Análisis de una red social geo-localizada

Favio Di Ciocco, Daniel Pinto, Diego Espejo

13 de diciembre de 2018

Resumen

En este trabajo se analizó una red social llamada Brightkite, fundada en el 2007, cuya principal característica que la destacaba era el uso de "Check-Ins" a la hora de postear contenidos o revisar el contenido de otros usuarios. Los Check-Ins son un registro de la ubicación del usuario de la red.

Los datos obtenidos contenían por un lado la red de amigos y por el otro el registro de los Check-Ins realizados por cada uno de los usuarios para un grupo limitado del total de usuarios de la red.

Se analizó:

1. La distribución de Grado y de Cantidad de lugares visitados de los usuarios
2. La existencia de una relación entre la cantidad de salidas y la cantidad de amigos de los usuarios
3. La existencia de una relación entre la Distancia máxima viajada y la cantidad de lugares visitados por los usuarios.
4. Si los amigos reportados en la red de amigos frecuentaban los mismo lugares utilizando tres métodos diferentes.
5. Si se pueden identificar los lugares del mapa frecuentados por las mismas personas a lo largo de 6 meses, dividido en períodos de semanas.

1. Objetivos

El objetivo de este trabajo fue estudiar una red social geo-localizada llamada Brightkite, para a partir de ella responder algunas preguntas planteadas tales como: *¿La gente que tiene más amigos es la que más viaja?*, *¿La gente que sale más es la que viaja más lejos?*, *¿Los amigos frecuentan los mismo lugares?*, *¿Podemos diferenciar amigos de compañeros de trabajo?*.

2. Brightkite

Brightkite era una red social creada en el 2007 cuya principal característica era el uso de Google maps para la geolocalización de sus usuarios. Los usuarios eran capaces de realizar "Check-Ins" en diversos lugares a través de mensajes de texto o posteando imágenes en la red, y de esta manera tenían acceso a conocer quienes se hallaban cerca del lugar en el cuál realizaron Check-in, así como de las personas que anteriormente habían visitado ese Check-In.

Figura 1: Página de inicio de la red Brightkite

Esta función era la característica que destacaba este servicio, que por otra parte contaba con las mismas funciones que otras redes sociales, como la de armar grupos de amigos, subir contenido a la red y revisar el contenido de amigos, como se puede apreciar en la figura 1

Para Diciembre de 2011 esta red fue retirada de la tienda de aplicaciones y en su página simplemente aparecía un mensaje de despedida.

3. Datos Relevados

Para poder trabajar sobre esta red se descargaron los datos que se hallan en la página de *Stanford Network Analysis Project*[1]. Estos datos contenían por un lado una red de amigos para un total de 58228 usuarios con una cantidad total de enlaces de 214078, mientras que por otro lado contenía la lista de Check-Ins realizados por un total de 50687 usuarios en un intervalo de tiempo de dos años y medio, que va desde Abril del 2008 hasta Octubre del 2010. En la Tabla 1 se puede ver la forma en que los datos de los Check-Ins venían presentados.

Usuario	Tiempo del Check-In	Latitud	Longitud	Id de la ubicación
0	2008-12-03T21:09:14Z	39.633321	-105.317215	ee8b88dea22411
0	2008-11-30T22:30:12Z	39.633321	-105.317215	ee8b88dea22411
0	2008-11-28T17:55:04Z	-13.158333	-72.531389	e6e86be2a22411
0	2008-11-26T17:08:25Z	39.633321	-105.317215	ee8b88dea22411
1	2008-08-14T21:23:55Z	41.257924	-95.938081	4c2af967eb5df8
1	2008-08-14T07:09:38Z	41.257924	-95.938081	4c2af967eb5df8

Tabla 1: Ejemplos de Check-Ins para dos usuarios, donde se puede ver que cada Check-In contiene información del momento en que se realizó, la ubicación por Latitud y Longitud y una identificación del lugar

Como se puede observar de la tabla 1, se contaba no sólo con la ubicación espacial sino con la temporal para cada registro realizado.

4. Análisis de la red

4.1. Distribución de grado y de cantidad de lugares visitados

Se buscó caracterizar el comportamiento de la red en las variables grado y la "variable cantidad de lugares visitados". Esta segunda variable representa la cantidad de lugares distintos visitados, representados por la cantidad total de ID's distintos de cada usuario.

