Universidad Nacional del Sur

Tesis de Doctor en Matemática

Estadística de Procesos Estocásticos y Aplicaciones a las Redes Sociales

Melina Valeria Guardiola

Bahía Blanca Argentina

Agosto, 2013

Para mis padres.

Prefacio

Esta tesis es presentada como parte de los requisitos para optar al grado académico de Doctor en Matemática de la Universidad Nacional del Sur y no ha sido presentada previamente para la obtención de otro título en esta Universidad u otras. La misma contiene resultados obtenidos en investigaciones llevadas a cabo en el Departamento de Matemática de la Universidad Nacional de Sur durante el período comprendido entre los meses de junio de 2008 y agosto de 2013, bajo la dirección del Dr. Gonzalo Perera, Profesor Titular Grado 5 de la Universidad de la República, Uruguay, y la supervisión de la Dra. Ana Tablar, Profesora Adjunta del Departamento de Matemática de la Universidad Nacional del Sur. Este trabajo fue financiado con una beca otorgada por ANPCyT y UNS en el marco del Proyecto de Formación de Doctores en Áreas Tecnológicas Prioritarias (PFDT) del Programa de Recursos Humanos (PRH 2007 - código 37).

Agradecimientos

Antes de comenzar esta tesis, quisiera agradecer a toda la gente que me acompañó durante la realización de la misma.

- En primer lugar, a mi director Dr. Gonzalo Perera, por su dedicación, apoyo y comprensión. Por ser además un excelente docente, por su hospitalidad y generosidad en mis estadías en Uruguay. Su paciencia y confianza hicieron posible la culminación de este trabajo.

- También quiero agradecer a Ana Tablar, José Bavio y Beatriz Marrón, por el compañerismo, los buenos consejos, los aportes y la colaboración en el desarrollo de esta tesis.

- Agradezco el apoyo y la predisposición brindados por las autoridades y personal del Departamento de Matemática de la UNS y también a la Agencia Nacional de Promoción Científica y Tecnológica por su financiación mediante la beca otorgada.

- Finalmente, un agradecimiento de otra índole va dirigido a mis padres por estar siempre y apoyarme en cada etapa y a mis hermanas por aconsejarme y acompañarme.

A todos, muchas gracias!

7 de Agosto de 2013

Departamento de Matemática.

UNIVERSIDAD NACIONAL DEL SUR

Resumen

Las redes sociales en internet tales como Facebook, Twitter o Linkedin se han transformado en uno de los medios de comunicación más utilizados por millones de personas en el mundo generando una combinación entre elementos virtuales y componentes del mundo real, razón por al cual es interesante su estudio.

Comenzamos este trabajo planteando un modelo probabilístico para describir el comportamiento a largo plazo de Facebook. Debido al importante nivel de interactividad existente en la red, para simplificar el modelo utilizamos una dinámica Markoviana.

Introducimos el concepto de Transversalidad Completa en una red social y proponemos estimadores para medir la interacción entre los usuarios. Luego, realizamos test de hipótesis sobre la transversalidad en la comunicación con los estimadores propuestos. Por otro lado, proponemos segmentar la red para estudiar comportamientos entre los segmentos y encontrar índices de performance adecuados para medir la calidad en la segmentación.

Nuestro objetivo final consiste en analizar la viabilidad de la Teoría de Mundo Pequeño en Facebook, mediante un estimador que nos permitirá estudiar la conectividad en la red y, tomando resultados sobre trazas de potencias de matrices de Wigner, hallaremos su distribución asintótica bajo cierto contexto comunicacional. Formulamos además, una definición rigurosa del grado medio de separación entre dos perfiles y realizamos su cálculo suponiendo distintos modos de comunicación, para poder así concluir acerca de la presencia del fenómeno de mundo pequeño en la red social.

Abstract

Virtual social networks such as Facebook, Twitter or Linkedin have become one of the means of communication most used by million of people worldwide generating combination of virtual elements and components of the real world, that is the reason for study them.

We begin this work suggesting a probabilistic model to describe the long-term behavior of Facebook. Due to the high level of interactivity on the web, to simplify the model we use a Markovian dynamics.

We introduce the concept of Complete Transversal Comunication in a social network, we propose estimators to measure the interaction between users and we make hypotesis testing on transversality in communication with the proposed estimators. On the other hand, we propose to segment the network, study behaviors between the segments and find suitable performance indices to measure the quality of segmentation.

Our ultimate goal is to analyze the viability of Small World Theory on Facebook by an estimator to study the network connectivity and, by results of power traces of Wigner matrices, we obtain its asymptotic distribution under certain communicative context. Also, we formulate a rigorous definition of the average degree of separation between two profiles and we perform the calculation for different communication modes that allows us to conclude about the small world phenomenon in the social network.

Índice general

Introducción

Los avances tecnológicos han sido determinantes en las nuevas formas de interacción humana permitiendo que la comunicación sea más eficiente y supere barreras de espacio y tiempo antes limitadas.

Las nuevas tecnologías de comunicación como Internet han derivado en distintas formas, entre ellas las redes sociales que se han revelado en los últimos años como una herramienta comunicacional de un impacto inesperado. El contacto frecuente entre las personas a través de estas redes da origen a relaciones virtuales que se van desarrollando de acuerdo a sus intereses. Los temas de conversación dominantes, las noticias más votadas, las marcas con más "me gusta", los personajes públicos con más seguidores o los videos más vistos, constituyen en su conjunto un extraordinario sistema de detección de tendencias sociales en tiempo real.

El gran desafío en la actualidad consiste en desarrollar herramientas lo suficientemente precisas para reconocer a los usuarios más influyentes por sectores y mercados y entender mejor las nuevas dinámicas de circulación de información. A su vez, difícilmente una red social pueda ser observada en su totalidad, por lo que adquieren gran importancia los problemas estadísticos y de modelización probabilística. Por otra parte, muchas veces las redes sociales aportan ejemplos de situaciones estadísticas de modelos no identificables o de datos censurados, lo cual los hace particularmente desafiantes.

En definitiva, nos encontramos con una amplia gama de inquietudes como las que hemos mencionado y esta tesis intenta brindar un aporte en este sentido, restringiendo el análisis a la red social "Facebook" o similar en cuanto a reglas de uso. Para ello desarrollamos un modelo que, si bien no describe la realidad en su totalidad, resultará de gran utilidad como una aproximación al estudio de la dinámica de la red. Para el desarrollo del mismo utilizamos herramientas matemáticas y de estadística de procesos estocásticos, tratando de responder interrogantes como el de la existencia de transversalidad en la comunicación, el de encontrar la mejor forma de segmentar la red para llegar a un público objetivo o el del análisis del fenómeno de mundo pequeño.

A continuación, detallamos la forma en que fue organizado este trabajo y dejamos constancia de que el tema considerado es sólo un pequeño ángulo del amplio panorama que el concepto de red social ofrece a la matemática y en particular a la probabilidad y

estadística, esperando que resulte de interés y genere nuevas inquietudes para continuar la investigación en esta línea.

En el primer Capítulo, hacemos una descripción de los orígenes, evolución y características de las redes sociales desde una perspectiva general y presentamos algunos conceptos y medidas de representación para su modelización. Concluimos el capítulo haciendo hincapié en la gran difusión de las redes sociales virtuales y en particular en el funcionamiento y crecimiento en el largo plazo de Facebook.

El segundo Capítulo aborda el modelado de Facebook y para ello presentamos previamente herramientas sobre Cadenas de Markov necesarias para el desarrollo del modelo. Introducimos el concepto de Transversalidad Completa en la comunicación e intentamos dentro de este contexto hallar la distribución de las funciones aleatorias que intervienen en el modelo.

El tercer Capítulo propone dos test de hipótesis para probar Transversalidad Completa en la red y Transversalidad Completa entre Segmentos de la red. Presentamos herramientas de la teoría de U-Estadísticos necesarias para hallar la distribución asintótica de los estimadores que utilizamos para realizar los test mencionados. Además, intentamos estudiar la relación entre distintos segmentos para luego definir un índice de performance de utilidad para medir la calidad en la segmentación.

En el cuarto Capítulo abordamos el "fenómeno de mundo pequeño" en Facebook. En primer lugar, intentamos medir el nivel de conectividad proponiendo un estimador de la proporción de ciclos de cierta longitud en muestras de perfiles y estudiamos su distribución asintótica bajo la hipótesis de Transversalidad Completa. Resultados como los obtenidos sobre trazas de potencias de matrices de Wigner serán fundamentales para el desarrollo de esta tarea. Por último, podremos calcular bajo dos contextos comunicacionales extremos, el grado medio de separación entre dos perfiles cualesquiera de Facebook y concluir sobre la viabilidad de la Teoría de Mundo Pequeño en la red, dejando como inquietud a partir de estos resultados el estudio de este fenómeno en contextos comunicacionales intermedios.

Capítulo 1

Redes Sociales

Esta tesis aborda problemas que son de interés en el campo de las redes sociales. Por consiguiente, en este Capítulo introducimos algunas ideas fundamentales que nos permitirán tener una perspectiva general de sus orígenes, evolución y características, profundizando en el caso de la red social Facebook. Además presentamos algunos conceptos básicos de gran utilidad para su modelización matemática y probabilística.

Existe una vasta literatura para el tratamiento del tema por lo que sólo heremos mención de los principales textos en que nos basamos para definir y caracterizar tanto al Análisis de Redes Sociales como a las Redes Sociales. Entre ellos, contamos con las definiciones de Wasserman S. y Faust K. (1994) [45]; White H., Wellman B. y Nazer N. (2004) [50]; Wetherell C., Plakans A. y Wellman B. (1994) [49]; Granovetter M.(1973) [24] y principalmente con el enfoque de Freeman L. (2006) [20]. Por otro lado, para el tratamiento de las redes sociales virtuales contamos con información obtenida en internet.

El Capítulo se organiza de la siguiente manera. En la primera Sección discutimos la definición y las características principales del Análisis de Redes Sociales y en la siguiente, proporcionamos un resumen sobre su origen y evolución. En la Sección 3, presentamos algunos conceptos básicos y medidas de representación de las redes sociales. La Sección 4 muestra una síntesis en materia de la investigación en redes realizada hasta el momento. Finalmente, en la Sección 5 hacemos hincapié en la utilidad y el crecimiento vertiginoso de las redes sociales en internet para concluir en la Sección 6 con la descripción de la historia, funcionamiento y comportamiento a largo plazo de la red social Facebook.

1.1. Definición y características

Una *Red Social* es formalmente definida como un conjunto de actores sociales conectados por uno o más tipos de relaciones, según Wasserman S. y Faust K. [45]. De acuerdo con Boorman S.A. y White H.D. [9], y con Zhu J. [53], los miembros de la red social son unidades y las relaciones mediante las que se conectan, siguen patrones que son objeto de

estudio de los investigadores sociales. Estas unidades son comunmente individuos, grupos u organizaciones, pero en principio cualquier unidad que tenga conexión con otra puede estudiarse como miembro de una red, tal como páginas web, blogs, e-mails, mensajes de texto, familias, artículos, vecinos o naciones. White H.D. y colaboradores [50], señalan que los investigadores de diversos campos académicos han demostrado que las redes sociales funcionan en varios niveles, desde las familias hasta el nivel de las naciones, y juegan un papel fundamental en la determinación de la forma en que se resuelven problemas, la forma en que funcionan las organizaciones y el grado en que los individuos alcanzan el éxito en sus objetivos.

La corriente principal de la investigación social tradicional se centra exclusivamente en el comportamiento de los individuos. Según Freeman L. [20], este enfoque deja de lado la parte social o la estructura de la conducta humana, es decir, la parte relacionada con la forma en que interactúan los individuos y la influencia que tienen unos sobre otros. Sin embargo, los analistas de redes sociales toman dichas partes como base de construcción fundamental del mundo social, no sólo recogen los tipos de datos particulares, también comienzan su análisis desde una perspectiva distinta a la adoptada por los individualistas.

El *Análisis de Redes Sociales*, (Social Network Analysis, SNA de aquí en más), también llamado análisis estructural, es el estudio de la interacción entre los actores sociales, Freeman L. [20]. Las relaciones que estudian los analistas de redes sociales son aquellas que vinculan a los seres humanos individuales, pues estos investigadores sostienen que además de las caracterísicas individuales, los vínculos relacionales y la estructura social son necesarios para comprender por completo los fenómenos sociales. Wetherell y colaboradores [49] consideran que el SNA:

-Conceptualiza la estructura social como una red con vínculos que conectan a sus miembros y canalizan recursos.

-Se focaliza en las características de los vínculos en lugar de focalizarse en las características de los actores individuales.

-Considera a las comunidades como "comunidades personales", es decir, como redes de relaciones individuales que las personas conservan y utilizan a lo largo de su vida.

Como ha señalado Freeman L. [20], el enfoque estructural no se limita al estudio de las relaciones sociales humanas, sino que también está presente en casi todos los campos de la ciencia, por ejemplo, los químicos moleculares examinan cómo interactúan los distintos tipos de átomos para formar diferentes tipos de moléculas, los ingenieros eléctricos estudian cómo las interacciones de ciertos componentes electrónicos tienen influencia en el flujo de corriente a través de un circuito, como así también los biólogos estudian las formas en que cada especie del ecosistema interactúa e incide en los demás.

Podemos mencionar distintos tipos de redes. En general, los analistas de redes sociales

distinguen las siguientes:

- *Redes modo-uno versus redes modo-dos:*

  Las primeras involucran relaciones entre un conjunto de actores similares, mientras las segundas se basan en relaciones entre dos conjuntos distintos de actores. Un ejemplo de red modo-dos puede ser el análisis de una red formada por una organización privada y sus vínculos con agencias sin fines de lucro en una comunidad, según Hawe P. y colaboradores [26]. Además las redes modo-dos, también llamadas redes de afiliación, se usan por ejemplo para investigar relaciones entre un conjunto de actores y una serie de eventos, es decir, aunque las personas no tengan lazos directos entre ellos pueden tener interés en eventos similares o actividades en la comunidad y esto permite la formación de lazos débiles, como señala Granovetter M. en [24].

- *Redes sociocéntricas versus redes egocéntricas:*

  Según Wellman B. y colaboradores [48], las primeras consisten en un mapa de todos los lazos relevantes entre todos los nodos estudiados, por ejemplo, lazos relacionales entre todos los profesores de un colegio secundario. Las redes egocéntricas, en cambio, se refieren a los lazos que tiene una persona específica en los distintos contextos en los que interactúa.

La forma que tiene una red social permite determinar la utilidad para los individuos que la componen. Las redes más abiertas con lazos débiles tienen mayor probabilidad de introducir nuevas ideas y oportunidades a sus miembros que las redes cerradas con muchos lazos redundantes. Para el éxito individual es mejor estar conectado con varias redes que tener muchas conexiones dentro de una misma red. Además, los individuos pueden influir o actuar de intermediarios en sus redes sociales haciendo de puente entre dos redes que no están directamente vinculadas, esto es, según Scott J. [41], rellenar agujeros estructurales.

