia en KM entre centroides de amigos	Distancia en KM entre centroides de amigos con más de un comentario	Distancia en KM entre centroides de amigos con solo un comentario
media	1,86	3,49	3,73	3,21	4,48
desvío estándar	2,44	2,3	2,43	2,04	2,98
mínimo	0	0,0017	0	0	0
1er cuartil	0	1,78	1,85	1,64	2,05
mediana	0	3,23	3,24	2,78	4,02
3er cuartil	3,35	4,76	5,11	4,37	6,23
máximo	15,44	15,44	16,41	15,11	16,41

Tabla 7.2: Medidas de tendencia central para intra y entre centroides

Si bien la distancia intra centroides, considerando todos los usuarios del grafo principal, es menor que la distancia entre centroides de amigos, esto cambia cuando solamente consideramos aquellos usuarios que han dejado más de un comentario.

La distancia media intra centroide de usuarios con más de un comentario es levemente mayor que la existente entre centroides de amigos con más de un comentario. Esto quiere decir que la dispersión existente dentro de cada centroide es mayor que la distancia existente entre centroides de distintos usuarios.

Esto podría indicar que los amigos circulan por espacios geográficos similares en CABA, de manera que sus centroides no están claramente diferenciados unos de los otros.

Esta relación, mayores distancias intra centroides que distancias entre centroides, se mantiene para el grupo de personas de otros idiomas pero a menor escala, en donde la distancia entre amigas extranjeras es de 2,69 km contra los 2,89 km de la distancia promedio intra centroide.

Una vez analizado la distancia entre comentarios y amigas, analizaremos en la siguiente sección si el hecho de ser amigas en la red social tiene alguna relación

con la distancia geográfica entre los centroides de las personas que componen el vínculo.

7.4 Distancia geográfica y social

En los trabajos citados en esta sección, se ha encontrado que la geolocalización de las personas que integran una red social está correlacionada a la distancia social³⁸ que mantienen con otros integrantes de la red.

Abordaremos esta hipótesis en dos partes, primero midiendo si la distancia entre amigas es significativamente diferente de la distancia promedio entre personas que no tienen una amistad en la red social. Posteriormente, analizaremos si la distancia geográfica entre amigas está correlacionada con la fortaleza de sus vínculos.

7.4.1 Distancia geográfica entre amigas y no amigas

La hipótesis es que aquellos usuarios que tienen una relación de amistad tienen una distancia euclídea más cercana entre sus centroides que personas que no presentan una relación de amistad, debido a que posiblemente frecuentan lugares más cercanos que con personas que no tienen un vínculo de amistad.

Para hacer dicho experimento se calculó la distancia euclídea entre los centroides de todos los nodos entre sí, es decir las 492.493.420³⁹ distancias medidas linealmente en kilómetros entre dos puntos expresados en latitud y longitud dados. Luego se comparó si la distancia promedio es diferente entre el subgrupo de nodos amigos del subgrupo de usuarios que no son amigos entre sí.

Podemos encontrar a continuación las medidas de tendencia central de distancia para ambos grupos, subdividiéndose entre personas hispanoparlantes y otros.

³⁸ En este trabajo se entiende por distancia social basado en la teoría sociológica de redes. Podemos entender que la fuerza de una relación está basada en la frecuencia con la que dos usuarios comparten amigos y/o venues en común.

³⁹ No se tienen en cuenta la distancia entre los mismos nodos y solo se contabiliza una vez la distancia entre mismos pares de nodos ya que se trata de un grafo no dirigido.

Entre	Amigos GO			No Amigos GO			Student's T-Test
	Media (Distancia km)	Desvio Std (Distancia km)	N	Media (Distancia km)	Desvio Std (Distancia km)	N	P-Value
Usuarios Hispanos	3,68	2,69	167429	4,36	2,7	301167153	<0,001
Usuarios hispanos y no hispanos	3,9	2,56	52835	4,41	2,78	167923206	<0,001
Usuarios no Hispanos	3,4	2,43	8960	4,1	2,79	23403061	<0,001
Usuarios totales	3,73	2,43	229224	4,37	2,73	492493420	<0,001

Tabla 7.3: Distancia entre centroides para amigos y no amigos por categoría idiomática

En la tabla 7.3 se puede observar que las distancias entre amigas (3,73 km) es 0,6 kms menor que para personas no amigas (4,37 km), la cual es significativa con un $p < 0,01$.

Recapitulando, hemos encontrado una diferencia significativa entre la distancia promedio que mantienen amigas respecto a la distancia promedio que mantienen aquellas personas que no son amigas, y estas diferencias se mantienen tanto para el grupo de comentaradores en castellano, otros lenguajes y en la totalidad de los nodos.