En la figura 2 se puede observar en verde la distribución de grado y en naranja la distribución de cantidad de lugares visitados. Estas distribuciones parecían comportarse siguiendo una ley de potencias, por lo que se les hizo un ajuste a cada uno, graficándose los respectivos ajustes en rojo y celeste.

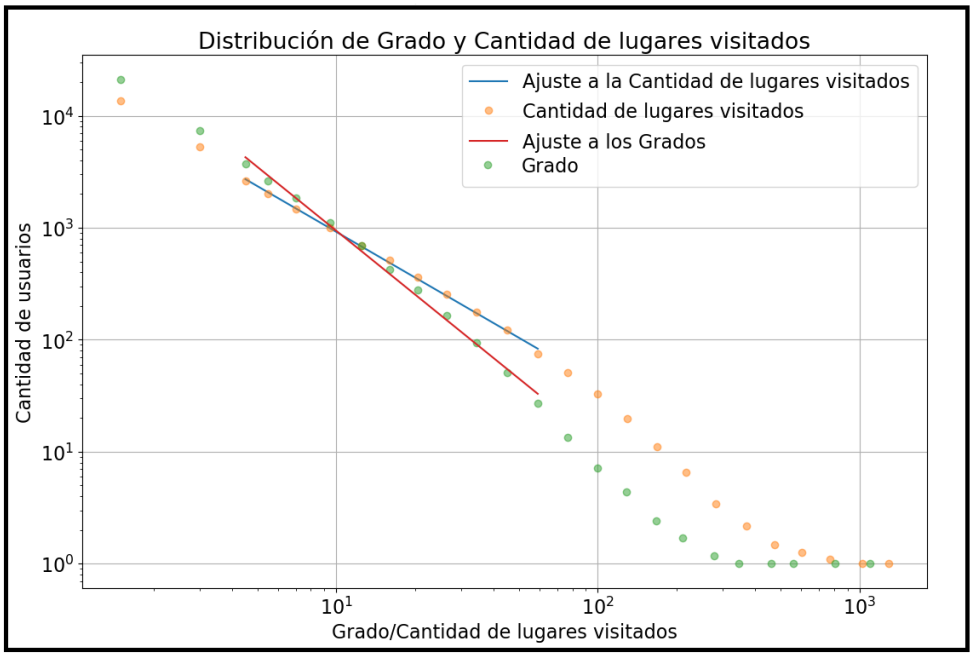

Figura 2: Distribución de Cantidad de Lugares visitados y del Grado de los usuarios

Para la distribución de grados se obtuvo un exponente de *-1.89*, mientras que para la cantidad de lugares visitados se obtuvo un exponente de *-1.35*.

4.2. ¿La gente que tiene más amigos es la que más viaja?

Dadas las dos distribuciones presentadas en la figura 2 se buscó analizar si existía alguna relación entre la çantidad de salidas de un usuarioz la cantidad de amigos que este tenía, donde la cantidad de salidas está representada por la cantidad total de Check-Ins realizados. En la figura 3 se relacionó la cantidad de Check-Ins totales de cada usuario con su grado y se graficó.