Las redes sociales han sido también utilizadas para examinar cómo las organizaciones interactúan unas con otras, caracterizando las conexiones más informales que ligan a los ejecutivos entre sí y las asociaciones y conexiones entre los empleados de las distintas organizaciones. Juegan un rol clave en la contratación, en el éxito en los negocios y en la performance del trabajo. Les da a las compañías formas de obtener información, disuadir a la competencia y reunirse para fijar precios o políticas, de acuerdo con Wasserman S. y Faust K. [45].

Por último, haremos mención de las características más importantes en que se basa el SNA desde el punto de vista de la investigación social, según Freeman L. [20]:

1) El SNA es el estudio de la estructura, ya que el enfoque de redes sociales se basa en la noción intuitiva de que el patrón de los vínculos sociales en los que están involucrados

los actores tiene consecuencias importantes para ellos. Esta es la característica más importante del SNA. Freeman L. argumenta que sociólogos como Auguste Comte, Ferdinand T*önnies*, Emile Durkheim, Sir Hebert Spencer, trataron de especificar los distintos tipos de vínculos sociales que ligaban a los individuos en distintas formas de colectividades sociales y que debido a que estaban interesados en los vínculos sociales, todos compartían una perspectiva estructural.

2) Siendo la intuición estructural el motor del SNA, éste se basa en datos empíricos sistemáticos especialmente relacionales o de red. Los datos relacionales se usan para explicar las relaciones entre dos o más actores sociales, digamos i y j. Estos difieren de los datos de atributos que describen las relaciones entre dos o más atributos de un solo actor social i ó j.

3) La tercer característica del SNA se basa en la Teoría de Grafos. Cuando un grafo es utilizado como modelo de una red social, los nodos o vértices representan a los actores y las relaciones que los conectan son los enlaces o aristas.

   Existen dos tipos de grafos, *dirigidos* y *no dirigidos*. Los *dirigidos* consisten en un conjunto de vértices y un conjunto de aristas y en este tipo de grafos la arista "e", es un par ordenado (i, j) representando la conexión de un vértice i a un vértice j, siendo i el vértice inicial y j el final. Si la dirección de una arista no es importante, o equivalentemente, la existencia de una arista de i a j implica necesariamente una arista de j a i, decimos que esa red es un grafo *no dirigido*. El *grado* de un vértice es el número de vértices adyacentes a él. El *diámetro* de un grafo es la mayor distancia entre dos nodos del mismo. Un *camino* de un vértice i a uno j es una sucesión de links distintos $(i, u_1)(u_1, u_2)...(u_k, j)$. La *longitud* de ese camino es el número de aristas, en este caso $k + 1$. Un *bucle* es una arista con extremos idénticos. Dos *aristas paralelas* son aristas que tienen los mismos extremos. Un *grafo simple* es un grafo sin bucles ni aristas paralelas. Un *ciclo* consiste en un camino cerrado en el que no se repite ningún vértice a excepción del primero que aparece dos veces como principio y fin del camino. Un *grafo conexo*, es un grafo tal que para toda partición de su conjunto de vértices en dos conjuntos no vacíos, existe una arista con un extremo en uno de esos conjuntos y el otro extremo en el otro conjunto. Un *árbol* es un grafo conexo que no contiene ciclos.

   Aunque los grafos constituyen una manera muy útil de representar información sobre redes sociales, cuando existen muchos actores y/o muchas clases de relaciones éstos pueden hacerse tan visualmente complicados que se hace muy difícil identificar estructuras. En su lugar, también es posible representar la información sobre redes sociales en forma matricial.

   La forma más común de matriz en el análisis de redes sociales es una matriz simple

compuesta por tantas filas y columnas como actores existan en el conjunto de datos y donde los elementos representan los vínculos entre los actores. La más simple y común de las matrices es la matriz binaria $A = (a_{ij})$, donde a_{ij} toma el valor 1 si existe un vínculo entre los vértices i y j, y 0 si no lo hay. Este tipo de matriz es el punto de partida de casi todos los análisis de redes y se denomina "matriz de adyacencia". Luego, un grafo no dirigido puede representarse por una matriz de adyacencia simétrica $A = (a_{ij})$. La Tabla 1.1 nos muestra una matriz de relaciones y la Figura 1.1 es el grafo que describe esta matriz de datos.

	A	**B**	**C**	**D**	**E**	**F**
A	-	1	1	0	0	0
B	0	-	1	1	0	0
C	1	1	-	0	0	0
D	0	0	0	-	1	1
E	0	0	0	0	-	1
F	0	0	0	1	1	-

Tabla 1.1

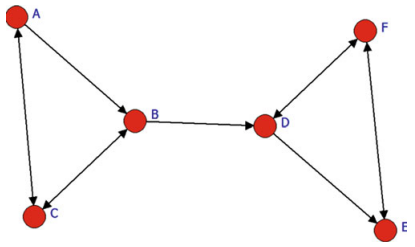


Figura 1.1

Por ejemplo, hay una línea imaginaria que conecta A (vértice inicial) con B (vértice final), y esta línea muestra una relación de orden. Estas dos formas equivalentes de representación nos permiten determinar tanto las características de la estructura como las propiedades de la posición de cada actor de la red.

4) Otra característica importante del SNA es el uso de modelos matemáticos y/o modelos computacionales. A diferencia de muchos otros enfoques de la investigación social, el análisis de redes se ha apoyado sistemáticamente en diversas ramas de la matemática tanto para aclarar sus conceptos como para explicar sus consecuencias en términos más precisos.

1.2. Historia del *SNA*

En esta sección, presentamos un breve resumen de la evolución de las redes sociales y el SNA siguiendo el texto de Linton Freeman [20].

- A comienzos del siglo XX, Georg Simmel (1858-1918) fue el primer estudioso que pensó directamente en términos de red social. Sus ensayos apuntan a la naturaleza del tamaño de la red sobre la interacción y a la probabilidad de interacción en redes ramificadas.

- En la década del 30, Jacob Levy Moreno (1889-1974), psicólogo social, inició el registro sistemático y el análisis de la interacción social en pequeños grupos, en especial en aulas y grupos de trabajo, dando lugar a la llamada Sociometría. En forma paralela en Harvard, W. Lloyd Warner y Elton Mayo exploraron las relaciones interpersonales en el trabajo.

- El período comprendido entre 1940 y 1960 se suele considerar como una "era oscura" en el SNA. Hubo aportes sustantivos como los de la escuela de Manchester, por ejemplo los estudios de parentesco de Elizabeth Bott en Inglaterra en los años 50 y entre 1950 y 1960 los estudios de urbanización de Max Gluckman, J. Clyde Mitchell et al. sobre las redes comunitarias en el sur de África, India y el Reino Unido. En particular en 1954, el antropólogo de Manchester, J.A. Barnes comenzó a estudiar sistemáticamente el término "red social" para mostrar patrones de lazos, abarcando conceptos más tradicionales como tribus, clanes, familias, género, etnia, clase social, etc.

- Entre los años 1960 y 1970 varios académicos trabajaron para incluir bajo la denominación SNA una amplia gama de temáticas. Entre ellos podemos citar a Harrison White, Ivan Chase, Bonnie Erickson, Harriet Friedmann, Mark Granovetter, Nancy Howell, Joel Levine, Nicholas Mullins, John Padgett, Michael Schwartz, Barry Wellman, Charles Tilly y Stanley Milgram, quien desarrolló la tesis de los "Seis Grados de Separación". A partir de 1970, al menos en los departamentos de ciencias sociales de las grandes universidades el SNA era ya un tema instalado y objeto de cursos sistemáticos.

- En la década del 80, un número de sociólogos comenzó a utilizar el SNA como una técnica analítica para examinar el fenómeno socioeconómico. A mediados de los 80's, Mark Granovetter, propuso el concepto de "arraigo" orientando al enfoque del SNA en la corriente principal del campo de la investigación social. Argumentó que el funcionamiento de la economía está sumergido en la estructura social, pero que sin embargo la estructura social central son las redes sociales de los individuos.

- Después de 1990, el *SNA* fue asociado en forma gradual con el capitalismo social, llamando la atención de los estudiosos del campo de la sociología, la política, la economía, las ciencias de la comunicación y otras disciplinas. El libro de Ronald Burt, "Structural Holes" [11], representa este período. Burt argumenta que el capital social no tiene relación con la longitud de los lazos, pero sí con la existencia de agujeros estructurales. Otro de los que estudió el *SNA* desde la perspectiva del capital social fue Lin Nan [32]. Más adelante, cuando tratemos la teorización del *SNA* en la Sección 1.4, volveremos sobre estos dos investigadores sociales.

1.3. Conceptos básicos del *SNA*

Vimos que una red social se refiere a un conjunto de actores sociales, nodos o vértices que son conectados por uno o más tipos de relaciones.

A través de la medición empírica de una red, se encontrarán diversos roles y grupos de actores y podemos preguntarnos: ¿quiénes son los conectores, líderes, puentes, islotes?, ¿dónde están los agrupamientos o clusters y quienes están en ellos?, ¿quién está en el centro de la red y quién en la periferia?

Para entender a la red y a sus miembros los investigadores debieron evaluar la ubicación de los actores de la red. Por ejemplo, para medir la ubicación de cada vértice, se puede usar el concepto de "Centralización" y otros conceptos relacionados que veremos a continuación.

1.3.1. Medidas generales de la estructura de la red

- Lazos:

  Los lazos, enlaces o aristas conectan a dos o más vértices dentro de un grafo. Muchos de los comportamientos humanos, tales como intercambio de información o préstamo de dinero a alguien, son lazos dirigidos mientras que compañeros de membresías son ejemplos de lazos no dirigidos. Los lazos dirigidos pueden ser en dos direcciones como el caso de dos personas que se visitan o pueden existir en una dirección como cuando sólo se da apoyo emocional a alguien, según Plickert G. y colaboradores [37]. Ambos tipos de lazos, dirigidos y no dirigidos, pueden medirse como lazos binarios que existen o no dentro de cada díada o lazos que pueden ser más o menos fuertes, transmitir más o menos recursos, o tener contactos más o menos frecuentes.

- Densidad:

  Describe el nivel general de vinculación entre los vértices o actores de un grafo y es un indicador de la conectividad existente en el mismo. Un grafo completo es aquel en el que los puntos son adyacentes unos con otros, es decir, cada punto está conectado directamente con todos los otros.

La densidad es la proporción de lazos existentes comparados con los posibles y puede definirse intuitivamente como el cociente entre el número de relaciones efectivas y el número de relaciones posibles.

- Clique:

  Un clique es un subconjunto de actores que están más fuertemente conectados mutuamente que lo que lo están con otros actores que no forman parte del grupo. Es decir, un clique en un grafo es un subgrafo en el cual cualquier vértice está directamente conectado con cualquier otro del subgrafo.

  La división de actores en cliques o subgrupos puede ser un aspecto muy importante de la estructura social. Pueden ser importantes para comprender el comportamiento de la red en su conjunto. Saber cómo un individuo está inmerso en la estructura de grupos en una red puede ser un aspecto crítico para la comprensión de su conducta. Por ejemplo, algunos pueden actuar como puentes entre grupos (los cosmopolitas, ampliadores de fronteras). Otros pueden tener todas sus relaciones dentro de un único clique (los locales o internos). Algunos actores pueden ser parte de una élite cerrada y densamente conectada, mientras otros estar completamente aislados de ese grupo. Tales diferencias en las formas que los individuos están inmersos en la estructura de grupos en una red pueden tener profundas consecuencias en las maneras en que esos actores perciben su funcionamiento.

1.3.2. Medidas de centralidad de los actores de la red

- Grado, intermediación y proximidad de los actores:

  El "grado" es el número de lazos de un actor en la red, la "intermediación" utiliza a un actor para comunicar a otros actores y la "proximidad" es la distancia entre un actor y el resto de la red. Qué atributo estructural se elige depende de lo que se quiera analizar. Una medida basada en el "grado" sirve para evaluar la capacidad de comunicación de que dispone un actor de la red; una medida basada en la "intermediación", para control de la comunicación y una medida basada en la "proximidad", si queremos evaluar la independencia de un actor.

- Centralidad:

  Si nos interesamos por un actor y los efectos de su posición estructural en la red, es necesario conocer su centralidad. Nos referiremos al actor focal como "ego" y a los otros como "alteregos". Los individuos centrales ocupan una posición privilegiada en los intercambios, en particular por comparación con aquellos que son rechazados a la periferia, ellos son los vértices más significativos de la red y es razonable pensar que esto se traduce en términos de poder.

a) Centralidad de grado

Esta medida organiza a los actores por el número efectivo de sus relaciones directas en el conjunto de la red. Trata de la centralidad local de un actor con respecto a los actores cercanos, pero dice poco sobre la importancia del actor en la red completa, y es muy sensible a variables como el tamaño del grafo. El grado de un vértice es útil como índice de su potencial de comunicación.

b) Centralidad de intermediación

La centralidad vista como intermediación se define como el nivel en que otros actores deben pasar a través de un actor focal para comunicarse con el resto de los actores. Sintetiza el control que cada uno de los actores tiene de los flujos relacionales en el conjunto de la red. Es decir, el individuo más conectado o centralizado es el más popular, eficiente o poderoso.

c) Centralidad de proximidad

Una tercera manera de medir la centralidad de un vértice consiste en medir su grado de proximidad con respecto a todos los demás individuos. Es una medida más global, utilizando no sólo las conexiones de un individuo a su vecindario sino también a su proximidad con los miembros de la red. Se basa en la noción de distancia y mide independencia.

1.3.3. Medidas de cohesión de la red

- Unipolaridad

  La unipolaridad indica el valor del grado del actor más central en relación al máximo de centralidad posible que podría tener ese actor. Es decir, si un actor juega un papel decisivo en la conexiones con los otros y lo hace directamente, la unipolaridad aumenta, representando el mayor grado efectivo de entre los de la red.

- Integración del grafo

  Corresponde a la suma del grado de todos los actores de un grafo. De modo estándar sería la razón entre la suma efectiva de los grados de todos y cada uno de los actores y el valor máximo de la suma de los grados posibles.

- Centralización

  Es la aplicación del concepto de "centralidad" a toda la red. Una red puede ser centralizada o descentralizada. Al igual que para la centralidad, existen tres medidas de centralización y cada una corresponde a una de las propiedades usadas para definir la centralidad de los actores de la red (grado, intermediación, proximidad). Se considera que una centralización de "grado fuerte" indica comunicación activa entre

todos los miembros de la red, mientras que una "fuerte intermediación o proximidad" traduce el hecho de que un número pequeño de actores controla esta comunicación.

1.4. Estado del arte

Antes de llevar a cabo un análisis de una red social, en especial antes de realizar la recolección de datos, el investigador debe decidir qué tipo de redes querrá analizar y qué tipo de relaciones está interesado en estudiar. Por ejemplo, deberá diferenciar entre las dos dimensiones importantes que hemos visto, a través de las cuales los datos de la red varían: redes unimodales vs. redes bimodales y redes completas vs. redes egocéntricas. Una vez tomada esta decisión debe seleccionar el método de recolección de datos que puede ser por observación, archivos, material histórico, resúmenes, entrevistas, experimentos, etc.