En el caso particular de las usuarias que dejan comentarios en otros idiomas, la distancia es menor tanto para amigas como no amigas, lo que indica que este grupo de usuarias se mueve en un terreno más acotado que las comentadoras en castellano, tal como ha sido planteado gráficamente al principio de este capítulo. Así mismo vemos que la distancia es mayor para el caso que el par de amigos de la relación sean hispanos y no hispanos, tanto para el subgrupo de amigos como para no amigos, reforzando la hipótesis de que los usuarios locales dejan comentarios en áreas distintas de la ciudad que los no hispanos.

Comprobado que los centroides de las usuarias amigas son más próximos que las que no son amigas, procederemos a continuación a analizar si la distancia geográfica entre las amigas tiene además un correlato con la distancia social que

tienen entre sí.

7.4.2 Distancia Social y geográfica entre amigas

La cantidad de amigas en común entre dos personas funciona como un indicador de la fuerza de su amistad [75]. Podemos entonces suponer que personas que poseen mayor cantidad de amigos en común tengan una distancia geográfica menor entre sus centroides, ya que seguramente compartirán espacios de socialización o se moverán por espacios geográficos similares, como ya lo hemos comprobado para el grupo de amigos y no amigos.

Para medir la “distancia social”, en primer lugar, se ha calculado para cada par de usuarias amigas la cantidad de amigas que suman en total ambos nodos así como la cantidad de amigas que comparten en común. Para calcular la cantidad de amigos que comparten en común dos nodos dados, se ha calculado dos variables considerando la red reducida y la extendida. La primera tiene en cuenta amigos de la red social GO, reducida al componente principal que han dejado comentarios en la CABA (31385 nodos, 229224 aristas), mientras que la segunda (GE) incluye amigos directos de estos 31385 nodos que no han comentado en lugares de CABA (671805 nodos y 2004852 aristas). Ambas redes han sido descritas en más detalle en la unidad 6 de la tesis. En la figura 7.6 se grafica como se ha calculado la distancia social base y extendida.

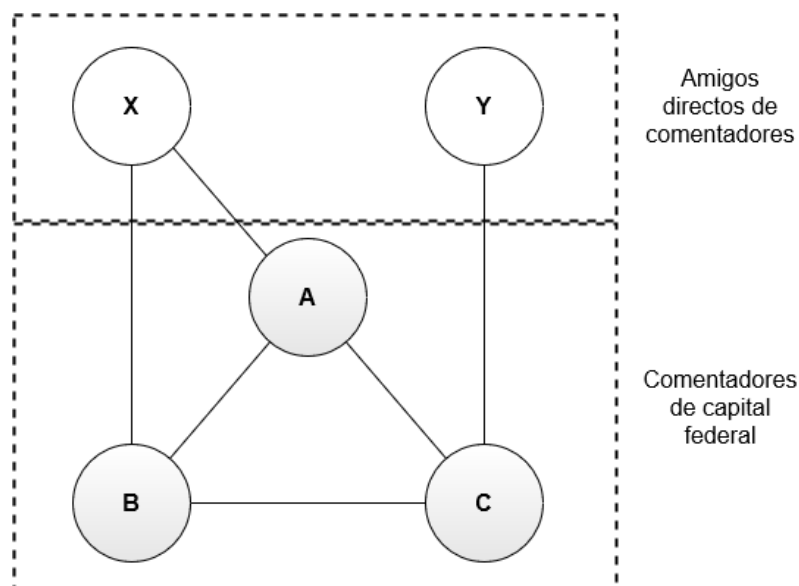


Figura 7.6: Diagrama de cálculo de distancia social entre amigas

A modo de ejemplo, para la red de comentadores de CABA, la relación entre los nodos A-C y B-C poseen solo un nodo en común, mientras que la relación A-B posee solo un nodo en común (C) si se considera la red base, pero dos nodos en común (C y X) si se consideran sus amigos directos que no pertenecen a la red base.

En la tabla 7.4 se calcularon las medidas de tendencia central para amigos en común tanto para GO como GE.

	Componente principal de CABA (GO)			Componente extendido CABA (GE)		
Medida	Amigos en común	Amigos totales de ambos nodos	Proporción de amigos totales y amigos en común	Amigos en común	Amigos totales de ambos nodos	Proporción de amigos totales y amigos en común
media	5,67	179,40	3,60%	10,80	385,03	2,68%
desvío estándar	13,40	167,00	5,20%	23,05	324,60	3,90%
mínimo	0	3	0,00%	0	3	0,00%
1er Cuartil	0	41	0,00%	1	135	0,26%
Mediana	1	112	1,50%	3	292	1,35%
3er Cuartil	5	283	5,20%	11	535	3,51%
máximo	499	1007	49,80%	707	2329	49,92%

Tabla 7.4: Amigos en común y totales por relación de amistad para GO y GE

Observamos que el promedio de amigas en común para la red base de comentadores de CABA es de 5,67 mientras que la mediana es de apenas 1 persona en común, y representa proporcionalmente el 3,6% del total de amigas de cada relación. Si incluimos a sus amigos directos que no son comentadores de capital federal, la proporción de amigos en común que tienen dos usuarios amigos cae a tan solo el 2,68%.