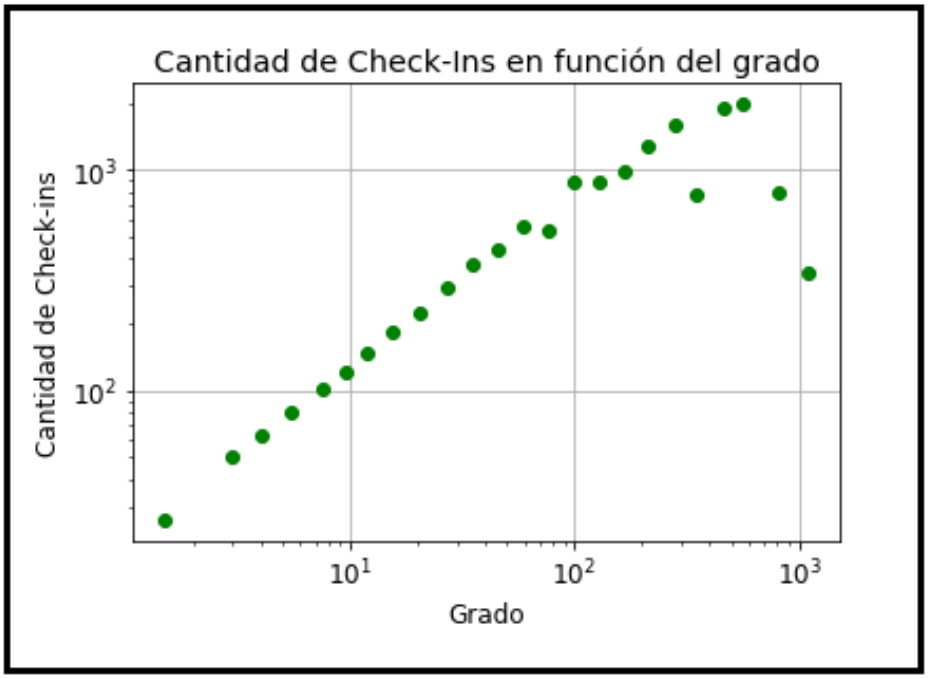

Figura 3: Cantidad de Check-Ins totales en función del grado

Como se puede observar de la figura, existe una relación creciente entre el grado y la cantidad de Check-Ins totales, con la excepción de tres puntos en la región de la derecha del gráfico. Esta caída se debe al hecho de que la cantidad de usuarios con un grado tan alto es muy baja, por lo que no hay otros usuarios con los cuales promediarlo y obtener así un comportamiento del grupo de usuarios con ese grado, sino que lo que se obtiene graficado es el comportamiento del único usuario con ese grado.

La relación mostrada entre la cantidad de Check-Ins y el Grado no demuestra que las personas que más salen se deba exclusivamente a que visiten a los amigos, pero si nos permite darnos cuenta de una conclusión razonable. A medida que una persona utilice más la red social, más Check-Ins registrará y agregará una mayor cantidad de amigos a su lista de amigos. Por tanto, podemos interpretar de la misma manera que la gente que tiene pocos amigos y pocos Check-Ins se debe a que simplemente utilizaron el servicio alguna vez por probar y luego lo dejaron.

4.3. ¿La gente que más sale es la que viaja más lejos?

Esta pregunta busca analizar si existe algún punto de saturación de lugares para visitar en el entorno del usuario. Es decir que a medida que la persona registra cada vez más lugares distintos, necesariamente comienza a realizar viajes más largos ya que no le quedan lugares cercanos de su interés que añadir. Para esto se calculó la distancia máxima recorrida por cada usuario, sin tomar en cuenta la distancia promedio de sus viajes ni la cantidad de veces que realizó el viaje máximo, y se lo comparó con la cantidad de lugares distintos visitados por cada usuario. Esto se graficó y se muestra en la figura 4.

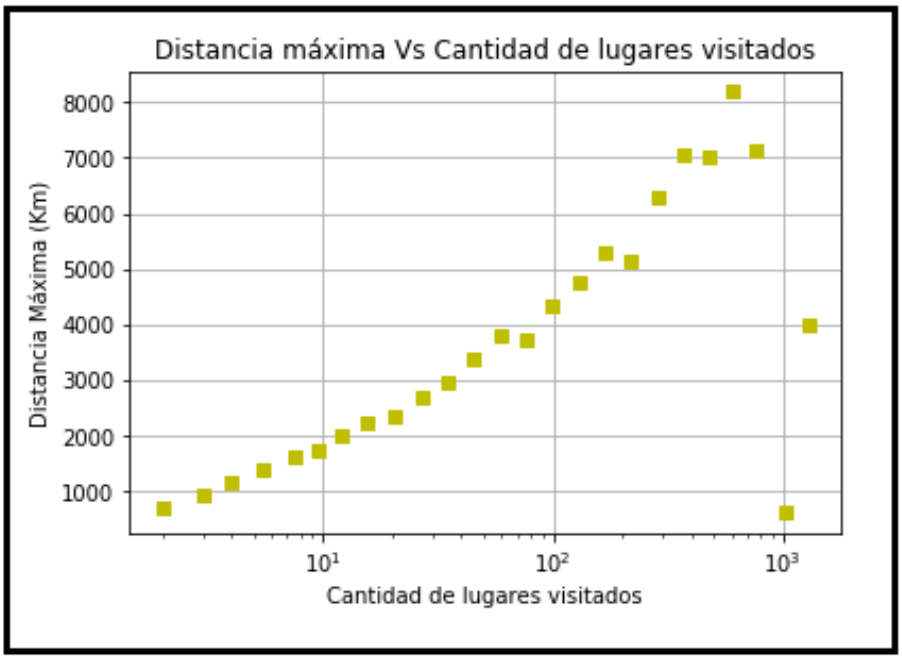

Figura 4: Distancia máxima recorrida en función de la cantidad de lugares distintos visitados

Se puede observar en este caso también que hay una relación creciente entre las distancias máximas recorridas por los usuarios y la cantidad de lugares distintos visitados por ellos, y además que la diferencia es sustancial, aumentando el valor desde un mínimo aproximado de 800 km hasta un máximo aproximado de 8100 km.