El objetivo básico, como en todos los campos de investigación de las ciencias sociales, es la teorización del SNA. Según Marin A. y Wellman B. [33], como perspectiva o paradigma de investigación el SNA toma como punto de partida la premisa de que la vida social se genera principalmente por las relaciones y los patrones que las componen y proporciona una manera de estudiar un problema pero no predice qué es lo que va a suceder. Sin embargo, en la etapa de diseño de la investigación los académicos deben hacer todo lo posible para llevar a cabo su trabajo con la aplicación de teorías relacionadas que apuntan a extender y modificar las teorías sociales, siendo varias de las teorías famosas reveladas bajo la perspectiva de las redes sociales.

El SNA se ha utilizado en epidemiología para ayudar a comprender cómo los patrones de contacto humano favorecen o impiden la propagación de enfermedades como el HIV en una población. A veces la evolución de una red social puede modelarse usando modelos basados en agentes, proporcionando información sobre la interacción de reglas de comunicación, propagación de rumores y estructuras sociales.

La teoría de Difusión de Innovaciones explora las redes sociales y su rol en la influencia de la difusión de nuevas ideas y prácticas. El cambio en los agentes y en la opinión del líder a menudo tienen un papel importante en el estímulo a la adopción de innovaciones, aunque también intervengan factores inherentes a las innovaciones.

Por su parte, Robin Dunbar sugirió que el tamaño típico en una red egocéntrica está limitado a unos 150 miembros, debido a los posibles límites de la capacidad del canal de comunicación humano. Esta norma surge de los estudios transculturales de la sociología y especialmente de la antropología sobre el número máximo de individuos de una aldea (en el lenguaje moderno mejor entendido como una ecoaldea). Esto está teorizado en la psicología evolutiva, que sostiene que este número puede ser una suerte de límite o

promedio de la habilidad humana para reconocer miembros y del seguimiento de hechos emocionales sobre todos los miembros de un grupo. Sin embargo, este puede deberse a la intervención de la economía y la necesidad de seguir a los "polizones", lo que hace que sea más fácil en grandes grupos sacar ventaja de los beneficios de vivir en una comunidad sin contribuir a ella.

Usando una perspectiva de redes, Granovetter M. [24] puso de manifiesto la teoría de la "fuerza de los lazos débiles". Encontró que un gran número de lazos débiles puede ser útil para la búsqueda de información e innovación. Los cliques tienen una tendencia a tener opiniones más homogéneas, así como a compartir rasgos comunes, por lo que esta tendencia homofílica es la razón por la cual los miembros de los cliques se atraen en principio y además cada uno conoce en cierto modo lo que conocen los demás. Para encontrar nueva información o ideas, las personas con frecuencia miran más allá del clique a otros amigos y conocidos. Sin embargo, el científico chino Bian Y. [4] revela la "teoría de los lazos fuertes", encontrando que en China las redes personales se utilizan para influenciar a las autoridades a asignar trabajo a sus contactos, y esta actividad ilegal es facilitada por los lazos fuertes caracterizados por confianza y obligación.

Otro ejemplo de teorización de la investigación de redes sociales es el "fenómeno de mundo pequeño". Este fenómeno será abordado en mayor profundidad en el Capítulo 4 y se refiere a la hipótesis sobre que la cadena de conocidos sociales necesaria para conectar a una persona arbitraria con otra persona arbitraria en cualquier parte del mundo, es generalmente corta. El concepto dio lugar a la famosa frase de "seis grados de separación" a partir de los resultados del "experimento de un mundo pequeño" hecho en 1967 por el psicólogo social Stanley Milgram. En el mismo, a una muestra de individuos en Estados Unidos se les pidió que hicieran llegar un mensaje a una persona objetivo en particular, pasándolo a lo largo de una cadena de conocidos. La duración media de las cadenas exitosas resultó ser de unos cinco intermediarios, o seis pasos de separación. Los métodos, y también la ética, del experimento de Milgram fueron cuestionados más tarde y algunas otras investigaciones para replicar los hallazgos de Milgram habrían encontrado que los grados de conexión necesarios podrían ser mayores. Investigadores continúan explorando este fenómeno dado que la tecnología de comunicación basada en Internet ha completado la del teléfono y los sistemas postales disponibles en los tiempos de Milgram. Un reciente experimento electrónico del mundo pequeño en la Universidad de Columbia, arrojó que cerca de cinco a siete grados de separación son suficientes para conectar a dos personas cualesquiera a través de e-mail, según Watts D. [47].

Luego de 1990 los investigadores extienden la teorización del SNA de manera importante. Entre los más destacados se encuentran Burt R. [11] con su teoría de los "agujeros estructurales" y Lin N. [32] con la "teoría del capital social". Burt argumenta que las cone-

xiones débiles entre grupos son agujeros en la estructura social de mercado. Estos agujeros en la estructura social crean una ventaja competitiva para individuos cuyas relaciones se extienden a los agujeros. Además los agujeros estructurales son una oportunidad para negociar el flujo de información entre las personas y controlar proyectos que reúnen a personas de ambos lados del agujero. Por otro lado Lin N. supone que la estructura social a nivel macro es un tipo de estructura jerárquica determinada por la asignación de recursos como riqueza, poder o status social. Explica la importancia del uso de conexiones sociales y relaciones sociales para alcanzar sus objetivos. El capital social así como los recursos accesibles a través de tales conexiones y relaciones son fundamentales para alcanzar metas individuales, para grupos sociales y comunidades.

Los "grafos de colaboración" pueden ser utilizados para ilustrar buenas y malas relaciones entre los seres humanos. Un enlace positivo entre dos vértices denota una relación positiva (amistad, alianza, citas) y un enlace negativo una relación negativa (odio, ira). Estos grafos de redes sociales signados pueden ser utilizados para predecir la evolución futura del grafo. En ellos, existe el concepto de ciclos "equilibrados" y "desequilibrados". Los primeros son aquellos en que el producto de todos los signos es positivo. Los grafos equilibrados o balanceados representan un grupo de personas con muy poca probabilidad de cambio en sus opiniones sobre las otras personas en el grupo, mientras que los desequilibrados representan un grupo de individuos que es muy probable que cambie sus opiniones sobre los otros en su grupo. Por ejemplo, en un grupo de tres personas, A, B y C, donde A y B tienen una relación positiva, B y C tienen una relación positiva, pero C y A tienen una relación negativa, es un ciclo de desequilibrio. Este grupo es muy probable que se transforme en un ciclo equilibrado, tal que B sólo tenga una buena relación con A, y tanto A como B tengan una relación negativa con C. Al utilizar el concepto de ciclos equilibrados y desequilibrados, puede predecirse la evolución de un grafo de red social.

1.5. Redes sociales en internet

Las redes sociales en internet han ganado su lugar de una manera vertiginosa convirtiéndose en negocios promisorios para empresas, artistas, marcas, freelance y sobre todo en lugares para encuentros personales.

En las redes sociales tenemos la posibilidad de interactuar con otros aunque no los conozcamos, el sistema es abierto y se va construyendo con lo que cada suscripto aporta a la red, cada nuevo miembro que ingresa transforma al grupo en otro nuevo. La red no es la misma si uno de sus miembros deja de formar parte.

Intervenir en una red social nos permite relacionarnos con personas con las cuales compartir nuestros intereses, preocupaciones o necesidades. Muchas veces suelen dar al anónimo popularidad, al discriminado integración, etc. Además, la fuerza del grupo per-

mite cambios sobre el individuo que de otro modo tal vez resulten difíciles y genera nuevos vínculos afectivos y también de negocios.

La causa del éxito y la popularidad que han adquirido tan velozmente las redes sociales en internet se debe en su mayor medida a los beneficios psicosociales que acabamos de mencionar.

El origen de las redes sociales virtuales se remonta al menos, a 1995, cuando Randy Conrads crea el sitio web classmates.com. Con esta red social se pretendía que la gente pudiera recuperar o mantener el contacto con antiguos compañeros del colegio, instituto, universidad, etc.

En el año 2002 comienzan a aparecer sitios web promocionando las redes de círculos de amigos en línea, y este concepto se hizo popular en el 2003 con la llegada de sitios tales como MySpace o Xing. Hay más de 200 sitios de estas redes sociales, aunque Friendster ha sido uno de los que mejor ha sabido emplear la técnica del círculo de amigos. La popularidad de estos sitios creció rápidamente y grandes compañías ingresaron en el espacio de las redes sociales en internet. Por ejemplo, Google lanzó Orkut el 22 de enero de 2004. Otros buscadores como KaZaZZ! y Yahoo crearon redes sociales en el 2005.

Las redes sociales continúan avanzando en internet a pasos agigantados, especialmente dentro de lo que se ha denominado Web 2.0 y Web 3.0, y dentro de ellas cabe destacar un nuevo fenómeno que pretende ayudar al usuario en sus compras por internet: las redes sociales de compras que son utilizadas por marcas y empresas para darse a conocer y hacer publicidad para grupos sociales con perfiles determinados. De esta forma aumenta la efectividad del boca a boca y las ventas, mejoran las relaciones con los clientes, es mucho más sencillo medir los resultados y segmentar al público. Además, brindan un espacio en el que los usuarios pueden consultar dudas sobre los productos en los que están interesados, leer opiniones y escribirlas, votar a sus productos favoritos, conocer gente con sus mismas aficiones y, por supuesto, comprar ese producto en las tiendas más importantes con un solo click. Esta tendencia se denomina Shopping 2.0.

El número de redes que existen en la actualidad en internet es muy elevado, pero lo realmente interesante es que, debido a la eficacia de estas páginas, han aparecido muchas otras especializadas en diferentes ámbitos. Las redes sociales en internet pueden definirse básicamente como horizontales o verticales. Llamamos redes sociales horizontales a aquellas cuyos usuarios tienen intereses afines. Por ejemplo, redes sociales de música, de fútbol o de libros. Conocemos como redes sociales verticales, a aquellas en las que los usuarios tienen intereses y motivaciones heterogéneas. Aquí entran Twitter y Facebook, donde hay usuarios de todos los tipos y se habla de todo. Por otro lado, encontramos que los tipos de redes sociales que existen en internet se pueden clasificar del siguiente modo:

-Las redes sociales globales, es decir, aquéllas que unen a grupos heterogéneos de personas con la intención principalmente de socialización. Entre ellas podemos destacar a Facebook, Twitter, MySpace y Google+.

-Las redes profesionales, que sirven para relacionar a profesionales de diferentes ámbitos en la búsqueda de objetivos laborales. La red profesional más grande del mundo es Linkedin que relaciona a más de 42 millones de usuarios de más de 200 países del mundo y de los que aproximadamente la mitad son de afuera de Estados Unidos. Xing, es otra de las redes que ha absorbido las plataformas profesionales más importantes, tiene una fuerte presencia en Alemania (su país de origen), España y China, con más de siete millones de usuarios en total.

-Las redes personales, que se caracterizan por ser la perfecta expresión de cada usuario, es decir, el usuario define personalmente lo que desea indicar en esa red. Incluyen un conjunto de ideas innovadoras como que la tecnología se aplica a la resolución de necesidades personales, lo cual permite mejorar la forma en que el usuario realiza sus actividades particulares de trabajo, estudio, ocio o por ejemplo las relaciones con la administración.

- Las redes privadas, que corresponden a las redes horizontales cerradas, a las que sólo se accede siendo miembro de un grupo u organización. Son espacios donde las personas pueden interactuar en línea, compartir contenidos y realizar otras actividades colectivamente. Sin embargo, por ser privadas, pueden ser adaptadas y dirigidas a cumplir con los objetivos que definan las organizaciones.

En la siguiente Sección profundizamos sobre la descripción y características de la red social Facebook, objeto de estudio en esta tesis.

1.6. La red social Facebook

1.6.1. Orígenes y expansión

Facebook es un sitio web gratuito de redes sociales. Se creó como una versión en línea de los "facebooks" de las universidades americanas. Los "facebooks" son publicaciones que hacen las universidades al comienzo del año académico, que contienen las fotografías y nombres de todos los estudiantes y que tienen como objetivo ayudar a los mismos a conocerse. Facebook llevó esta idea a internet, primero para los estudiantes americanos y abrió sus puertas a cualquier persona que contara con una cuenta de correo electrónico.

Facebook nació en 2004 como un hobby de Mark Zuckerberg, en aquél momento estudiante de Harvard, y como un servicio para los estudiantes de su universidad. En su primer mes de funcionamiento, contaba con la suscripción de más de la mitad de los estudiantes de Harvard, y se expandió luego a otras universidades tales como MIT, Boston University y Boston College y las más prestigiosas instituciones de Estados Unidos.

Un año después, Facebook tenía más de un millón de usuarios, una oficina en Palo Alto, California y había recibido apoyo financiero de 500 mil dólares por parte de Peter Thiel, co-fundador de Pay-Pal. Ese mismo año incorporó a los alumnos de más de 25 mil

escuelas secundarias y dos mil universidades de Estados Unidos y el extranjero, logrando un total de 11 millones de usuarios.

En 2006, Facebook introdujo más universidades extranjeras y desarrolló nuevos servicios en su plataforma, tales como Facebook Notes (una herramienta de blogging con tagging, imágenes y otras utilidades) o la importación de blogs de servicios como Xanga, LiveJournal o Blogger, y ya en 2007 Facebook Marketplace, que compite con Craigslist. También implementó acuerdos comerciales con iTunes y recibió una inversión de capital adicional de 25 millones de dólares por parte de Peter Thiel, Greylock Partners y Meritech Capital Partners.

Facebook se hace público en 2006 permitiendo que, no sólo los estudiantes de determinadas universidades o escuelas americanas participaran, sino que todas las personas que tuvieran correo electrónico pudieran formar parte de su comunidad. Se convirtió entonces en una comunidad de comunidades, en la que se conectaban estudiantes, empresas y gente que podía elegir participar en una o más redes. Es una comunidad creada por y en función de sus miembros.

En febrero de 2007 llegó a tener la mayor cantidad de usuarios registrados en comparación con otros sitios web orientados a estudiantes de nivel superior, teniendo más de 19 millones de miembros en todo el mundo. En julio de ese año, Facebook anunció su primera adquisición, Parakey, Inc. de Blake Ross y Joe Hewitt y en agosto se le dedicó la portada de la prestigiosa revista Newsweek, además de una integración con YouTube.

A fines de octubre de 2007 la red de redes vendió una parte, el 1,6 por ciento, a Microsoft a cambio de 240 millones de dólares, con la condición de que Facebook se convirtiera en un modelo de negocio para marcas de fábrica en donde se ofrecieran sus productos y servicios.

Una fuerte inyección de capital a Facebook (27,5 millones de dólares) fue liderada por Greylock Venture Capital. Uno de los socios de Greylock es Howard Cox que, según el diario The Guardian, pertenece al fondo de inversión en capital de riesgo de la CIA.