Adicionalmente a estas métricas de amigos en común, también se calculó el coeficiente de Jaccard, el cual difiere levemente del cálculo anteriormente realizado. Mientras que, en el anterior cálculo, la sumatoria de amigos total, era la sumatoria de todos los nodos vecinos de los dos nodos considerados en la relación, en el coeficiente de Jaccard solo se contabilizan los nodos únicos.

El coeficiente de Jaccard para los nodos u y v se define como:

$$\frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$$

en donde $\Gamma(u)$ significa el conjunto de nodos amigos de u .

En el siguiente grafo de ejemplo el cálculo de Jaccard entre los nodos A y C se definen de la siguiente manera:

$$\text{jacc_coeff}(A, C) = \frac{|\{B, D\}|}{|\{B, D, E, F\}|} = \frac{2}{4} = \frac{1}{2}$$

Coefficiente de Jaccard

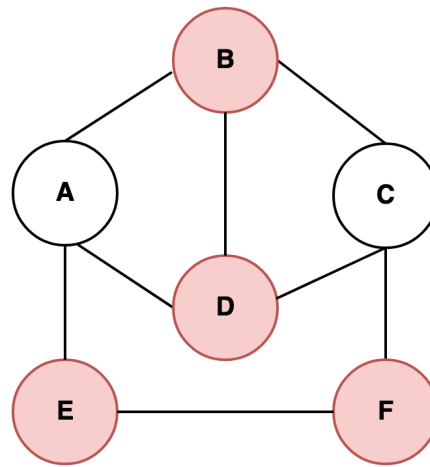


Figura 7.7: Cálculo de coef. de Jaccard entre nodos amigos

Adicionalmente se observó que las proporciones de amigas en común que tienen dos personas amigas del grafo de CABA varía según los idiomas de las personas que componen la relación. La proporción asciende levemente según el grupo de amigos que hablan español o distintos lenguajes entre sí (3,60%) respecto a aquellos pares de amigos que hablan solamente idioma extranjero (3,9%). Si bien ya se ha encontrado una diferencia significativa en la distancia promedio para grupos de hispanos y no hispanos, en tanto amigos en común, no difieren significativamente.

Finalmente se calculó la cantidad de amigas en común entre nodos que no eran amigas, pero este porcentaje promedio fue casi cero, ya que como hemos visto la densidad de la red es baja. Este punto reforzó la hipótesis de que la cantidad de amigas en común que tienen dos nodos aleatorios del grafo aumenta según se sea amigo.

Posteriormente se procedió a analizar correlaciones entre las distancias geográficas que tienen los 229224 amigos de la red social y las medidas descritas de “distancia

social” de manera de comprobar si hay una relación entre la distancia geográfica y todas las métricas de distancia social calculadas.

Se ha demostrado la distribución de relaciones triádicas en reciprocidad con la distancia geográfica en otras redes como Twitter [76], sin embargo en este estudio no se ha podido encontrar alguna relación significativa entre los indicadores de distancia social entre dos nodos amigos y su distancia geográfica.

Diferentes sub grupos fueron considerados, como por ejemplo solo personas con más de un comentario, distintos lenguajes, y otras variables calculadas *ad hoc* como por ejemplo cantidad de comentarios en común que tienen dos usuarios en base a la localización de los mismos.

Para el caso de utilizar el coeficiente de Jaccard, si bien este coeficiente le da más relevancia a los amigos en común entre dos nodos, tampoco se ha observado ninguna relación significativa entre distancia social y geográfica.

7.5 Probabilidad de Amistad

En diferentes estudios se ha encontrado que la probabilidad de los vínculos es más propicia a existir entre usuarios cercanos geográficamente que usuarios distantes. La misma sigue aproximadamente a una distribución de potencias, con diferentes valores de alfa [68] para diferentes tipos de redes sociales, tales como mensajes de correo electrónico, vínculos de Facebook, etc.

Para entender si esta propiedad similar se encuentra presente en la red social analizada, se puso a prueba si la distribución de distancia entre los centroides de las personas que tienen vínculos entre sí sigue la distribución de potencias.

Se encontró que la misma no sigue una distribución de potencias. En la figura 7.4, se observa que la distribución se asemeja más a una distribución de Poisson que de potencias, en donde la distancia media es de 3,73 km. Cuando se ajusta la distribución a la ley de potencias, Figura 7.8, nos encontramos con una forma que dista lejos de ser una distribución lineal como la esperada y explicada anteriormente en el capítulo de distribución de potencias.