5. Estudio de Similaridad entre la red de amigos y redes creadas a partir de los registros de Check-Ins

Este análisis surgió como un intento de responder la pregunta: ¿Los amigos frecuentan los mismos lugares? Para esto utilizamos tres métodos distintos:

1. Utilizamos una red proyectada a partir de la red bipartita pesada de usuarios y lugares visitados.
2. Construimos una red entre usuarios según un criterio de Similaridad.
3. Construimos una red entre usuarios vinculándolos cada vez que coincidieran según un criterio tanto espacial como temporal.

En los tres casos lo que se hizo fue calcular la cantidad de enlaces que aparecían en ambas redes, y dependiendo del caso se normalizaba utilizando el tamaño de la red original o el de la red creada.

5.1. Red Bipartita y red proyectada

Se construyó una red bipartita en la cual los dos conjuntos de nodos distintos eran los usuarios y los lugares visitados, los cuales podían diferenciarse utilizando el ID asignado a cada uno. De esta manera se obtuvo una red pesada que vinculaba usuarios con los lugares que visitaron y con un peso igual a la cantidad de veces que visitaron cada lugar.

Esta red luego se proyectó sobre los usuarios, de manera de que cada usuario se vincule con todo otro usuario con el que haya compartido lugares a los que fue, sin importar la cantidad de veces que hayan ido. Se obtuvo una red no dirigida sin peso. Nuestra suposición sobre esta red es que si en verdad los amigos

visitan los mismos lugares, independientemente del tiempo o de la cantidad de veces que visitan los lugares, entonces la red proyectada debe estar sobreestimando las relaciones de amistad.

En cantidad de enlaces, la red proyectada era más de diez veces mayor a la red original, alcanzando una cantidad de 7760000 enlaces aproximadamente, mientras que la red de amigos tenía unos 220000 aproximadamente. Nuestra primera intuición fue que la mayor parte de los enlaces de la red original estarían en la red proyectada, por lo que se calculó la intersección de enlaces, y luego se lo normalizó utilizando la cantidad de enlaces de la primer red, ya que lo que se quería obtener era la fracción de enlaces de la primer red que aparecían en la segunda.

Al hacer este cálculo, se obtuvo que sólo el 27 % de los enlaces de la primer red aparecen en la segunda, lo cual es un número muy bajo. Es importante destacar que el máximo posible de enlaces obtenible en la segunda red era 91 % y no 100 %. Esto se debe a que en la red de amigos aparecen ciertos usuarios que nunca hicieron Check-Ins, y como la segunda red está conectada en función de los lugares a los que fueron los usuarios, aquellos que nunca hicieron Check-Ins no estarían conectados a nadie. La cantidad de enlaces de la red de amigos producidos por estos usuarios sin Check-In, representa el 9 %.

Estos resultados nos permiten mostrar que únicamente la coincidencia espacial de los usuarios no permite reconstruir los enlaces que realmente existen en la red de amigos.

5.2. Criterio de Similaridad

En este caso el enfoque es al revés del anterior. Se busca estudiar si se puede crear una red que sea más chica que la red original, pero cuyos enlaces sean en una gran proporción únicamente enlaces existentes en la red real de amigos. Para esto primero se define la Similaridad entre usuarios como la intersección del conjunto de lugares visitados de cada uno, dividido la unión de estos conjuntos. Esto se representa en la ecuación 1

$$S_{(u,v)} = \frac{Lug_{(u)} \cap Lug_{(v)}}{Lug_{(u)} \cup Lug_{(v)}} \tag{1}$$

Donde $S_{(u,v)}$ es el valor de Similaridad asignado al enlace entre el usuario u y el v, que varía entre 0 y 1 y $Lug_{(u)}$ es el conjunto de lugares visitados por el usuario u.