En 2008 lanzó su versión en francés, alemán y español para impulsar su expansión fuera de Estados Unidos, ya que sus usuarios se concentraban en Estados Unidos, Canadá y Gran Bretaña. La mayor cantidad de usuarios de Iberoamérica, proviene de Colombia, superando a países con mayor población como México, Brasil y Argentina.

Facebook compite por abrirse espacio entre empresas de éxito como Google y MySpace, por lo que se enfrenta a grandes desafíos para lograr crecer y desarrollarse. Una de las estrategias de Zuckerberg ha sido abrir la plataforma Facebook a otros desarrolladores. La propuesta económica es que quienes construyan algo sobre Facebook se quedarán con el dinero generado por la publicidad o por las transacciones.

Lo más importante es la dimensión viral del sistema: cuando un amigo agrega una aplicación aparece en su página y en su perfil. Cliquear lleva a la aplicación y permite interactuar directamente con ella. Todos los amigos ven la elección y la consideran como

un voto a favor, lo cual los alienta a probarla ellos también. Así lo demuestran, además, los hechos. A los 10 días de lanzamiento, el número de aplicaciones disponibles habían pasado de 85 a más de 300. iLike, la más popular, comenzó con mil abonados a la mañana siguiente del lanzamiento y a los dos días ya eran 300.000.

En Facebook la información es filtrada por los amigos y las redes. El modelo no descansa sobre un motor de búsqueda, sino sobre las redes sociales. Casi cualquier persona con conocimientos informáticos básicos puede tener acceso a todo este mundo de comunidades virtuales.

Facebook está prohibido en Irán, Birmania y Bután.

En julio de 2009 Mark Zuckerberg, hizo público que la red social había alcanzado los 250 millones de usuarios. En julio de 2010, declara 500 millones de usuarios, y traducciones a 70 idiomas y en mayo de 2011, 600 millones. Su infraestructura principal está formada por una red de más de 50.000 servidores sobre GNU/Linux . Los números para el segundo cuatrimestre de 2012 indicaban que 955 millones de usuarios entraban al menos una vez al mes en la plataforma, lo que mostró un crecimiento del 29 por ciento respecto al año anterior y a fines de 2012 este número había alcanzado los 1.060 millones de usuarios.

En la actualidad, a pesar de las especulaciones sobre una supuesta baja de actividad en Facebook, la red social hizo públicos sus resultados del primer trimestre de 2013 y las cifras son bastante positivas. No sólo no han perdido usuarios sino que han conseguido aún más, llevando el total a 1.110 millones de usuarios activos cada mes. Con respecto a la cantidad total de usuarios en Argentina, ésta asciende a 20.501.120 y, de acuerdo a las estimaciones de Socialbakers, se encuentra entre los 12 países con mayor cantidad de usuarios.

En cuanto a usuarios que acceden a la red social desde dispositivos móviles, la cifra ascendería a 751 millones, lo que sigue confirmando la tendencia de que el futuro de Facebook está en el móvil.

1.6.2. Funcionamiento y características

El funcionamiento de Facebook es similar al de cualquier otra red social. Los usuarios se registran y publican información en su perfil (una página web personal dentro de Facebook). Allí pueden subir textos, videos, fotografías y cualquier otro tipo de archivo digital. El usuario tiene la posibilidad de compartir dichos contenidos con cualquier otro usuario o sólo con aquellos que forman parte de su red de contactos o amigos.

Principales servicios:

- Lista de amigos:

  En ella el usuario puede agregar a cualquier persona que conozca y esté registrada, siempre que acepte su invitación, hasta un límite de cinco mil amigos. En Facebook

se pueden localizar personas con quienes se perdió el contacto o agregar otros nuevos con quienes intercambiar fotos o mensajes. Para ello, el servidor de la red posee herramientas de búsqueda y de sugerencia de amigos.

- Grupos y páginas:

  Es una de las utilidades de mayor desarrollo. Se trata de reunir personas con intereses comunes. En los grupos se pueden añadir fotos, videos, mensajes, etc. Las páginas se crean con fines específicos y, a diferencia de los grupos, no contienen foros de discusión ya que están encaminadas hacia marcas o personajes específicos y no hacia algún tipo de convocatoria. Además, los grupos también tienen su normativa, entre la cual se incluye la prohibición de grupos con temáticas discriminatorias o que inciten al odio y falten al respeto y la honra de las personas. Si bien esto no se cumple en muchas ocasiones, existe la opción de denunciar y reportar los grupos que vayan contra esta regla y, para tal fin, Facebook incluye un enlace en cada grupo que se dirige hacia un cuadro de reclamos y quejas.

- Muro:

  Es un espacio en cada perfil de usuario que permite que los amigos escriban mensajes para que el usuario los vea. Solamente es visible para usuarios registrados. Se puede reproducir en el muro propio contenidos del muro de un amigo: "compartir", o hacer aparecer en el muro de un amigo algo que se difunde en el propio: "etiquetar".

- Fotos y videos:

  Cada usuario puede agregar a su cuenta álbumes fotográficos, digitales y videos.

- Regalos:

  Son pequeños íconos con un mensaje. Los regalos dados a un usuario aparecen en el muro con el mensaje del donante, a menos que el donante decida dar el regalo en privado, en cuyo caso el nombre y el mensaje del donante no se exhibe a otros usuarios.

- Aplicaciones:

  Son pequeñas aplicaciones o cuestionarios utilizados únicamente con el fin de entretener. Por ejemplo, descubrir cosas de la personalidad, averiguar quién es el mejor amigo, etc.

- Juegos:

  La mayoría de las aplicaciones encontradas en Facebook se relacionan con juegos de rol, juegos de trivias o juegos de habilidades.

- Chat:

  Facebook tiene una plataforma de chat que, si bien no es competitiva frente a otras, brinda un servicio bastante completo.

Polémicas sobre Facebook:

1) Una fuertísima inyección de capital a Facebook de 27,5 millones de dólares, pudo verificarse que fue realizada por Greylock Venture Capital (fondo de inversión con fuerte vínculo con la CIA). Dichos fondos de inversión están desde el principio vinculados a Facebook.

2) No hay ninguna seguridad de privacidad. Hay compañías de marketing global que han accedido a contenidos aún de chat, comprando el acceso a Facebook.

3) En el momento de aceptar el contrato de términos de uso de la comunidad, el usuario cede la propiedad exclusiva y perpetua de toda la información e imágenes que agregue a la red social.

4) Nadie sabe realmente cuantas personas lo usan, lo que motivó a desarrollar el modelo que presentaremos en el siguiente Capítulo de la tesis. Se producen las siguientes situaciones:

   a) Una misma persona puede tener varios perfiles de Facebook, los cuales no son distinguibles ya que lo que identifica a un perfil es una dirección de e-mail que puede utilizarse una sola vez. Es decir, muchos perfiles pueden corresponder a una misma persona.

   b) Núcleos sociales pequeños (por ejemplo, la juventud de determinado grupo político, social o deportivo de determinada ciudad) suelen compartir el uso de un mismo perfil compartiendo la contraseña. Por lo tanto, un perfil puede corresponder a muchas personas.

   En cualquier contexto comunicacional, lo que cuentan son las personas, no sus perfiles. En la actualidad, el gran desafío consiste en reconocer a los usuarios más influyentes por sectores y mercados y comprender la dinámica de circulación de información entre los mismos. Se pretende llegar, y con eco favorable, a las personas o público objetivo. Podemos observar que los fenómenos a) y b), claramente contrapuestos, y cuyo balance es completamente desconocido, generan un modelo de red "en dos pisos", que es fuertemente no-identificable donde lo que realmente interesa es lo que llamamos el "nivel inferior", nivel en el que se encuentran las personas, a partir de lo que es observable, o sea los perfiles, que se encuentran en lo que llamamos el "nivel superior".

5) Las teorías sociológicas sobre las Redes Sociales como las que plantea De Ugarte D. en [16] ven en Facebook una revolución cultural, social y comunicacional, sugiriendo la transversalización de la intercomunicación, punto muy polémico y que fue otra motivación de estudio en esta tesis.

1.6.3. Facebook a largo plazo

¿Cómo crece una red social?

Aunque hasta el momento no hemos observado ciclos de vida completos de una red social de éxito, los datos que hasta ahora tenemos nos permiten establecer un crecimiento en forma de curva logística (o en forma de "S") en tres etapas: inicio, explosión y cima (aún por llegar para la mayoría). En la Figura 1.2 vemos un posible ejemplo de aplicación al crecimiento de una red social.

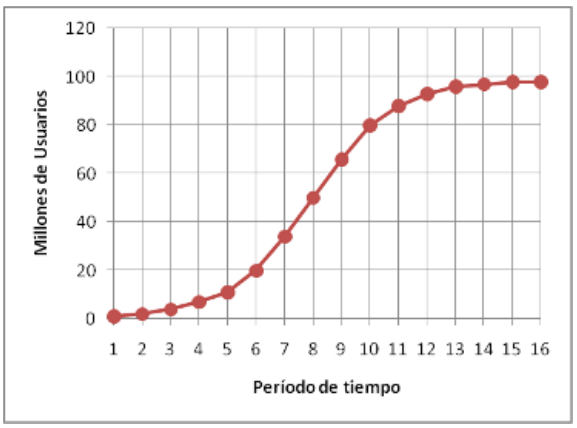


Figura 1.2

En la etapa de inicio, períodos 1-5 en la Figura 1.2, los usuarios comienzan a invitar a potenciales usuarios, con una tasa de éxito importante que conlleva un crecimiento exponencial. Aunque es importante contar con una buena fuente de usuarios, no es algo definitivo. Por ejemplo, mientras Facebook surgió prácticamente de la nada, MySpace contó con una fuente tan importante como eUniverse, que en aquel momento disponía de unos 20 millones de usuarios y suscriptores por e-mail. En este punto, la tasa de abandono de los usuarios suele ser baja.

La etapa de explosión, períodos 6-10 en la Figura 1.2, viene marcada por una bajada de la tasa de éxito, escondida por el gran número de usuarios ya disponibles. El crecimiento en este período deja de ser exponencial para convertirse en lineal. Su duración depende del nivel de saturación de la red y de la tasa de abandono de los usuarios. En los casos de Facebook y MySpace, el comienzo de la saturación se vio disipado por una gran expansión a nivel internacional. No sabemos aún si se ha llegado a la cima en cuanto al número de

usuarios registrados, pero puede ser que sí estemos asistiendo a esta fase en lo que a número de usuarios activos (usuarios que acceden al menos una vez al mes) se refiere.

¿Se puede prever el radio de crecimiento?

Según publica Carrero F. [12], cuando se lanza una red social se realizan estimaciones sobre su crecimiento, más concretamente, sobre cuántos usuarios se espera tener en las diferentes fases. Se pueden construir modelos para justificar de alguna manera el rápido crecimiento que se ha observado de determinados momentos, con el objeto de prevenir situaciones de riesgo. Sin embargo, al comienzo de la vida de la red es muy difícil determinar cuál será la tasa de crecimiento en las primeras fases y cuál será el valor que determinará su saturación, y no será hasta entrar en la fase de la explosión cuando se pueda prever con mayor precisión dónde está el techo.

Existen múltiples razones para querer predecir el crecimiento de una red social. Una de ellas es determinar si se está en el camino de alcanzar las expectativas iniciales y, si no es así, buscar un remedio para potenciar el crecimiento; ver si se podrán sobrepasar esas expectativas; anticipar el momento en que podemos empezar a llegar a la saturación de la red y buscar alternativas para no abandonar la fase de explosión.

Y aún si se llega a entrar en fase de cima, esto no debe tomarse como un límite máximo, ya que podrían darse situaciones en las que la red se estabiliza un tiempo en la cima y posteriormente vuelve a crecer (por ejemplo, por una mejora del producto, una expansión de mercado, una fusión o adquisición de otra empresa, etc.).

Ciclo de vida de Facebook

Facebook sigue creciendo, aunque a un ritmo más lento de lo que lo hacía al principio. Si bien hemos visto que al momento la cantidad de usuarios fue registrada en 1.110 millones de usuarios activos, la curva de crecimiento ha tendido a suavizarse en estos últimos años y la red social parece respetar un crecimiento en forma de curva logística a lo largo del tiempo como el que hemos expuesto.

De acuerdo con lo publicado en el 2011 por Closa G. [14], en marketing y economía, el ciclo vital de un producto tiene cuatro fases:

- introducción/lanzamiento
- crecimiento
- madurez
- declive

Sin embargo, Facebook no funciona así, parece tener su propio ciclo de vida. Mientras el resto de los productos gozan de cuatro fases, en esta red social parece que la fase

de introducción/lanzamiento reaparece cuando la fase de la madurez alcanza un estadío avanzado. Esto es posible manteniendo a los usuarios en vilo con pequeños cambios en su interfaz en períodos de tiempo estudiados. Cualquier actualización de diseño o nueva característica de la red es válida para hacer de ello un relanzamiento del producto.

En el mundo 2.0 se sabe que en un período de tiempo aparecerá un producto que mejorará al que estaba anteriormente. En Facebook, en cambio, parece lógico que con tanta actualización, aunque el crecimiento no sea exponencial, se tardará más en llegar a la fase de declive.

Por su parte, los representantes de Facebook justifican la falta de crecimiento en los mercados más importantes con el argumento de que en Estados Unidos y el Reino Unido todos aquellos que deseaban abrir una cuenta en Facebook ya lo han hecho. Los usuarios estadounidenses no sólo tienen actualmente menos tendencia a revisar sus perfiles, sino que también pasan menos tiempo en la página, ya que prefieren las nuevas aplicaciones para teléfonos inteligentes.

Sin embargo, los últimos datos indican que Facebook ha obtenido un beneficio neto de 219 millones de dólares en los tres primeros meses de 2013, lo que supone un aumento del 6,8 por ciento respecto al mismo período del año anterior, según ha anunciado la propia empresa. El beneficio total de Facebook ha alcanzado los 1.458 millones de dólares, una subida interanual de un 37,8 por ciento, de los cuales un 49 por ciento se originaron en Estados Unidos y Canadá, y otro 29 por ciento en Europa.

Tratando de hacer frente a la nueva competencia, Mark Zuckerberg se ha centrado en los "smartphones". A principios de abril de 2013 Facebook presentó la nueva interfaz, "Facebook Home", para el sistema operativo Android. "Facebook Home" permite a sus usuarios intercambiar mensajes en la pantalla del teléfono, incluso cuando la aplicación no está activada. Esta novedad ha contribuido a que 751 millones de usuarios hayan entrado al servicio a través de su "smartphone", un 54 por ciento más que hace un año. Así mismo, la popular aplicación fotográfica Instagram, adquirida por Facebook por 1.000 millones de dólares en 2012, alcanzó la cifra de 100 millones de usuarios activos.

Con el fin de sacar más rendimiento de la red social, la compañía de Zuckerberg estrenó tres nuevos servicios destinados a departamentos comerciales de empresas para facilitar que las marcas encuentren su público objetivo dentro de la plataforma.