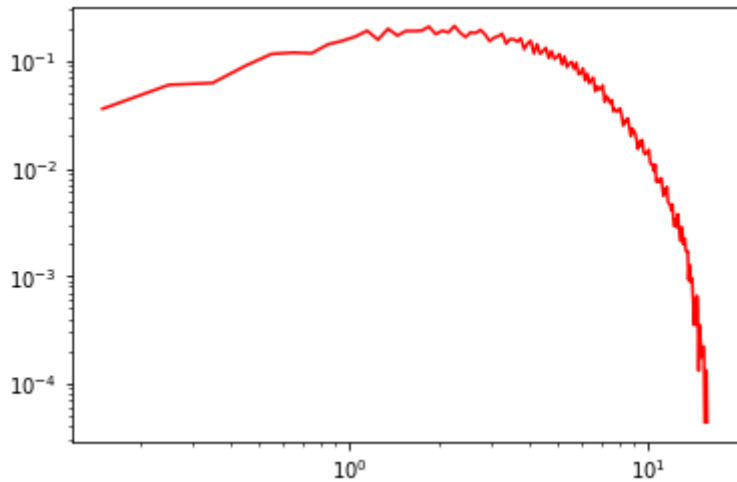


Figura 7.8: Función de densidad de probabilidad de potencias para distancia en km entre centroides de amigas (eje x)

Por el contrario, al comparar el ajuste de la distribución entre una función exponencial y de potencias, utilizando la razón de verosimilitud para identificar cuál de las dos distribuciones se ajustaban mejor, esta se asemeja más a la exponencial. Tal como lo planteado en la introducción de este capítulo, la evidencia actual sobre una distribución de potencias y el factor explicativo de la distancia geográfica para los vínculos sociales, es una discusión vigente.

Particularmente en el caso de Foursquare hay evidencia de tal distribución solamente en rangos de distancias mucho mayores ($>1000\text{km}$), que exceden ampliamente el máximo de 17km de área de CABA analizados en esta tesis. Así mismo en otros estudios que comparan distintas redes sociales geolocalizadas, se observa que Foursquare presenta los usuarios que tienen vínculos más alejados entre sí [77].

Los lazos sociales de larga distancia en LBSN tienen una probabilidad de ocurrencia mayor que otras redes sociales, como las redes de teléfonos móviles. Las conversaciones telefónicas están mucho más afectadas por la distancia geográfica que en redes como Facebook y Foursquare [64].

7.6 Resultados y discusión sobre distancia social y geográfica entre amigos

A lo largo de este capítulo se ha podido encontrar evidencia de que las personas que tienen una amistad en la plataforma sociabilizan más cerca, en términos geográficos, que aquellas personas que no son amigas.

Sin embargo no se ha podido demostrar relación significativa entre distancia geográfica y social entre amigos. A grandes rasgos, puede haber dos razones para explicar por qué no se cumple esta última premisa, la primera debido a las características de los datos trabajados en esta tesis y limitar el armado de la red social a comentaristas en CABA, y otra razón debido a las características propias de

la plataforma, las cuales no son mutuamente excluyentes.

En relación a las razones metodológicas en la recolección de datos y de limitaciones de la información utilizada en este trabajo que podrían estar impidiendo encontrar una relación significativa entre distancia social y geográfica, se encuentran:

- Como vimos en el capítulo 5, la proporción entre check-ins y comentarios para las personas que dejaron comentarios en capital federal, es de apenas un 2%. Esto nos lleva a pensar que una gran proporción de usuarios no dejan comentarios, por lo que la selección de usuarios de CABA basada en comentarios no sería suficiente para describir la distancia social de las amistades.
- Como anteriormente fue visto, la dispersión de los comentarios de los usuarios era mayor a la distancia que estos tienen con sus amigos. El centroide, que representa el atributo geoespacial característico de cada usuario, no es un indicador tan representativo y único de cada nodo.
- Adicionalmente, como vimos en el capítulo 5, los integrantes del grafo comentadores de CABA no hacen un uso exclusivo de Foursquare en la ciudad, solamente un 40% de sus comentarios están ubicados en la misma, por lo que no es posible calcular con exactitud su centroide ni su cercanía social ni geográfica.
- La transitividad⁴⁰ del grafo, la cual es 0,0134, es relativamente baja, por lo que la probabilidad de tener amigos en común es muy baja, consecuentemente el indicador de distancia social basado en amigos en común no es suficientemente fiable para estimar el grado de amistad.
- Dejar comentarios en lugares compartidos con otros usuarios no está correlacionado con ninguna otra variable, ni distancia geográfica entre centroides ni cantidad de amigos en común. Estos puntos geográficos o lugares que conforman los centroides no representan necesariamente que la amistad se caracteriza por transitar los mismos espacios, y por ende los centroides no son tan buen indicador de la distancia entre amigos. Los lazos de amistad no necesariamente están en base a actividad geolocalizada ya que aproximadamente el 73% de las amistades de los nodos considerados son personas que no han dejado algún comentario en CABA⁴¹.
Si bien se demostró que la distancia entre amigas y no amigas es diferente,

⁴⁰ Las redes sociales tienden a tener un coeficiente de transitividad más alto que el que se observa en esta red.