Para calcular la intersección de lugares se considera la cantidad total de Check-Ins de cada nodo, y para cada lugar en el cual los usuarios considerados coincidieron, el valor de la intersección será el mínimo de las visitas que ambos hicieron a ese lugar. La unión en cambio será considerar el máximo de las visitas que ambos hicieron a cada lugar considerado.

En la figura 5 se esquematiza la forma de calcular la similaridad para dos usuarios que sólo visitan dos lugares.

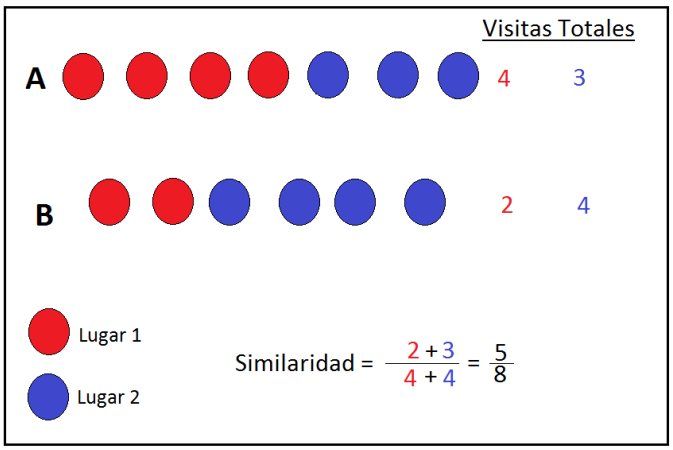

Figura 5: Cálculo de Similaridad para dos usuarios que visitan dos lugares cada uno

Este criterio de Similaridad asigna a cada enlace entre usuarios un valor que refleja que tan parecidos son los hábitos de los usuarios, ya que toma en consideración la frecuencia con la que visitan cada lugar.

5.3. Red de Similaridad

Esta red posee enlaces pesados entre los usuarios, donde el peso es el valor de similaridad que hay entre los usuarios enlazados.

A partir de esta red lo que se hizo fue barrer todos los valores de similaridad desde el mínimo hasta al máximo y para cada valor de similaridad utilizado se removieron los enlaces cuya similaridad fuera menor de la utilizada y se calculó la intersección de enlaces entre la red de Similaridad y la red de amigos, normalizando con la cantidad de enlaces de la red de Similaridad. En la figura 6 se muestran los datos obtenidos utilizando este criterio para la red considerando sólamente los Check-Ins de los primeros 9 meses. No se utilizó la totalidad del tiempo de la red debido al gran tiempo de calculo requerido.

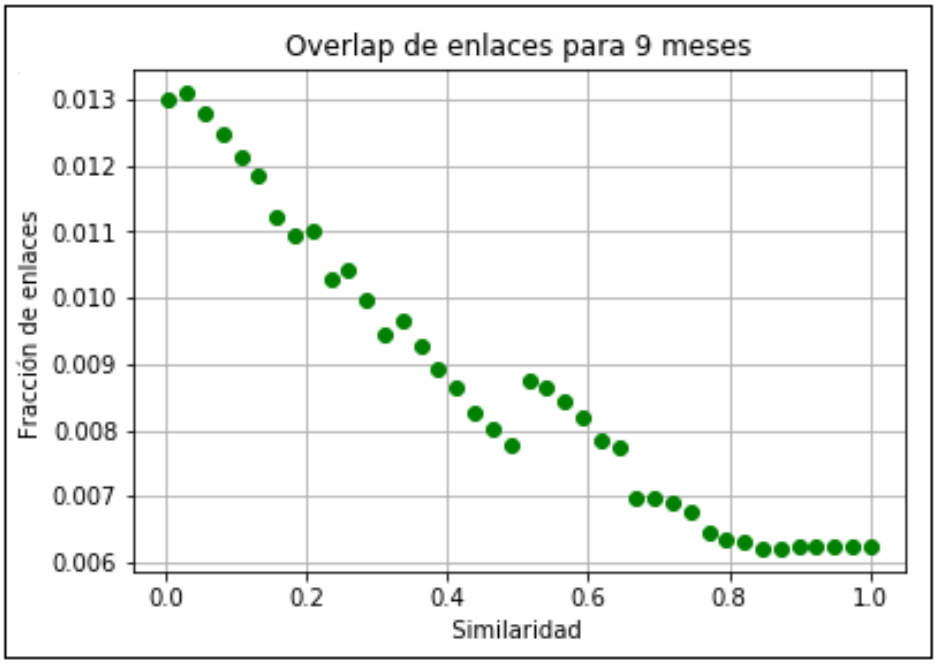

Figura 6: Fracción de enlaces de la red de amigos que aparecen en la red de Similaridad a medida que se vuelve más restrictivo el criterio de Similaridad