En conclusión, si bien Facebook hace unos años que no tiene el crecimento vertiginoso que supo tener en su fase de expansión, con las actualizaciones mencionadas parece estar lejos de la fase de declive.

1.7. Conclusiones

- El *SNA* es el estudio de la estructura social. Los analistas de redes sociales están interesados en cómo los individuos están sumergidos dentro de una estructura y

cómo la estructura emerge de las relaciones micro entre partes individuales.

- La gran ventaja del SNA es que considera cómo la estructura de red comunicacional de un grupo influye en el conocimiento y comportamiento individual.

- Como una aproximación a la investigación social, el SNA muestra características tales como: intuición estructural, datos relacionales sistemáticos, imágenes gráficas y modelos matemáticos o computacionales.

- En más de 70 años de historia los analistas de redes sociales revelaron un gran número de maneras precisas y formales de definir términos tales como "relación", "densidad", "centralidad", "clique", entre otros, que pueden aplicarse sin ambigüedad a los datos en poblaciones de individuos.

- Reconociendo que estamos todos conectados como una red que no podemos ver, según Mickenberg R. y Dugan J. [35], el SNA se hace más y más popular entre investigadores de varios campos como sociología, matemática, ciencias de la computación, economía, ciencias de la comunicación y psicología alrededor del mundo.

- En ciencias sociales, la teorización del SNA ha mejorado en las dos décadas recientes, a pesar de haber sido criticada antes de los 80's. Sin embargo, las nuevas tecnologías (internet, telefonía móvil, transmisión digital, etc.) han hecho la recolección de datos sociales más fácil a mayor escala y a costos más bajos que los métodos convencionales, mejorando así problemas relacionados con el análisis e interpretación de los datos.

- Las técnicas existentes parecen ser inadecuadas para manejar los nuevos tipos de datos de redes sociales que son continuos, dinámicos y multinivel. Gracias a la situación actual, el desarrollo y búsqueda de nuevas técnicas, las herramientas que resuelven estos problemas deben estar en las agendas de investigación de los científicos sociales.

- Las redes sociales en internet han ganado velozmente un lugar, conjugando pluralidad y comunidad. Las herramientas informáticas para potenciar la eficacia de las redes sociales online (software social), operan en tres ámbitos: la comunicación, la comunidad y la cooperación.

  Las redes sociales continúan avanzado en internet, por ejemplo convirtiéndose en un sitio de consulta y compra, dando lugar a la tendencia Shopping 2.0.

- Desde sus comienzos la red social Facebook ha crecido de manera vertiginosa hasta convertirse en una de las redes sociales más importantes con 1.110 millones de usuarios activos hasta el momento. En la actualidad, si bien sigue creciendo, lo hace

a un ritmo más lento de lo que lo hacía al principio. Este comportamiento a lo largo del tiempo, permite representar a su evolución en forma de curva logística. Sin embargo, cabe destacar que debido a las constantes actualizaciones y mejoras parece estar lejos de la etapa de declive.

Por otra parte, las polémicas que ha generado en cuanto a la falta de privacidad, la transversalidad de la comunicación o la incertidumbre sobre cuántas personas son las que realmente están detrás de cada perfil, resultaron la principal motivación para su análisis y modelado en esta tesis.

Capítulo 2

Modelo: Dinámica de Facebook a Largo Plazo

En este Capítulo presentamos un modelo probabilístico para estudiar la dinámica a largo plazo de Facebook. Proporcionamos previamente algunos conceptos sobre Cadenas de Markov, que serán de utilidad para el desarrollo del mismo. Nuestro interés se centrará en estudiar el comportamiento comunicacional de los "usuarios" de la red social que estarán representados por los "perfiles" e identificados por una dirección de e-mail.

Incorporamos además, el concepto de "Transversalidad Completa" y su aplicación en el modelo.

2.1. Modelos Markovianos

Las cadenas de Markov y los procesos de Markov son un tipo especial de procesos estocásticos que poseen la siguiente propiedad:

Propiedad de Markov: Conocido el estado del proceso en un momento dado, su comportamiento futuro no depende del pasado. Dicho de otro modo, "dado el presente, el futuro es independiente del pasado".

Definición 2.1 *Una cadena de Markov a tiempo discreto (CMTD) es un proceso estocástico* $\{X_t : t = 0, 1, 2, \dots\}$ *que toma valores en un espacio discreto* S*, llamado espacio de estados, y que satisface la propiedad de Markov, esto es, para cualquier* $t = 0, 1, 2, \dots$ *y para cualquier conjunto de estados* $\{x_0, \dots, x_{t+1}\} \subseteq S$*, se cumple:*

$$P(X_{t+1} = x_{t+1} | X_t = x_t, \dots, X_0 = x_0) = P(X_{t+1} = x_{t+1} | X_t = x_t).$$

Cuando el espacio de estados S de una cadena de Markov es un conjunto finito se dice que la cadena es *finita*.

Probabilidades de Transición: La probabilidad $P(X_{t+s} = j|X_s = i)$, denotada por $P_{ij}(s, t+s)$, representa la probabilidad de pasar del estado i a tiempo s al estado j a tiempo $t + s$. Estas probabilidades se conocen como las *probabilidades de transición en t pasos*. Cuando $P(X_{t+s} = j|X_s = i) = P(X_t = j|X_0 = i)$ decimos que la $CMTD$ es *homogénea* en el tiempo y expresamos a estas probabilidades como $P_{ij}(t)$.

La probabilidad $P(X_{t+1} = j|X_t = i)$, denotada por $P_{ij}(t, t+1)$ representa la probabilidad de transición del estado i a tiempo t, al estado j a tiempo $t+1$. Estas probabilidades se conocen como las *probabilidades de transición en un paso* y, si la $CMTD$ es homogénea en el tiempo, las denotamos por P_{ij}.

De aquí en más, los resultados que presentaremos supondrán que las $CMTD$ son homogéneas el tiempo.

Comenzando en el estado i la $CMTD$ irá a algún estado j (incluyendo la posibilidad de que $j = i$), por lo que se deduce que $\sum_{j \in S} P_{ij} = 1$, con $0 \leq P_{ij} \leq 1$. Las probabilidades de transición en un paso P_{ij} se resumen usualmente en una *matriz de transición estocástica*, es decir, una matriz $\mathbf{P}$ tal que la suma de los elementos de cada fila es uno:

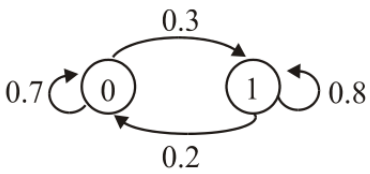
$$\mathbf{P} = (P_{ij}) = \begin{pmatrix} P_{00} & P_{01} \cdots \\ P_{10} & P_{11} \cdots \\ \vdots & \vdots \end{pmatrix}.$$

Gráficamente, una $CMTD$ con espacio de estados finito se representa mediante un *diagrama de transición de estados*, es decir, un grafo finito dirigido donde el estado i de la cadena es representado por un vértice y la transición en un paso del estado i al estado j, por una arista marcada por la probabilidad de transición P_{ij}.

Ejemplo 2.2 *Sea $\{X_t\}$ una CMTD con dos estados 0 y 1, con matriz de probabilidades de transición en un paso dada por:*

$$\boldsymbol{P} = \begin{pmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{pmatrix}, \tag{2.1}$$

el correspondiente diagrama de transición de estados es:



Partiendo del estado 0, se realiza una transición desde el mismo al estado 1 con probabilidad 0.3, y con probabilidad 0.7 la cadena permanece en 0 en el siguiente paso. Del mismo

modo, se produce una transición del estado 1 al 0 con probabilidad 0.2, y con probabilidad 0.8 la cadena permanece en el estado 1 en el siguiente paso.

Ecuación de Chapman-Kolmogorov: La transición del proceso del estado i a tiempo s al estado j a tiempo $t+s$, puede descomponerse en subtransiciones del estado i a tiempo s a un estado intermedio, digamos h, a tiempo r y de allí al estado j a tiempo $t+s$, donde $s < r < t+s$. Esta condición viene dada por la denominada *Ecuación de Chapman-Kolmogorov*:

$$P_{ij}(s, t+s) = \sum_{h \in S} P_{ih}(s, r)\ P_{hj}(r, t+s),$$

con $0 \leq s < r < t+s$.

Para $CMTD$ homogéneas en el tiempo las Ecuaciones de Chapman-Kolmogorov se simplifican de la siguiente manera:

$$P_{ij}(t) = \sum_{h \in S} P_{ih}(r)\ P_{hj}(t-r), \tag{2.2}$$

con $0 < r < t$. Debido a que la ecuación (2.2) se verifica para todo $r < t$, tomando en particular $r = 1$ obtenemos

$$P_{ij}(t) = \sum_{h \in S} P_{ih}\ P_{hj}(t-1).$$

Además, siendo $\mathbf{P}^{(t)}$ la matriz de probabilidades de transición en t pasos, la ecuación (2.2) puede ser reescrita en forma matricial para el caso particular en que $r = 1$ como $\mathbf{P}^{(t)} = \mathbf{P}\ \mathbf{P}^{(t-1)}$ y, aplicando este procedimiento en forma recursiva, resulta $\mathbf{P}^{(t)} = \mathbf{P}^t$.

Distribución de Probabilidad de una *CMTD*: Es de interés calcular la función de masa de probabilidad de la variable aleatoria X_t, es decir, las probabilidades $\pi_i(t) = P(X_t = i)$ de que la $CMTD$ se encuentre en el estado i, en el instante t.

Dada la matriz de probabilidades de transición en t pasos $\mathbf{P}^t$, la distribución de probabilidad en t pasos:

$$\vec{\pi}(t) = (\pi_0(t), \pi_1(t), \dots),$$

puede obtenerse de la siguiente manera:

$$\vec{\pi}(t) = \vec{\pi}(t-1)\ \mathbf{P} = \vec{\pi}(0)\ \mathbf{P}^t, \quad t = 0, 1, 2, \dots, \tag{2.3}$$

donde $\vec{\pi}(0) = (\pi_0(0), \pi_1(0), \dots)$ es la distribución de probabilidad inicial.

De esta forma se obtiene una sucesión $\vec{\pi}(0), \vec{\pi}(1), \vec{\pi}(2), \dots$ de distribuciones de probabilidad en donde cada una de ellas, excepto la primera, se obtiene de la anterior multiplicada

a derecha por la matriz de probabilidades de transición en un paso. Es natural preguntarse si tal sucesión converge a una distribución límite. Más adelante estudiaremos este problema y enunciaremos las condiciones bajo las cuales existe un único límite para la misma.

Clasificación de las *CMTD*: Las $CMTD$ se categorizan de acuerdo a la clasificación de sus estados.

Definición 2.3 *Un estado j se dice accesible desde otro estado i, con $i, j \in S$, si es posible pasar del estado i al j en una cantidad finita de pasos de acuerdo a la matriz de probabilidades de transición dada. Es decir, para algún $t \in \mathbb{N}$, $P_{ij}(t) > 0$. Además, decimos que los estados i y j están comunicados si j es accesible desde i e i es accesible desde j.*

La comunicación es una relación de equivalencia y en consecuencia induce una partición del espacio de estados, es decir, dos estados pertenecen a la misma clase de equivalencia si y sólo si tales estados están comunicados.

Definición 2.4 *Una $CMTD$ se dice irreducible si todos los estados se comunican entre sí, esto es, si existe sólo una clase de comunicación.*

Definición 2.5 *El estado $i \in S$ se dice absorbente si y sólo si una vez que se accede al mismo ningún otro estado de la $CMTD$ puede ser alcanzado desde él, es decir, si $P_{ii} = 1$.*

Observemos que una $CMTD$ que contiene al menos un estado absorbente no puede ser irreducible.

Definición 2.6 *Una colección de estados no vacía $\mathcal{C}$ de una $CMTD$ es cerrada si ningún estado fuera de $\mathcal{C}$ es accesible desde algún estado dentro de $\mathcal{C}$. Es decir, $i \in \mathcal{C}$ y $j \notin \mathcal{C}$, i y j no están comunicados.*

Definición 2.7 *Dado un estado $i \in S$, la probabilidad de recurrencia en t pasos, $f_i^{(t)}$, se define como la probabilidad del primer retorno al mismo en exactamente $t \in \mathbb{N}$ pasos después de haberlo dejado. Luego, la probabilidad f_i de regresar al estado i está dada por: $f_i = \sum_{t=1}^{\infty} f_i^{(t)}$.*

Definición 2.8 *Cualquier estado $i \in S$ al que la $CMTD$ retorne con probabilidad $f_i = 1$ se llama estado recurrente; en otro caso, si $f_i < 1$, i se llama estado transitorio.*

Definición 2.9 *Dado cualquier estado recurrente i, el tiempo medio de recurrencia m_i, del estado i de una CMTD está dado por:*

$$m_i = \sum_{t=1}^{\infty} t\, f_i^{(t)}.$$

Si $m_i < \infty$, i se dice recurrente positivo o no nulo; en otro caso, si $m_i = \infty$, i se dice recurrente nulo.

Definición 2.10 *Se define el período del estado recurrente $i \in S$, d_i, como el máximo común divisor del conjunto de enteros positivos t, tal que $P_{ii}(t) > 0$.*
Un estado recurrente se dice aperiódico si $d_i = 1$, y periódico de período d_i si $d_i > 1$.

Feller W. [19] demostró que todos los estados que pertenecen a una misma clase son del mismo tipo, es decir, todos son periódicos, aperiódicos, transitorios, recurrentes positivos o recurrentes nulos.

Definición 2.11 *Si uno de los estados $i \in S$ de una CMTD irreducible es aperiódico, entonces lo son todos los demás estados $j \in S$, es decir $d_j = 1$, para todo $j \in S$, y la CMTD se dice aperiódica.*

Definición 2.12 *Una CMTD irreducible, aperiódica con todos sus estados recurrentes positivos se dice ergódica.*

Observación 2.13 *Los estados de una CMTD irreducible finita son todos recurrentes positivos, en consecuencia una CMTD irreducible, aperiódica y finita es ergódica.*

***CMTD* en el Largo Plazo**: Para estudiar las propiedades de las $CMTD$ en el largo plazo nos basamos específicamente en las $CMTD$ ergódicas. En este tipo de cadenas se puede ir de un estado i hasta otro estado j en t pasos aunque no exista un arco dirigido de un estado i a un estado j, es decir, hay libertad de pasar de un estado a cualquier otro en t pasos.