⁴¹ Eso no excluye que parte de sus amistades sean personas que hacen check-in principalmente en lugares de la CABA, pero debido a las limitaciones de privacidad de Foursquare, no se ha podido saber la localización de estos check-ins

no significa que la distancia geográfica entre distintas amigas pueda ser cuantificada por las amigas en común.

Adicionalmente a estas razones que imposibilitan una medición adecuada de la métrica de distancia social en relación con la geográfica, también hay otras razones asociadas a las características de la plataforma analizada, entre las cuales se encuentran:

- Como se ha citado anteriormente, en algunos estudios se describe que las relaciones de amistad en Foursquare son de un rango geográfico mayor a lo que es considerado en este trabajo, por lo que las amistades consideradas en este trabajo no sean representativas para analizar la fuerza de un vínculo.
- En algunos estudios se ha encontrado que los movimientos geográficos de las personas en una distancia corta son asociados a la rutina, y no son impactadas por la estructura de la red social. Por el contrario, los viajes de larga distancia son más influenciados por los lazos sociales [78]. Debido a las limitaciones geográficas de este estudio, circunscritas a CABA, entonces no sería posible medir la distancia geográfica en relación con la fuerza de un vínculo.

Conclusiones y trabajos futuros

A continuación se presentan las conclusiones de esta tesis y los posibles trabajos futuros para ampliar los hallazgos obtenidos.

Conclusión

En este trabajo se hace uso del análisis de redes sociales (SNA) y otras técnicas propias de la disciplina para el estudio de las interacciones geoespaciales entre personas y lugares dentro del ámbito de la Ciudad Autónoma de Buenos Aires. Adicionalmente se describe una metodología para la recolección, procesamiento y análisis utilizando fuentes de datos disponibles públicamente en el estado y plataformas online.

En tanto características de la red social analizada, se ha encontrado los siguientes aspectos estructurales y propiedades:

- Se descarta que la ley de potencias sea la mejor distribución teórica que ajuste a la distribución de grados del grafo analizado, encontrando la distribución log-lineal como más adecuada, en consonancia con la literatura académica más reciente al respecto.
- Se comprueba que la red presenta un bajo grado de afinidad selectiva negativa, dentro de los parámetros que presentan otras redes sociales similares.
- De igual manera, se comprueba la estructura de mundo pequeño, coincidiendo con otros estudios.
- Por último, se encuentra una estructura de comunidad alta y se describen las características estructurales de las distintas comunidades presentes.

En tanto red social circunscripta a un espacio geográfico específico, se encuentran los siguientes aspectos:

- Se halla que los lugares representados en el medio online se concentran especialmente en zonas con alta densidad poblacional y tienen su correlato en la distribución de la riqueza a nivel de la ciudad.
- Se describen patrones temporales y características de lugares predominantes, comerciales o de entretenimiento, para cada comuna de la ciudad.

En términos de sociabilidad y características de la interacción entre las personas integrantes de la red social analizada, en este trabajo se ha mostrado que:

- Se ha podido generar evidencia de que las personas que tienen una amistad en la plataforma sociabilizan más cerca, en términos geográficos, que aquellas personas que no son amigas.
- No se ha podido encontrar alguna relación significativa entre indicadores de

distancia social, medido en términos de personas amigas en común, y distancia geográfica de sociabilización entre dos personas amigas.

- Se comprueba, basado en los registros en línea, que las personas siguen patrones de sociabilidad esperables según horarios socialmente determinados, como almuerzo, cena, días laborales y horarios recreativos durante el fin de semana según las características de las comunas.

- Personas presuntamente turistas, catalogadas como tales por los comentarios registrados, tienen más amistades y usan más intensamente la red social que las personas de habla hispana.

- Se encuentra evidencia de que las personas no hispano parlantes frecuentan sectores más concentrados geográficamente de la ciudad que personas hispano parlantes.

- Se encuentra que las mujeres publican mayor cantidad de comentarios por persona, lo que va en consonancia con otros estudios, pero publican menos fotos en comparación con personas que se identifican como de género masculino.

Trabajos futuros

Se han descrito aspectos metodológicos de recolección de datos públicos, características estructurales de la red social y de los lugares como también de la sociabilidad de las personas dentro de la ciudad. Aplicaciones en turismo, comercio, mercadotecnia, investigación académica y desarrollo de políticas gubernamentales pueden ser realizadas profundizando diferentes aspectos de este trabajo, con los datos ya recolectados o bien otras LBSN.

Los hallazgos presentados han demostrado ser valiosos y de difícil acceso mediante otros medios, por lo que seguir profundizando el estudio de las redes sociales geolocalizadas en CABA puede permitir el desarrollo de políticas públicas en transporte, turismo y urbanismo considerando la relación entre las personas, desplazamiento geográfico, patrones temporales y características de comercios y servicios en cada una de las comunas.