A partir de la figura se puede observar que a medida que el criterio de similaridad se vuelve más estricto y se remueven enlaces de la red de Similaridad, la fracción de enlaces de la segunda red que se pueden hallar en la red de amigos disminuye. Esto nos marca que el criterio de similaridad empleado no representa correctamente la amistad entre los usuarios, y que por tanto, la amistad entre los usuarios no está dada simplemente por una coincidencia de hábitos.

5.4. Criterio espacio-temporal

Habiendo visto que el criterio de similaridad usado anteriormente no nos permitió rearmar la red de amistad a partir de los *check-ins* se decidió introducir al análisis la variable temporal. Para ello se hizo lo siguiente: primero se armé un grafo vacío, que llamaremos *grafo proyectado*; luego, tomando nodos de a pares, se revisó aquellos *check-ins* de cada nodo que estuvieran a una distancia menor a diez metros, luego si esa condición se cumplía se pedía además que el tiempo entre ambos *check-ins* fuera menor a media hora, finalmente si esa última condición se cumplía, se añadía un enlace entre ambos nodos al grafo proyectado. Siguiendo este procedimiento para cada par de nodos de nuestra lista de *check-ins* se pudo armar el nuevo grafo. Hecho eso se volvió a medir la fracción de enlaces en una y otra red como en el caso del criterio

anterior, al mismo tiempo que se calculó el número de nodos y de enlaces para poder analizar las relaciones entre ambos grafos. Este procedimiento se realizó para los intervalos de tiempo de 3, 6, 9 y 12 meses desde que se comenzaron a relevar los Check-Ins y los resultados son expuestos en la tabla 2

	Grafo Original		Grafo Proyectado		
	Nodos	Enlaces	Nodos	Enlaces	Fracción
3 Meses	9028	39984	2083	2439	0,6902
6 Meses	16512	67254	4495	6336	0,7054
9 Meses	23804	92318	9138	14731	0,7306
12 Meses	—	—	—	—	0,7377

Tabla 2: Tabla de fracción de enlaces para ambos grafos.

Puede verse como la diferencia entre la cantidad de nodos y enlaces entre un grafo y otro es muy grande y esto aumenta a medida que consideramos mayor el intervalo de tiempo. Es por esto que la normalización de la fracción de enlaces la hacemos respecto de la cantidad de enlaces en el grafo proyectado. Así se puede ver que alrededor el 70 % de los enlaces que se obtuvieron en el grafo proyectado según el criterio empleado se repiten en la red de amistad original, mientras que si se hubieran normalizado de acuerdo a la cantidad de enlaces de la red original ese porcentaje hubiera sido mucho menor teniendo en cuenta la gran cantidad de enlaces que tiene el grafo original respecto del grafo proyectado. De todas formas no podemos concluir que sea un buen criterio el usado dado la gran cantidad de nodos y enlaces que quedan fuera del nuevo grafo proyectado; pese al 70 % de enlaces obtenidos, este número se vuelve menor al 10 % si se considera la red original.

6. Análisis geográfico de la red a partir de los registros de Check-Ins

Nos preguntamos si, a partir de la información de los check-ins, podemos identificar qué lugares en el mapa que sean frecuentados por las mismas personas, conforme pasa el tiempo.

Comenzamos por clasificar los lugares en tres categorías:

- Casas: lugares que suelen tener sólo un usuario que hace check-ins.
- Trabajos: lugares que suelen tener muchos usuarios que hacen check-ins frecuentemente.
- Otros (esparcimiento): lugares que suelen tener varios usuarios esporádicos.

El criterio para decir cuáles son las casas y trabajos fue el siguiente: primero, fijarse los dos lugares más frecuentados por cada usuario; segundo, fijarse cuáles de esos lugares se repiten para una cantidad determinada de usuarios. Aquellos lugares que se repitan serán los trabajos. En otras palabras, si un lugar se encuentra entre los dos más visitados para, por ejemplo, 5 o más usuarios, diremos que se trata de un trabajo. Acá asumimos que los usuarios suelen hacer check-ins en sus casas y en sus trabajos, y que tienen 1 sólo trabajo.