Definición 2.14 *Una distribución de probabilidad $\vec{\pi} = (\pi_0, \pi_1, \dots)$ de una CMTD se dice estacionaria si cualquier transición de acuerdo con la matriz de probabilidades de transición en un paso $\boldsymbol{P} = (P_{ij})$ no tiene efecto sobre esas probabilidades, es decir, si para todo $j \in S$, se cumple que $\pi_j = \sum_{i \in S} \pi_i\, P_{ij}$.*
En términos matriciales, $\vec{\pi}$ es estacionaria si verifica:

$$\vec{\pi} = \vec{\pi}\, \mathbf{P}. \tag{2.4}$$

La condición $\vec{\pi} = \vec{\pi}\ \mathbf{P}$ tiene como consecuencia el hecho de que para cualquier $t \in \mathbb{N}$ se cumpla que $\vec{\pi} = \vec{\pi}\ \mathbf{P}^t$, es decir, $\vec{\pi}$ es también una distribución estacionaria para $\mathbf{P}^t$. Esto significa que si la variable aleatoria inicial X_0 tiene esa distribución $\vec{\pi}$, entonces la distribución de X_t también es $\vec{\pi}$ pues $P(X_t = j) = \sum_{i \in S} \pi_i\ (\mathbf{P}^t)_{ij}$, es decir, esta distribución no cambia con el paso del tiempo.

Observaciones:

1) El vector de ceros cumple la condición de $\vec{\pi} = \vec{\pi}\ \mathbf{P}$ pero no corresponde a una distribución de probabilidad.

2) Para encontrar una posible distribución estacionaria de una $CMTD$ con matriz de probabilidades de transición $\mathbf{P}$, un primer método consiste en resolver el sistema de ecuaciones (2.4).

3) Pueden darse sólo tres situaciones sobre la existencia de distribuciones estacionarias para una $CMTD$ cualquiera: o no existe ninguna, o existe una y es única o existen infinitas.

Cuando existe una distribución estacionaria, ésta tiene como soporte el conjunto de estados recurrentes positivos. Esto se resume en la siguiente proposición.

Proposición 2.15 *Sea $\vec{\pi}$ una distribución estacionaria. Si i es un estado transitorio o recurrente nulo entonces $\pi_i = 0$.*

Presentamos ahora sin demostración dos condiciones que en su conjunto garantizan la existencia y unicidad de una distribución estacionaria.

Proposición 2.16 *Toda $CMTD$ que es irreducible y recurrente positiva tiene una única distribución estacionaria dada por:*

$$\pi_j = \frac{1}{m_j} > 0,$$

con m_j el tiempo medio de recurrencia del estado j.
En particular, toda cadena finita e irreducible tiene una única distribución estacionaria.

Definición 2.17 *La distribución de probabilidad límite de una $CMTD$ se define por:*

$$\vec{\pi}(\infty) = \lim_{t \to \infty} \vec{\pi}(t) = \lim_{t \to \infty} \vec{\pi}(0)\boldsymbol{P}^t = \vec{\pi}(0)\ \lim_{t \to \infty} \boldsymbol{P}^t = \vec{\pi}(0)\ \boldsymbol{P}^{(\infty)}. \tag{2.5}$$

Supongamos que bajo ciertas condiciones la sucesión de distribuciones de probabilidad $\vec{\pi}(0), \vec{\pi}(1),\ \vec{\pi}(2), \ldots$ converge a una distribución de probabilidad límite $\vec{\pi}(\infty)$. Entonces esta distribución límite verificará las siguientes propiedades:

1) $\vec{\pi}(\infty)$ no dependerá de la distribución inicial y estará dada por el límite de las potencias de $\mathbf{P}$, $\mathbf{P}^{(\infty)}$, pues si se toma como distribución inicial a aquella concentrada en el i-ésimo estado, entonces el j-ésimo elemento de la distribución límite es $\pi_j = \lim_{t\to\infty}(\mathbf{P}^t)_{ij}$.

2) El límite de las potencias de $\mathbf{P}$ es una matriz con todas las filas idénticas, siendo esta fila la distribución límite.

3) De la ecuación (2.3) tenemos que $\vec{\pi}(t) = \vec{\pi}(t-1)\ \mathbf{P}$, por lo que si existe el límite para $\vec{\pi}(t)$ cuando $t \to \infty$, resulta:

$$\vec{\pi}(\infty) = \lim_{t\to\infty} \vec{\pi}(t) = \lim_{t\to\infty} \vec{\pi}(t-1)\ \mathbf{P} = \vec{\pi}(\infty)\ \mathbf{P}.$$

Por lo tanto, si existe la distribución límite, ésta es estacionaria. La afirmación inversa no es necesariamente cierta.

En espacios de estados finitos o infinitos, si existen los límites de las probabilidades $(\mathbf{P}^t)_{ij}$, cuando $t \to \infty$, y no dependen de i, la distribución límite podría ser una distribución estacionaria. Esto es sólo una posibilidad ya que los límites podrían ser todos cero. Sin embargo, en el caso finito tales límites conforman una distribución de probabilidad verdadera. La siguiente proposición refleja estas ideas.

Proposición 2.18 *Sea la CMTD con probabilidades de transición P_{ij} tales que los límites $\pi_j = \lim_{t\to\infty}(\boldsymbol{P}^t)_{ij}$ existen para cada j y no dependen de i. Entonces*

i) $\sum_{j\in S} \pi_j \leq 1$

ii) $\pi_j = \sum_{i\in S} \pi_i (\boldsymbol{P}^t)_{ij}$.

Cuando el espacio de estados es finito se cumple la igualdad en el inciso i) obteniéndose una distribución de probabilidad verdadera.

El siguiente resultado es el recíproco del anterior, supone la existencia de una distribución estacionaria para concluir que los límites de las probabilidades existen.

Teorema 2.19 *Dada una CMTD irreducible, aperiódica y con distribución estacionaria $\vec{\pi}$. Entonces para cualquier par de estados i y j, $\lim_{t\to\infty}(\boldsymbol{P}^t)_{ij} = \pi_j$.*

Finalmente establecemos las condiciones suficientes para la existencia del límite de las probabilidades de transición, asegurando además que se trata de una distribución de probabilidad.

Teorema 2.20 *Dada una CMTD irreducible, aperiódica y recurrente positiva (es decir, ergódica), las probabilidades límite* $\pi_j = \lim_{t\to\infty}(\boldsymbol{P}^t)_{ij} = \frac{1}{m_j}$ *existen y constituyen la única solución al sistema de ecuaciones*

$$\pi_j = \sum_{i\in S} \pi_i \, P_{ij}, \tag{2.6}$$

sujeto a las condiciones $\pi_j \geq 0$, *y* $\sum_{j\in S} \pi_j = 1$.

Resumiendo los resultados principales para la clasificación de $CMTD$ tenemos que:

1) Dada una $CMTD$ *aperiódica*, los límites $\vec{\pi}(\infty) = \lim_{t\to\infty} \vec{\pi}(t)$ existen.

2) Para cualquier $CMTD$ *irreducible* y *aperiódica* existe la distribución de probabilidad límite $\vec{\pi}(\infty)$ y es independiente de la distribución de probabilidad inicial $\vec{\pi}(0)$.

3) Para una $CMTD$ *ergódica*, la distribución de probabilidad límite $\vec{\pi}(\infty)$ existe y comprende el único vector de probabilidades estacionarias $\vec{\pi}$.

2.2. Descripción del modelo

Como hemos mencionado en el Capítulo 1 cuando nos referimos a las polémicas sobre Facebook, en cualquier contexto comunicacional lo que realmente interesa es estudiar el comportamiento de los usuarios. En Facebook estos usuarios son representados por los perfiles y por lo tanto resulta difícil saber cuántas personas hay detrás de un perfil o cuántos perfiles tiene una misma persona. Hemos señalado que estos fenómenos contrapuestos generan un modelo de red en "dos niveles" que es fuertemente "no-identificable", y en el que el interés se centra en el "nivel inferior" representado por las personas o usuarios a partir de lo observable que es el "nivel superior" y está representado por los perfiles.

Comenzaremos por definir algunos conjuntos y funciones que utilizaremos para describir las ideas mencionadas.

Trabajaremos a tiempo discreto t, es decir, t será un número natural o cero. Llamaremos $\mathcal{P}_t$ al conjunto de usuarios de internet en el instante t y $\mathcal{F}_t$ al conjunto de perfiles de Facebook en el instante t. Es claro que $\mathcal{P}_t \subseteq \mathcal{P}_{t+1}$. Además, supondremos que una vez que un perfil es creado en Facebook no puede darse de baja ya que borrar toda la información que el mismo tiene incorporada en la red, nombre, fotos, videos, mensajes, dirección de e-mail, etc., no es tan fácil como abrir una cuenta y, por ser ésta una opción definitiva, está bastante oculta en Facebook, luego $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$.

Notaremos como $\mathcal{P}_\infty$ y $\mathcal{F}_\infty$ a los conjuntos $\mathcal{P}_\infty = \bigcup_{t=0}^{\infty} \mathcal{P}_t$ y $\mathcal{F}_\infty = \bigcup_{t=0}^{\infty} \mathcal{F}_t$ respectivamente.

Para definir el instante en que nace un perfil $f \in \mathcal{F}_\infty$, haremos uso de un conjunto de variables aleatorias independientes e idénticamente distribuidas $\{\xi_t(f)\}$ con $E(\xi_t(f)) = \mu$, $E(\xi_t^2(f)) = \nu^2$ y soporte no acotado, es decir, $P(\xi_t(f) < \lambda) > 0$, para todo λ.

Definición 2.21 *Sea $\delta_N > 0$, definimos el tiempo de nacimiento de un perfil $f \in \mathcal{F}_\infty$ como el menor instante $s \in \mathbb{N}$ tal que $\xi_s > \delta_N$. Esto equivale a que $f \notin \mathcal{F}_t$, $f \in \mathcal{F}_s$, para todo $s \geq t+1 \Leftrightarrow \xi_{t+1}(f) > \delta_N$.*

Es decir, a tiempo t el perfil f todavía no ha sido creado, pero se creará a tiempo s, posterior a t, si y sólo si la variable ξ_t supera cierto umbral establecido δ_N.

Observación 2.22 *Para todo $f \in \mathcal{F}_\infty$, existe una variable aleatoria $t(f,w)$ tal que $f \notin \mathcal{F}_{t(f,w)}$, $f \in \mathcal{F}_s$, para todo $s \geq t(f,w)+1$. Esto es, para todo perfil por nacer existe un tiempo que depende de dicho perfil y es aleatorio, tal que para todo tiempo posterior, el perfil será creado.*

Esta forma de definir el nacimiento de un perfil nos permitirá modelar la creación de los perfiles y el crecimiento de la red desde una perspectiva aleatoria.

Dados los conjuntos de usuarios y de perfiles de Facebook en el instante t, $\mathcal{P}_t$ y $\mathcal{F}_t$ respectivamente, notaremos con $\mathcal{P}_t^{(k)}$ y $\mathcal{F}_t^{(k)}$, $k \in \mathbb{N}$, a la familia de todos los subconjuntos con k elementos de $\mathcal{P}_t$ y $\mathcal{F}_t$ respectivamente. Además, notaremos con $\mathcal{P}_t^*$ y $\mathcal{F}_t^*$ a los conjuntos $\mathcal{P}_t^* = \bigcup_{k=1}^{\infty} \mathcal{P}_t^{(k)} \cup \emptyset$ y $\mathcal{F}_t^* = \bigcup_{k=1}^{\infty} \mathcal{F}_t^{(k)} \cup \emptyset$. Luego, un elemento de $\mathcal{P}_t^*$ o es el conjunto vacío o es un subconjunto de usuarios de internet con k elementos en el instante t, donde k es un número natural. Análogamente, un elemento de $\mathcal{F}_t^*$ o es el conjunto vacío o es un subconjunto de perfiles de Facebook con k elementos en el instante t, con k un número natural.

Vamos a definir ahora las funciones que modelarán los "dos niveles" en que será representada la red.

Definición 2.23 *Definimos la función que a cada usuario de internet asocia el conjunto de perfiles que administra en el instante t, como la función $\varphi_t : \mathcal{P}_t \to \mathcal{F}_t^*$ dada por:*

$$\varphi_t(p) = \begin{cases} \emptyset, & \text{si } p \text{ no tiene perfil de Facebook,} \\ \{f_1, \ldots, f_k\} & \text{si } f_1, \ldots, f_k \text{ son los perfiles de Facebook que administra } p. \end{cases}$$

Definición 2.24 *Definimos la función que a cada perfil de Facebook asigna los usuarios que lo co-administran en el instante t, como la función $\mu_t : \mathcal{F}_t \to \mathcal{P}_t^*$ dada por:*

$$\mu_t(f) = \{p_1, \ldots, p_n\}, \text{ con } p_1, \ldots, p_n \text{ los usuarios que co-administran el perfil } f.$$

Trivialmente se verifican las siguientes Ecuaciones de Consistencia:

1) Si $\varphi_t(p) \neq \emptyset \Rightarrow p \in \mu_t(\varphi_t(p))$.

2) Para todo $f \in \mathcal{F}_t, f \in \varphi_t(\mu_t(f))$.

Debido a que nuestro análisis se centrará en el comportamiento de los perfiles, definiremos la siguiente función que nos permitirá estudiar la dinámica entre pares de perfiles.

Definición 2.25 *Sean $f, g \in \mathcal{F}_t$. Definimos la función aleatoria de "amistad" en el instante t como la función $\alpha_t : \mathcal{F}_t \times \mathcal{F}_t \to \{0, 1\}$, tal que:*

$$\alpha_t(f,g) = \begin{cases} 1, & \textit{si } f \textit{ y } g \textit{ son amigos en el instante } t, \\ 0, & \textit{si } f \textit{ y } g \textit{ no son amigos en el instante } t. \end{cases}$$

Observación 2.26 *La amistad es una relación simétrica, por lo que la función α_t es simétrica.*

Nuestro siguiente objetivo consiste en estudiar el comportamiento de todos los posibles pares de perfiles que integran la red. Para ello, definiremos una matriz en la que cada elemento refleje el comportamiento del correspondiente par de perfiles que representa.

Notaremos con $M = card\{\mathcal{F}_\infty\}$ al cardinal del conjunto de perfiles de Facebook a tiempo infinito y con $\mathcal{M}_{M\times M}$ al conjunto de las matrices binarias simétricas de orden M.

Definición 2.27 *Definimos la matriz aleatoria de "amistad" en el instante t como la matriz $\mathcal{A}_t \in \mathcal{M}_{M\times M}$ cuyos elementos son los valores que toma la función α_t para cada par de perfiles $(f, g) \in \mathcal{F}_\infty \times \mathcal{F}_\infty$. Es decir,*

$$\mathcal{A}_t = \Big(\alpha_t(f,g)\Big)_{(f,g)\in\mathcal{F}_\infty\times\mathcal{F}_\infty}.$$

Propiedades de $\mathcal{A}_t$

- $\mathcal{A}_t$ es simétrica, ya que la función α_t lo es.
- Los elementos diagonales tomarán el valor 1 si a tiempo t el perfil correspondiente ya ha sido creado y 0 si no. Es decir,

$$\alpha_t(f,f) = \begin{cases} 1, & \text{para todo } f \in \mathcal{F}_t, \\ 0, & \text{si } f \in \mathcal{F}_\infty - \mathcal{F}_t, \end{cases}$$

  con lo cual a tiempo $t \approx \infty$ todo perfil ya ha sido creado.