Se han identificado diferentes características de las personas y comunidades no hispanas que pueden ayudar a mejorar servicios públicos y privados orientados a turistas, tales como la promoción de lugares y áreas geográficas dentro de CABA que no son tan frecuentados por el turismo, de manera de potenciar económicamente sectores específicos de la ciudad.

En un futuro análisis es posible identificar qué rol cumplen las distintas comunidades en la difusión de lugares e identificar si las comunidades se anclan en partes específicas de la ciudad y hacen un uso diferenciado de la misma para acceder a bienes y servicios.

En términos de estudio y políticas de género, el entendimiento cualitativo del uso de

las redes sociales geolocalizadas por parte de mujeres y hombres, ayudaría a entender las causas de las diferencias en el uso de plataformas online.

Profundizar en la relación entre lazos sociales y geolocalización puede habilitar a la implementación de mejores algoritmos en plataformas online, publicidad digital o servicios públicos para mejorar la oferta de bienes y servicios. Por ejemplo, los mecanismos por el cual sugerir lugares en un servicio similar a Foursquare, pueden estar basados sobre las propiedades de la comunidad a la que pertenece o amigos del individuo, teniendo en cuenta las características de los lugares que frecuentan. Finalmente, la realización de este trabajo con datos públicos de internet demuestra el efecto potencial que puedan tener las políticas de datos abiertos de empresas y gobierno para investigadoras, tomadores de decisiones, explotación comercial y organizaciones civiles.

Bibliografía

- [1] Panzarasa, P., Opsahl, T. and Carley, K. M. (2009), Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. J. Am. Soc. Inf. Sci., 60: 911–932. doi: 10.1002/asi.21015
- [2] Miniwatts Marketing Group (2020a). Internet Usage and Population Statistics for South America, December 31, 2017. Recuperado el 2 de Septiembre del 2020 de: <http://www.internetworldstats.com/stats15.htm#south>
- [3] INDEC (2018). Acceso y uso de tecnologías de la información y la comunicación. EPH. Cuarto trimestre de 2019
- [4] Wellman, B. (2004). The three ages of internet studies: ten, five and zero years ago. New media & society, 6(1), 123-129.
- [5] Simmel, G. (2002). Cuestiones fundamentales de sociología. Edición en Español. Gedisa Editorial.
- [6] Smith, C. E. (2012). Checking in: A phenomenological study of active users of geolocation tagging services. Proceedings of the American Society for Information Science and Technology, 49(1), 1-5.
- [7] Location-based service (s.f.). En Wikipedia. Recuperado el 26 de diciembre de 2019 de http://en.wikipedia.org/wiki/Location-based_service
- [8] Triadic closure (s.f.). En Wikipedia. Recuperado el 26 de julio de 2021 de https://en.wikipedia.org/wiki/Triadic_closure
- [9] Zickuhr, K. (2012). *Three-quarters of smartphone owners use location-based services*. Pew Internet & American Life Project.
- [10] World Bank. Information, Communication Technologies, & infoDev (Program). (2012). Information and communications for development 2012: Maximizing mobile. World Bank Publications.
- [11] INDEC (2013). Encuesta Nacional sobre Acceso y Uso de Tecnologías de la Información y la Comunicación (ENTIC) - 2013. Buenos Aires: INDEC
- [12] Miniwatts Marketing Group (2020). Facebook users in the world. Recuperado el 2 de Septiembre del 2020 de: <https://www.internetworldstats.com/facebook.htm>
- [13] Foursquare (2020). What is Foursquare City Guide?. Recuperado el 18 de

Septiembre del 2020 de:

<https://support.foursquare.com/hc/en-us/articles/201015194-What-is-Foursquare-City-Guide->

[14] Talend (s.f.) Open Studio for Data Integration.

<https://www.talend.com/products/data-integration/data-integration-open-studio/>

[15] Gobierno de la Ciudad de Buenos Aires (2015). Manzanas de la Ciudad de Buenos Aires. Recuperado el 12 de Enero del 2015 de:

<http://data.buenosaires.gob.ar/dataset/manzanas>

[16] Hussain, A., & Keshavamurthy, B. N. (2019). Analyzing Online Location-Based Social Networks for Malicious User Detection. In *Recent Findings in Intelligent Computing Techniques* (pp. 463-471). Springer, Singapore.

[17] Vasconcelos, M. A., Ricci, S., Almeida, J., Benevenuto, F., & Almeida, V. (2012, February). Tips, dones and todos: uncovering user profiles in foursquare. In *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 653-662).

[18] Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In *2008 eighth ieee international conference on data mining* (pp. 413-422). IEEE.

[19] Nakatani, S. (2010). Language detection library for java, 2010. URL: <https://github.com/shuyo/language-detection>.

[20] Wilson, R. J. (1979). *Introduction to graph theory*. Pearson Education India.

[21] Marin, A., & Wellman, B. (2011). Social network analysis: An introduction. *The SAGE handbook of social network analysis*, 11, 25.