De acá en adelante, los análisis que haremos sólo tendrán en cuenta los lugares de esparcimiento.

Se tomaron los registros de los primeros 6 meses, cerca de la ciudad de Nueva York (NY), ya que ahí se registró la mayor densidad espacial de check-ins. Se tomaron los datos de todos los lugares dentro de un radio de 200 km desde el centro de NY, que fue arbitrariamente elegido, y que se corresponden con las coordenadas del Central Park: (40.7829,-73.9654)

Se filtraron los lugares visitados para tener únicamente los lugares de esparcimiento, según el criterio mencionado anteriormente. Además, se eliminaron todos los lugares que sólo hayan sido visitados por 1 usuario, ya que se consideró que eran hogares.

Dividimos el período de aproximadamente 6 meses (desde el 21-03-2008 hasta el 27-08-2008), en 20 intervalos de 1 semana.

Para cada semana, se armó un grafo pesado, cuyos nodos eran los lugares de esparcimiento, y el peso de los enlaces entre dos lugares era dado por la cantidad de usuarios que durante esa semana visitaron al menos una vez esos dos lugares. Así, mediante el algoritmo Infomap, obtuvimos particiones del grafo en clusters de lugares, para cada semana. Hay que observar que por lo general la cantidad de componentes conexas era grande, tal que en promedio había 3 lugares en cada componente conexa. Por eso, también los clusters obtenidos mediante Infomap contenían pocos lugares. Se analizó si, en cada cluster, la distancia geográfica media entre sus lugares era menor que la distancia geográfica media a los lugares de otro cluster. En la tabla 3, se muestra esta distancia media entre los 8 clusters. Se ve que la distancia media entre lugares de un mismo cluster no siempre es la más pequeña, comparandola con las distancias medias a otros clusters.

	0	1	2	3	4	5	6	7
0	4.37	3.85	2.58	141.93	131.56	128.44	4.8	142.5
1	3.85	4.63	3.67	141.32	130.94	127.82	4.56	141.89
2	2.58	3.67	0.67	141.4	130.71	127.58	6.14	141.99
3	141.93	141.32	141.4	11.66	21.62	23.2	143.69	41.52
4	131.56	130.94	130.71	21.62	1.64	3.22	133.74	43.28
5	128.44	127.82	127.58	23.2	3.22	0.5	130.61	42.55
6	4.8	4.56	6.14	143.69	133.74	130.61	4.53	144.19
7	142.5	141.89	141.99	41.52	43.28	42.55	144.19	81.43

Tabla 3: Distancias medias entre los clusters de la quinta semana. En cada fila se resaltaron los valores mínimos.

Una situación similar ocurre en todas las semanas. Aunque las auto-distancias medias, que serían las de la diagonal de la tabla 3, son chicas en comparación al resto, en varios casos no son las más chicas, lo que nos dice que hay lugares lejanos en el mapa, que tienen usuarios en comun. Es decir, que algunos lugares compartan usuarios en una semana no significa que todos los lugares estén cerca unos de otros.

También se estudió cómo varía la cantidad media de lugares por cluster a lo largo de las semanas.

$$r = \frac{\#\text{Lugares}}{\#\text{Clusters}}$$

Un valor alto significaría que hay pocos clusters con muchos lugares cada uno, lo que a su vez significaría que muchos lugares comparten usuarios. Un valor bajo significaría que las personas circularon por los mismos lugares, rara vez yendo a otros lados.