- Debido a que en redes sociales existe un importante nivel de interactividad, es habitual suponer markovianidad. Luego, supondremos que $\{\mathcal{A}_t\}$ es una cadena de Markov matricial finita de dimensión $M \times M$. Esto significa que para conocer el comportamiento de la cadena $\{\mathcal{A}_t\}$ en el estado actual, basta conocer su comportamiento en el instante anterior.

Calcularemos para la cadena $\{\mathcal{A}_t\}$, las probabilidades de transición en un paso. Para ello, tendremos en cuenta que al ser $\mathcal{A}_t$ simétrica, bastará con calcular las probabilidades de transición de los elementos correspondientes a la subdiagonal inferior de la misma.

Definición 2.28 *Dado $\mathcal{F}_\infty = \{f_1, \ldots, f_M\}$, si $f_i, f_j \in \mathcal{F}_\infty$, decimos que f_i precede a f_j si $i < j$ y escribimos $f_i \prec f_j$.*

Introducimos a continuación las siguientes notaciones de utilidad para el cálculo de las probabilidades mencionadas:

- Si $\mathcal{A}_t = \mathbb{A}$, entonces el estado de amistad de la cadena $\{\mathcal{A}_t\}$ a tiempo t está dado por $\mathbb{A}$.
- Si $\mathbb{A} \in \mathcal{M}_{M\times M}$, $SD(\mathbb{A}) = \{\mathbb{A}_{i,j} : i > j\}$ representa el conjunto de elementos de la subdiagonal inferior de $\mathbb{A}$.

Luego, dados dos estados de $\{\mathcal{A}_t\}$, $\mathbb{A}$ y $\mathbb{B}$, queremos calcular las probabilidades de transición en un paso de $\mathbb{A}$ a $\mathbb{B}$, calculando las probabilidades de que la subdiagonal inferior de la cadena a tiempo $t+1$ sea la subdiagonal inferior de $\mathbb{B}$ dado que a tiempo t el estado correspondiente a la subdiagonal inferior de la cadena sea la subdiagonal inferior de $\mathbb{A}$. Es decir,

$$\begin{aligned} p_{t,t+1}^{\mathbb{A},\mathbb{B}} &= P(\mathcal{A}_{t+1} = \mathbb{B}/\mathcal{A}_t = \mathbb{A}) \\ &= P\left[SD(\mathcal{A}_{t+1}) = SD(\mathbb{B})/SD(\mathcal{A}_t) = SD(\mathbb{A})\right]. \end{aligned} \tag{2.7}$$

Pueden darse las siguientes posibilidades para los elementos de la subdiagonal inferior en la transición de $\mathbb{A}$ a $\mathbb{B}$:

- $\mathbb{A}_{i,j} = 0$ y $\mathbb{B}_{i,j} = 0$, caso en el que los perfiles no son amigos a tiempo t y no se hacen amigos en el siguiente instante.
- $\mathbb{A}_{i,j} = 0$ y $\mathbb{B}_{i,j} = 1$, caso en el que los perfiles no son amigos en el instante t pero pasan a serlo en el siguiente instante.
- $\mathbb{A}_{i,j} = 1$ y $\mathbb{B}_{i,j} = 0$, caso en el que los perfiles son amigos a tiempo t pero rompen amistad en el siguiente instante.

- $\mathbb{A}_{i,j} = 1$ y $\mathbb{B}_{i,j} = 1$, caso en el que los perfiles son amigos a tiempo t y siguen siéndolo en el instante siguiente.

Observación 2.29 *Hemos supuesto que una vez que un perfil es creado en Facebook no puede darse de baja. Luego, si $\mathbb{A}_{i,i} = 1$ y $\mathbb{B}_{i,i} = 0$, para algún i, la probabilidad de transición en un paso de $\mathbb{A}$ a $\mathbb{B}$ es cero.*

Para describir cómo cambian de un instante al siguiente los elementos de $SD(\mathcal{A}_t)$, definiremos ciertos índices que medirán el nivel de afinidad entre pares y ternas de perfiles.

Sea A un conjunto cualquiera. Notaremos por

$$A \times A - D(A \times A) = \{(a, a') : a, a' \in A, a \neq a'\},$$

al conjunto de pares de elementos de A distintos, y por

$$A \times A \times A - D(A \times A \times A) = \{(a, a', a'') : a, a', a'' \in A, a \neq a', a \neq a'', a' \neq a''\},$$

al conjunto de ternas de elementos de A distintos dos a dos.

Definición 2.30 *Sea $\mathcal{P}_\infty$ el conjunto de usuarios a tiempo infinito, y sean $p, p' \in \mathcal{P}_\infty$, con $p \neq p'$. Definimos el "índice de imagen" del usuario p sobre el usuario p' como la función $X : \mathcal{P}_\infty \times \mathcal{P}_\infty - D(\mathcal{P}_\infty \times \mathcal{P}_\infty) \to \mathbb{R}$ tal que:*

$$\begin{cases} X(p, p') > 0, & \text{si } p \text{ tiene una imagen favorable de } p', \\ X(p, p') = 0, & \text{si } p \text{ es indiferente a } p', \\ X(p, p') < 0, & \text{si } p \text{ tiene una imagen desfavorable de } p'. \end{cases}$$

Es decir, X asigna a cada par de usuarios distintos un número positivo, un cero o un número negativo según la imagen que tenga el primero con respecto al segundo sea favorable, indiferente o desfavorable respectivamente.

Observaciones:

1) Suponemos que la red se encuentra en la etapa de madurez, instancia en la que la imagen que tiene una persona sobre otra a largo plazo ya está formada, razón por la cual X no depende de t.

2) La imagen que tiene un usuario p con respecto a otro p' no necesariamente coincide con la que tiene p' sobre p, es decir, X es no simétrica.

3) Como los usuarios no se pueden observar si no a través de los perfiles, X es no observable.

4) Cuanto más "favorable" ("desfavorable") sea la imagen que tenga p sobre p', mayor (menor) va a ser el valor de $X(p, p')$, es decir, existe monotonía en la imagen en relación a X.

Dados dos perfiles distintos, $f, g \in \mathcal{F}_\infty$, vamos a definir el índice de imagen de f sobre g promediando los índices de imagen de todos los pares posibles de usuarios que los administran, transformados por una función de regresión monótona desconocida, junto con un ruido blanco entre los perfiles, que es aleatorio, depende del tiempo y es independiente del ruido blanco entre cualquier otro par de perfiles distinto de la siguiente manera:

Definición 2.31 *Sean $f, g \in \mathcal{F}_\infty$, $f \neq g$. Definimos el "índice de imagen" del perfil f sobre el perfil g como la función*

$$Y_t : \mathcal{F}_\infty \times \mathcal{F}_\infty - D(\mathcal{F}_\infty \times \mathcal{F}_\infty) \to I\!R,$$

dada por

$$Y_t(f, g) = \frac{1}{m * l} \sum_{i=1}^{m} \sum_{j=1}^{l} \Phi\left(X(p_i, p'_j)\right) + \varepsilon_t(f, g), \tag{2.8}$$

con $\{p_1, \ldots, p_m\}$ y $\{p'_1, \ldots, p'_l\}$, los conjuntos de usuarios que administran a los perfiles f y g respectivamente, $\Phi : I\!R \to I\!R$ una función de regresión monótona, desconocida, y $\{\varepsilon_t(f, g)\}_{t=0,1,\ldots}$ una sucesión de ruido blanco entre f y g, con $\varepsilon_t(f, g)$ independiente del ruido blanco entre otro par de perfiles distinto.

Ahora definiremos los índices de imagen de ternas ordenadas de usuarios y de perfiles.

Definición 2.32 *Sean $p, p', p'' \in \mathcal{P}_\infty$. Definimos el "índice de imagen" de la terna ordenada (p, p', p'') como la función $U : \mathcal{P}_\infty \times \mathcal{P}_\infty \times \mathcal{P}_\infty - D(\mathcal{P}_\infty \times \mathcal{P}_\infty \times \mathcal{P}_\infty) \to I\!R$ que asigna a cada terna ordenada de usuarios distintos dos a dos un número real, que representa el grado de aceptación de la acción: p le sugiere a p' el amigo p''.*

Observaciones:

1) Como hemos mencionado, suponemos que la red se encuentra en la etapa de madurez en la cual las afinidades entre los usuarios ya están formadas, razón por la cual al igual que X, U no depende de t.

2) Cada usuario en la terna tiene un rol determinado, por lo tanto U es no simétrica.

3) U es no observable, por la misma razón que no lo es X.

Además, dados tres perfiles distintos dos a dos f, g y h, definiremos el índice de imagen de la terna ordenada (f, g, h) promediando los índices de imagen de todas las posibles ternas ordenadas de usuarios que los administran, transformados por una función de regresión monótona desconocida, junto con un ruido blanco entre los perfiles que es aleatorio, depende del tiempo y es independiente del ruido blanco entre cualquier otra terna ordenada de perfiles distintos dos a dos distinta, de la siguiente manera:

Definición 2.33 *Sean $f, g, h \in \mathcal{F}_\infty$ distintos dos a dos. Definimos el "índice de imagen" de la terna ordenada (f, g, h) como la función*

$$W_t : \mathcal{F}_\infty \times \mathcal{F}_\infty \times \mathcal{F}_\infty - D(\mathcal{F}_\infty \times \mathcal{F}_\infty \times \mathcal{F}_\infty) \to I\!R,$$

dada por

$$W_t(f, g, h) = \frac{1}{m * k * l} \sum_{i=1}^{m} \sum_{j=1}^{k} \sum_{r=1}^{l} \Psi\left(U(p_i, p'_j, p''_r)\right) + \eta_t(f, g, h), \tag{2.9}$$

con $\{p_1, \ldots, p_m\}$, $\{p'_1, \ldots, p'_k\}$ y $\{p''_1, \ldots, p''_l\}$, los conjuntos de usuarios que administran a los perfiles f, g y h respectivamente, $\Psi : I\!R \to I\!R$ una función de regresión monótona, desconocida, y $\{\eta_t(f, g, h)\}_{t=0,1,\ldots}$ una sucesión de ruido blanco entre f, g y h, con $\eta_t(f, g, h)$ independiente del ruido blanco entre otra terna ordenada de perfiles distintos dos a dos distinta.

Entonces, siendo $\mathbb{A}$ y $\mathbb{B} \in \mathcal{M}_{M\times M}$ dos estados de la cadena $\{\mathcal{A}_t\}$ y, dados $f, g, h \in \mathcal{F}_\infty$, con el objetivo de calcular las probabilidades de transición de $\mathbb{A}$ a $\mathbb{B}$ definidas por (2.7), describiremos las acciones que puede provocar el perfil particular f e inciden en dicha transición a partir de los índices de imagen (2.8) y (2.9). Para ello, supondremos que existen ciertos números a los que llamaremos "umbrales" $\delta_B > 0$, $\delta_R > 0$ y $\delta_I > 0$, tales que se verifican las siguientes condiciones:

i) "f *provoca la ruptura* de amistad con g" si y sólo si el índice de imagen de f sobre g, $Y_t(f, g)$, o el índice de imagen de g sobre f, $Y_t(g, f)$, resulta negativo e inferior al umbral $-\delta_B$.

Observemos que la ruptura de amistad de f con g puede darse porque f decide eliminar a g de su lista de amigos o porque cierto comportamiento de f provoca que g lo elimine de sus amigos.

ii) "f *solicita con éxito* amistad a g" si y sólo si el índice de imagen de f sobre g, $Y_t(f, g)$, supera el umbral δ_R y el índice de imagen de g sobre f, $Y_t(g, f)$, supera el umbral δ_A.

La solicitud de amistad con éxito se produce cuando f solicita amistad a g y g acepta como amigo a f.

iii) "f *sugiere con éxito* a g, el amigo h" si y sólo si el índice de imagen de la terna ordenada de perfiles (f, g, h), $W_t(f, g, h)$, supera el umbral δ_I.

La sugerencia de amistad con éxito viene dada cuando f le sugiere a g su amigo h, g le solicita amistad a h y h acepta ser amigo de g.

Observación 2.34 *Parece razonable que $\delta_R > \delta_I$, esto es, el umbral que debe superar el índice de imagen necesario para que un perfil solicite amistad a otro en forma directa debería ser más alto que el umbral que debe superar el índice de imagen de la terna ordenada de perfiles correspondiente a cuando este mismo perfil solicita amistad a otro por sugerencia de un amigo.*

Luego, para el estudio de las acciones del perfil f que inciden en la transición de $\mathbb{A}$ a $\mathbb{B}$, debemos considerar sus acciones o intervenciones con o sin efecto en relación con todos los demás perfiles de la red, y las mismas pueden resumirse en una unión disjunta de los siguientes eventos:

$D_t^1(f)$: "*rupturas de f*",
$D_t^2(f)$: "*no rupturas de f*",
$D_t^3(f)$: "*solicitudes de amistad de f*",
$D_t^4(f)$: "*sugerencias de amistad de f*",
es decir,

$$I_t(f) = \bigcup_{i=1}^{4} D_t^i(f). \tag{2.10}$$

Calcularemos en primer lugar las probabilidades de los eventos $D_t^1(f)$ y $D_t^2(f)$, para esto expresamos a los mismos en términos de la condición *i)*. De este modo, $D_t^1(f)$ representa la intersección sobre todos los perfiles g que preceden al perfil f tales que f y g son amigos a tiempo t pero dejan de serlo en el instante siguiente porque el índice de imagen de f sobre g o de g sobre f resulta negativo e inferior al umbral $-\delta_B$, es decir,

$$D_t^1(f) = \bigcap_{\substack{g \prec f: \\ \mathbb{A}_{f,g}=1, \\ \mathbb{B}_{f,g}=0}} \{Y_t(f, g) < -\delta_B\} \cup \{Y_t(g, f) < -\delta_B\},$$

y $D_t^2(f)$ representa la intersección sobre todos los perfiles g que preceden a f tales que f y g son amigos a tiempo t y siguen siéndolo en el instante siguiente debido a que tanto el índice de imagen de f sobre g como el de g sobre f no alcanzó a ser tan bajo como para que se produjera la ruptura, es decir,

$$D_t^2(f) = \bigcap_{\substack{g \prec f: \\ \mathbb{A}_{f,g}=1, \\ \mathbb{B}_{f,g}=1}} \{Y_t(f, g) \geq -\delta_B,\ Y_t(g, f) \geq -\delta_B\}.$$