[22] Freeman, L. (2004). The development of social network analysis. *A Study in the Sociology of Science*, 1(687), 159-167.

[23] Arif, Tasleem. (2015). The Mathematics of Social Network Analysis: Metrics for Academic Social Networks. *International Journal of Computer Applications Technology and Research*. 4. 889-893. 10.7753/IJCATR0412.1003.

[24] Oliveira, M., & Gama, J. (2012). An overview of social network analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(2), 99-115.

[25] Estrada, E. (2016). When local and global clustering of networks diverge. *Linear Algebra and its Applications*, 488, 249-263.

[26] Oliveira, M., & Gama, J. (2012). An overview of social network analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(2), 99-115.

[27] Leskovec, J., Kleinberg, J., & Faloutsos, C. (2007). Graph evolution:

Densification and shrinking diameters. *ACM transactions on Knowledge Discovery from Data (TKDD)*, 1(1), 2-es.

[28] Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10), P10008.

[29] Carr, A. (2010). Foursquare's Business Chief on Revenue Plans, Google AdWords, and Why Marketers Shouldn't Delay on Geo-location. *FastCompany*. Recuperado el 2 de Noviembre del 2020 de: <http://www.fastcompany.com/1675626/>

[30] Chen, Y., Yang, Y., Hu, J., & Zhuang, C. (2016, March). Measurement and analysis of tips in foursquare. In *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)* (pp. 1-4). IEEE.

[31] Ye, M., Janowicz, K., Mülligann, C., & Lee, W. C. (2011, November). What you are is when you are: the temporal dimension of feature types in location-based social networks. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 102-111).

[32] Noulas, A., Scellato, S., Mascolo, C., & Pontil, M. (2011). An empirical study of geographic user activity patterns in foursquare. *ICWSM*, 11(70-573), 2.

[33] Marcos, M., Mera, G. S., & Di Virgilio, M. M. (2015). Contextos urbanos de la ciudad de Buenos Aires: una propuesta de clasificación de la ciudad según tipos de hábitat. *Papeles de población*, 21(84), 161-196.

[34] Sarraute, C., Lang, C., Ponienman, N. B., & Anapolsky, S. (2017). The city pulse of Buenos Aires. *arXiv preprint arXiv:1707.01032*.

[35] Hong, I. (2015). Spatial analysis of location-based social networks in seoul, korea. *Journal of Geographic Information System*, 7(03), 259.

[36] Zhang, J., & Luo, Y. (2017, March). Degree centrality, betweenness centrality, and closeness centrality in social network. In *2017 2nd International Conference on Modelling, Simulation and Applied Mathematics (MSAM2017)*. Atlantis Press.

[37] Zinoviev, D. (2018). *Complex network analysis in Python: Recognize-construct-visualize-analyze-interpret*. Pragmatic Bookshelf.

[38] Li, C., Wang, H., & Van Mieghem, P. (2012, May). Degree and principal eigenvectors in complex networks. In *International Conference on Research in Networking* (pp. 149-160). Springer, Berlin, Heidelberg.

[39] Valente, T. W., Coronges, K., Lakon, C., & Costenbader, E. (2008). How correlated are network centrality measures?. *Connections (Toronto, Ont.)*, 28(1), 16.

- [40] Freeman, L. C., Roeder, D., & Mulholland, R. R. (1979). Centrality in social networks: II. Experimental results. *Social networks*, 2(2), 119-141.
- [41] Agryzkov, T., Martí, P., Tortosa, L., & Vicent, J. F. (2017). Measuring urban activities using Foursquare data and network analysis: a case study of Murcia (Spain). *International Journal of Geographical Information Science*, 31(1), 100-121.
- [42] Ferreira, A. P. G., Silva, T. H., & Loureiro, A. A. F. (2015, November). Beyond sights: Large scale study of tourists' behavior using foursquare data. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)* (pp. 1117-1124). IEEE.
- [43] Newman, M. E. (2003). Mixing patterns in networks. *Physical review E*, 67(2), 026126.
- [44] Meghanathan, N. (2016). Assortativity Analysis of Real-World Network Graphs based on Centrality Metrics. *Computer and Information Science*, 9(3), 7-25.
- [45] Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007, October). Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement* (pp. 29-42).
- [46] Noldus, R., & Van Mieghem, P. (2015). Assortativity in complex networks. *Journal of Complex Networks*, 3(4), 507-542.
- [47] Peel, L., Delvenne, J. C., & Lambiotte, R. (2018). Multiscale mixing patterns in networks. *Proceedings of the National Academy of Sciences*, 115(16), 4057-4062.
- [48] Muchnik, L., Pei, S., Parra, L. C., Reis, S. D., Andrade Jr, J. S., Havlin, S., & Makse, H. A. (2013). Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Scientific reports*, 3(1), 1-8.
- [49] Broido, A. D., & Clauset, A. (2019). Scale-free networks are rare. *Nature communications*, 10(1), 1-10.
- [50] Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439), 509-512.
- [51] Oliveira, M., & Gama, J. (2012). An overview of social network analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*.
- [52] Easley, D., & Kleinberg, J. (2012). *Networks, crowds, and markets: Reasoning about a highly connected world*.
- [53] Alstott, J., Bullmore, E., & Plenz, D. (2014). powerlaw: a Python package for analysis of heavy-tailed distributions. *PloS one*, 9(1), e85777.