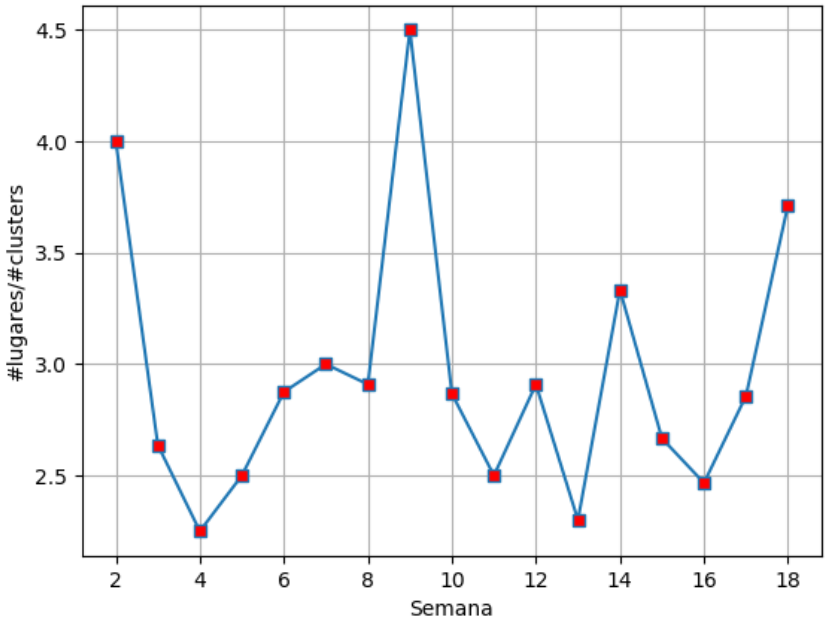

Figura 7: Cantidad media de lugares por cluster en función de la semana.

En la figura 7 se ven los cambios de r a lo largo del tiempo. Se observó que estos cambios no están correlacionados con la cantidad de check-ins de esa semana. La fracción r se mantiene relativamente constante. Hay que decir que la cantidad de check-ins semanales rondaba los 150 y variaba semana a semana en 50 check-ins aproximadamente, en promedio El bajo valor de r puede deberse al bajo uso de la red en general, la poca cantidad de usuarios, así como también que el haber sacado los lugares de trabajo haya sacado lugares de esparcimiento importantes, como por ejemplo restaurantes, bares, etc..

7. Conclusiones

- Se hizo un análisis sobre la distribución de Grado y Cantidad de lugares visitados de la red, y se obtuvo que ambas distribuciones responden a una ley de potencias, con exponentes *-1,89* y *-1,35* respectivamente.

- Se logró comprobar, como se puede observar en la figura 3 que a medida que el número de amigos registrados de los usuarios aumenta, también lo hace el número de Check-Ins que realizan. Esto nos lleva a la conclusión de que los usuarios que tienen más amigos salen más seguido, o que por lo menos hacen más registros en un mismo intervalo de tiempo.

- Se observó de la red que los usuarios que más lugares distintos visitaron son aquellos que mayor distancia máxima en algunos de sus viajes recorrieron, como muestra la figura 4. Esto nos marca que a medida que la cantidad de lugares distintos visitados aumenta, es más seguro que el usuario comience a agotar las salidas cercanas y comience a viajar a lugares cada vez más lejanos.

- De los resultados obtenidos intentando recrear la red de amigos usando la red Bipartita y la red de Similaridad se puede concluir que no es posible representar los enlaces de amistad entre los usuarios considerando únicamente la distribución espacial de los Check-Ins.

- Se consideró un criterio de distancia entre eventos espacio-temporal para añadir enlaces entre nodos que lo cumplieran, se observó que el 70 % de los enlaces del grafo proyectado se repetían en el grafo original pero debido a la gran diferencia entre el número de enlaces de ambos grafos(mucho mayor en

el original) no se pudo concluir que este criterio sea el correcto para rearmar la red de amistad original a partir de los *check-ins* de cada usuario.

- Se realizó un análisis geográfico, en donde se vió que el hecho de que dos lugares compartan varios usuarios semanales indica, aunque no concluyentemente, que los lugares se encuentran cerca geográficamente.

8. Aspectos a mejorar en el análisis

En el análisis de la red de Similaridad no se consideran dos casos extremos. El primero es el de aquellos usuarios con uno o dos Check-Ins, ya que estos casos son usuarios que aparejados con otros usuarios que realizaron pocos Check-Ins, podrían considerar una alta similaridad, cuando en realidad la información de Check-Ins de estos usuarios es muy baja como para siquiera distinguir un verdadero hábito de visitas. Es por esto que muy probablemente el alto número de usuarios con bajos Check-Ins puede haber introducido mucho ruido en los resultados.

El otro caso extremo es el de aquellos lugares que son muy concurridos, por números muy grandes de personas muy frecuéntemente, como estaciones de trenes, o centros comerciales importantes, donde la gente que va a estos lugares no conoce a la gran mayoría de gente que allí concurre.

Es considerando estos dos casos que se podría considerar rehacer este análisis pero filtrando aquellos casos que cumplan con lo anteriormente descrito, y comparar si los resultados obtenidos en ese caso siguen manteniendo la misma distribución que la ya obtenida.

9. Bibliografía

[1]: *https://snap.stanford.edu/data/loc-Brightkite.html*