- [54] Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007, October). Measurement and analysis of online social networks. In Proceedings of the 7th ACM SIGCOMM conference on Internet measurement.
- [55] Li, H., Zhao, Y., & Zhang, N. (2013). Structural and Dynamic Analysis of Foursquare Network.
- [56] Malevergne, Y., Pisarenko, V., & Sornette, D. (2005). Empirical distributions of stock returns: between the stretched exponential and the power law?. *Quantitative Finance*, 5(4), 379-401.
- [57] Scellato, S., & Mascolo, C. (2011, April). Measuring user activity on an online location-based social network. In 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS) (pp. 918-923). IEEE.
- [58] Wei, W., Zhu, X., & Li, Q. (2014, April). LBSNSim: Analyzing and modeling location-based social networks. In IEEE INFOCOM 2014-IEEE Conference on Computer Communications (pp. 1680-1688). IEEE.
- [59] Small-world experiment (s.f.). En Wikipedia. Recuperado el 26 de julio de 2021 de https://en.wikipedia.org/wiki/Small-world_experiment
- [60] Backstrom, L., Boldi, P., Rosa, M., Ugander, J., & Vigna, S. (2012, June). Four degrees of separation. In Proceedings of the 4th Annual ACM Web Science Conference (pp. 33-42).
- [61] Easley, D., & Kleinberg, J. (2012). Networks, crowds, and markets: Reasoning about a highly connected world. *Significance*, 9, 43-44.
- [62] Ugander, J., Karrer, B., Backstrom, L., & Marlow, C. (2011). The anatomy of the facebook social graph. arXiv preprint arXiv:1111.4503.
- [63] Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *nature*, 393(6684), 440-442.
- [64] Scellato, S., Noulas, A., Lambiotte, R., & Mascolo, C. (2011). Socio-spatial properties of online location-based social networks. *ICWSM*, 11, 329-336.
- [65] Leskovec, J., & Horvitz, E. (2008, April). Planetary-scale views on a large instant-messaging network. In Proceedings of the 17th international conference on World Wide Web (pp. 915-924).
- [66] Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12), 7821-7826.
- [67] Carrasco, J. A., & Miller, E. J. (2009). The social dimension in action: A multilevel, personal networks model of social activity frequency between individuals.

Transportation Research Part A: Policy and Practice, 43(1), 90-104.

[68] Levy, M., & Goldenberg, J. (2014). The gravitational law of social interaction. *Physica A: Statistical Mechanics and its Applications*, 393, 418-426.

[69] Backstrom, L., Sun, E., & Marlow, C. (2010, April). Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web* (pp. 61-70).

[70] Wahid-UI-Ashraf, A., Budka, M., & Musial, K. (2019). How to predict social relationships—Physics-inspired approach to link prediction. *Physica A: Statistical Mechanics and its Applications*, 523, 1110-1129.

[71] Mahmoudi, A., Yaakub, M. R., & Bakar, A. A. (2019). The relationship between online social network ties and user attributes. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(3), 1-15.

[72] Ahmad, I., Akhtar, M. U., Noor, S., & Shahnaz, A. (2020). Missing Link prediction using common neighbor and centrality based parameterized Algorithm. *Scientific Reports*, 10(1), 1-9.

[73] Pontes, T., Vasconcelos, M., Almeida, J., Kumaraguru, P., & Almeida, V. (2012, September). We know where you live: privacy characterization of foursquare behavior. In *Proceedings of the 2012 ACM conference on ubiquitous computing* (pp. 898-905).

[74] Kulshrestha, J., Kooti, F., Nikraves, A., & Gummadi, K. P. (2012, May). Geographic dissection of the Twitter network. In *Sixth international AAAI conference on weblogs and social media*.

[75] Ahmad, W., Riaz, A., Johnson, H., & Lavesson, N. (2010, September). Predicting friendship intensity in online social networks. In *21st International Tyrrhenian Workshop on Digital Communications*.

[76] Rout, D., Bontcheva, K., Preotiu-Pietro, D., & Cohn, T. (2013, May). Where's@wally? a classification approach to geolocating users based on their social ties. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media* (pp. 11-20).

[77] Haldar, N. A. H., Li, J., Reynolds, M., Sellis, T., & Yu, J. X. (2019). Location prediction in large-scale social networks: an in-depth benchmarking study. *The VLDB Journal*, 28(5), 623-648.

[78] Cho, E., Myers, S. A., & Leskovec, J. (2011, August). Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1082-1090).