Entonces, si $g, g' \in \mathcal{F}_\infty$, con $g \prec f$, $g' \prec f$, $g \neq g'$, son independientes $\{Y_t(f,g) < -\delta_B\}$ de $\{Y_t(f,g') < -\delta_B\}$, $\{Y_t(g,f) < -\delta_B\}$ de $\{Y_t(g',f) < -\delta_B\}$, $\{Y_t(f,g) \geq -\delta_B\}$ de $\{Y_t(f,g') \geq -\delta_B\}$ y $\{Y_t(g,f) \geq -\delta_B\}$ de $\{Y_t(g',f) \geq -\delta_B\}$. Esto se debe a la independencia entre los ruidos blancos $\varepsilon_t(f,g)$ y $\varepsilon_t(f,g')$ y entre los ruidos blancos $\varepsilon_t(g,f)$ y $\varepsilon_t(g',f)$, correspondientes a cada índice de imagen. En consecuencia, usando propiedades de conjuntos, tenemos que:

$$P(D_t^1(f)) = \prod_{\substack{g \prec f: \\ \mathbb{A}_{f,g}=1, \\ \mathbb{B}_{f,g}=0}} 1 - P(Y_t(f,g) \geq -\delta_B) \cdot P(Y_t(g,f) \geq -\delta_B),$$

y

$$P(D_t^2(f)) = \prod_{\substack{g \prec f: \\ \mathbb{A}_{f,g}=1, \\ \mathbb{B}_{f,g}=1}} P(Y_t(f,g) \geq -\delta_B) \cdot P(Y_t(g,f) \geq -\delta_B).$$

Para el cálculo de la probabilidad de $D_t^3(f)$ expresamos a dicho evento en función de la condición *ii)*, siendo entonces $D_t^3(f)$ la unión sobre todos los perfiles g que preceden a f tales que f y g no son amigos a tiempo t pero pasan a serlo o no en el instante siguiente. Se hacen amigos cuando el índice de imagen de f sobre g supera el umbral de afinidad δ_R y el índice de imagen de g sobre f supera el umbral de aceptación δ_A, y no se hacen amigos debido a que, aunque f solicite amistad a g, el índice de imagen de g sobre f no supera el umbral de aceptación δ_A, es decir,

$$D_t^3(f) = \bigcup_{\substack{g \prec f: \\ \mathbb{A}_{f,g}=0, \\ \mathbb{B}_{f,g}=1}} \{Y_t(f,g) > \delta_R, Y_t(g,f) > \delta_A\} \quad \cup \bigcup_{\substack{g \prec f: \\ \mathbb{A}_{f,g}=0, \\ \mathbb{B}_{f,g}=0}} \{Y_t(f,g) > \delta_R, Y_t(g,f) \leq \delta_A\}.$$

Luego, utilizando propiedades de conjuntos tenemos que

$$P(D_t^3(f)) = 1 - P\Bigg(\bigcap_{\substack{g \prec f: \\ \mathbb{A}_{f,g}=0, \\ \mathbb{B}_{f,g}=1}} \{Y_t(f,g) > \delta_R, Y_t(g,f) > \delta_A\}^c \Bigg) \times$$
$$\times P\Bigg(\bigcap_{\substack{g \prec f: \\ \mathbb{A}_{f,g}=0, \\ \mathbb{B}_{f,g}=0}} \{Y_t(f,g) > \delta_R, Y_t(g,f) \leq \delta_A\}^c \Bigg),$$

ya que si $g' \in \{g \prec f : \mathbb{A}_{f,g} = 0,\ \mathbb{B}_{f,g} = 1\}$ entonces $g' \notin \{g \prec f : \mathbb{A}_{f,g} = 0,\ \mathbb{B}_{f,g} = 0\}$. Además, $\{Y_t(f,g) \leq \delta_R,\ Y_t(g,f) \leq \delta_A\}^c$ y $\{Y_t(f,g') \leq \delta_R,\ Y_t(g',f) \leq \delta_A\}^c$ son independientes y $\{Y_t(f,g) \leq \delta_R,\ Y_t(g,f) > \delta_A\}^c$ y $\{Y_t(f,g') \leq \delta_R,\ Y_t(g',f) > \delta_A\}^c$ también, ya que son independientes los ruidos blancos $\varepsilon_t(f,g)$ y $\varepsilon_t(f,g')$ correspondientes a los índices

de imagen $Y_t(f,g)$ e $Y_t(f,g')$ respectivamente. Por lo tanto,

$$P(D_t^3(f)) = 1 - \prod_{\substack{g \prec f: \\ \mathbb{A}_{f,g}=0, \\ \mathbb{B}_{f,g}=1}} [1 - P(Y_t(f,g) > \delta_R)\, P(Y_t(g,f) > \delta_A)] \times$$
$$\times \prod_{\substack{g \prec f: \\ \mathbb{A}_{f,g}=0, \\ \mathbb{B}_{f,g}=0}} [1 - P(Y_t(f,g) > \delta_R)\, P(Y_t(g,f) \leq \delta_A)].$$

Finalmente calcularemos la probabilidad del evento $D_t^4(f)$ que, expresado en términos de la condición *iii)*, representa la intersección sobre todos los perfiles g que preceden a f tales que f y g son amigos en el instante t y en el instante siguiente h y g se hacen amigos o no. Se hacen amigos cuando el índice de imagen de la terna ordenada (f,g,h) supera el umbral δ_I necesario para que g y h se hagan amigos por sugerencia de f y no se hacen amigos cuando ese índice no supera el umbral δ_I, es decir,

$$D_t^4(f) = \bigcap_{\substack{g \prec f: \\ \mathbb{A}_{f,g}=1}} \left(\bigcup_{\substack{h \prec f: \\ \mathbb{A}_{f,h}=1, \\ \mathbb{B}_{g,h}=1}} \{W_t(f,g,h) > \delta_I\} \quad \cup \bigcup_{\substack{h \prec f: \\ \mathbb{A}_{f,h}=1, \\ \mathbb{B}_{g,h}=0}} \{W_t(f,g,h) \leq \delta_I\} \right)$$

y

$$\begin{aligned} P(D_t^4(f)) &= P\Big[\bigcap_{\substack{g \prec f: \\ \mathbb{A}_{f,g}=1}} \left(\bigcup_{\substack{h \prec f: \\ \mathbb{A}_{f,h}=1, \\ \mathbb{B}_{g,h}=1}} \{W_t(f,g,h) > \delta_I\} \quad \cup \bigcup_{\substack{h \prec f: \\ \mathbb{A}_{f,h}=1, \\ \mathbb{B}_{g,h}=0}} \{W_t(f,g,h) \leq \delta_I\} \right) \Big] \\ &= \prod_{\substack{g \prec f: \\ \mathbb{A}_{f,g}=1}} P\Big(\bigcup_{\substack{h \prec f: \\ \mathbb{A}_{f,h}=1, \\ \mathbb{B}_{g,h}=1}} \{W_t(f,g,h) > \delta_I\} \quad \cup \bigcup_{\substack{h \prec f: \\ \mathbb{A}_{f,h}=1, \\ \mathbb{B}_{g,h}=0}} \{W_t(f,g,h) \leq \delta_I\} \Big) \\ &= \prod_{\substack{g \prec f: \\ \mathbb{A}_{f,g}=1}} \Big[1 - P\Big(\bigcap_{\substack{h \prec f: \\ \mathbb{A}_{f,h}=1, \\ \mathbb{B}_{g,h}=1}} \{W_t(f,g,h) \leq \delta_I\} \cap \bigcap_{\substack{h \prec f: \\ \mathbb{A}_{f,h}=1, \\ \mathbb{B}_{g,h}=0}} \{W_t(f,g,h) > \delta_I\} \Big) \Big], \end{aligned} \tag{2.11}$$

donde (2.11) se obtiene utilizando propiedades de conjuntos.

Además, $\bigcap_{\substack{h \prec f: \\ \mathbb{A}_{f,h}=1, \\ \mathbb{B}_{g,h}=1}} \{W_t(f,g,h) \leq \delta_I\}$ es independiente de $\bigcap_{\substack{h \prec f: \\ \mathbb{A}_{f,h}=1, \\ \mathbb{B}_{g,h}=0}} \{W_t(f,g,h) > \delta_I\}$, ya que si $h' \in \{h \prec f : \mathbb{A}_{f,h} = 1,\ \mathbb{B}_{g,h} = 1\}$ entonces $h' \notin \{h \prec f : \mathbb{A}_{f,h} = 1,\ \mathbb{B}_{g,h} = 0\}$. Por otro lado, para f y g fijos, $h \neq h'$, $h \prec f$, $h' \prec f$, son independientes $\{W_t(f,g,h) \leq \delta_I\}$ de $\{W_t(f,g,h') \leq \delta_I\}$ y $\{W_t(f,g,h) > \delta_I\}$ de $\{W_t(f,g,h') > \delta_I\}$, ya que son independientes los ruidos blancos $\eta_t(f,g,h)$ y $\eta_t(f,g,h')$ correspondientes a los índices de imagen

$W_t(f,g,h)$ y $W_t(f,g,h')$ respectivamente. Por lo tanto, la expresión (2.11) para el cálculo de la probabilidad de $D_t^4(f)$ resulta

$$P(D_t^4(f)) = \prod_{\substack{g \prec f: \\ \mathbb{A}_{f,g}=1}} \Bigg[1 - \prod_{\substack{h \prec f: \\ \mathbb{A}_{f,h}=1, \\ \mathbb{B}_{g,h}=1}} P(W_t(f,g,h) \leq \delta_I) \prod_{\substack{h \prec f: \\ \mathbb{A}_{f,h}=1, \\ \mathbb{B}_{g,h}=0}} P(W_t(f,g,h) > \delta_I)\Bigg].$$

Retomando el cálculo de las probabilidades de transición en un paso del estado de amistad $\mathbb{A}$ al $\mathbb{B}$ de la cadena de Markov representada por la matriz de "amistad" $\mathcal{A}_t$, $p_{t,t+1}^{\mathbb{A},\mathbb{B}}$, (2.7) puede expresarse como la probabilidad de la intersección sobre todos los perfiles $f \in \mathcal{F}_\infty$, de las acciones o intervenciones de cada uno de ellos que inciden en la transición de $\mathbb{A}$ a $\mathbb{B}$, siendo estas acciones representadas por la unión disjunta de los eventos descriptos $D_t^i(f)$, $i = 1,2,3,4$, es decir,

$$p_{t,t+1}^{\mathbb{A},\mathbb{B}} = P\Bigg(\bigcap_{f \in \mathcal{F}_\infty} I_t(f)\Bigg), \tag{2.12}$$

con $I_t(f) = \bigcup_{i=1}^{4} D_t^i(f)$.

Luego, dado $f' \in \mathcal{F}_\infty$, $f \neq f'$, veremos que $I_t(f')$ es independiente de $I_t(f)$. Como $I_t(f')$ e $I_t(f)$ son uniones disjuntas de los conjuntos $D_t^i(f')$ y $D_t^i(f)$ respectivamente, $i = 1,2,3,4$, basta ver que $D_t^i(f')$ es independiente de $D_t^i(f)$, para $i = 1,2,3,4$. En efecto, en $D_t^i(f')$, la primera coordenada tanto del índice $Y_t(f',g)$ como del índice $W_t(f',g,h)$ es fija, y es además la primera coordenada de los ruidos blancos $\varepsilon_t(f',g)$ y $\eta_t(f',g,h)$ que intervienen en dichos índices respectivamente. Entonces, son independientes $Y_t(f',g)$ de $Y_t(f,g)$ y $W_t(f',g,h)$ de $W_t(f,g,h)$, puesto que son independientes los ruidos blancos $\varepsilon_t(f',g)$ de $\varepsilon_t(f,g)$ y $\eta_t(f',g,h)$ de $\eta_t(f,g,h)$. Por lo tanto, para $i = 1,2,3,4$, $D_t^i(f')$ es independiente de $D_t^i(f)$ y la probabilidad (2.12) resulta

$$p_{t,t+1}^{\mathbb{A},\mathbb{B}} \quad = \quad \prod_{f \in \mathcal{F}_\infty} P(I_t(f)), \tag{2.13}$$

con $I_t(f) = \bigcup_{i=1}^{4} D_t^i(f)$, por lo que aplicando el principio de inclusión-exclusión para los eventos $D_t^i(f)$, $i = 1,2,3,4$, podemos expresar a (2.13) como

$$\begin{aligned} p_{t,t+1}^{\mathbb{A},\mathbb{B}} \quad = \quad & \prod_{f \in \mathcal{F}_\infty} \Bigg(\sum_{i=1}^{4} P(D_t^i(f)) - \sum_{i=1}^{3} \sum_{j=i+1}^{4} P(D_t^i(f) \cap D_t^j(f)) + \\ & + \quad \sum_{i=1}^{2} \sum_{j=i+1}^{3} \sum_{k=j+1}^{4} P(D_t^i(f) \cap D_t^j(f) \cap D_t^k(f)) - P\Bigg(\bigcap_{i=1}^{4} D_t^i(f)\Bigg)\Bigg). \end{aligned} \tag{2.14}$$

Veamos finalmente que los eventos $D_t^i(f)$, $i = 1, 2, 3, 4$, son independientes entre sí, para un mismo perfil f.

$D_t^4(f)$ es independiente de $D_t^1(f)$, de $D_t^2(f)$ y de $D_t^3(f)$, ya que el ruido blanco $\eta_t(f, g, h)$ correspondiente al índice de imagen $W_t(f, g, h)$ que interviene en $D_t^4(f)$ es independiente del ruido blanco $\varepsilon_t(f, g)$ correspondiente al índice $Y_t(f, g)$ que interviene en $D_t^1(f)$, en $D_t^2(f)$ y en $D_t^3(f)$.

$D_t^1(f)$ es independiente de $D_t^2(f)$ y de $D_t^3(f)$, ya que si un perfil $f \in \mathcal{F}_\infty$ aporta a la intersección de $D_t^1(f)$ no aporta ni a la intersección de $D_t^2(f)$ ni a las uniones de $D_t^3(f)$. Por la misma razón $D_t^2(f)$ es independiente de $D_t^3(f)$.

Por lo tanto, la probabilidad de transición en un paso del estado $\mathbb{A}$ al $\mathbb{B}$ de la cadena de Markov $\{\mathcal{A}_t\}$ (2.14) está dada por:

$$\begin{aligned} p_{t,t+1}^{\mathbb{A},\mathbb{B}} &= \prod_{f \in \mathcal{F}_\infty} \Bigg(\sum_{i=1}^{4} P(D_t^i(f)) - \sum_{i=1}^{3} \sum_{j=i+1}^{4} P(D_t^i(f))\, P(D_t^j(f)) + \\ &+ \sum_{i=1}^{2} \sum_{j=i+1}^{3} \sum_{k=j+1}^{4} P(D_t^i(f))\, P(D_t^j(f))\, P(D_t^k(f)) - \prod_{i=1}^{4} P(D_t^i(f)) \Bigg). \end{aligned} \quad (2.15)$$

2.3. Transversalidad Completa

El concepto de *Transversalidad Completa*, (TC de aquí en más), en una red social está asociado a un modo de comportamiento comunicacional en el que cada perfil se relaciona con cualquier otro con la misma probabilidad, es decir, los perfiles se comunican unos con otros sin preferencias.

2.3.1. Homogeneidad de $\{\mathcal{A}_t\}$

En términos del modelo desarrollado en la Sección anterior, el comportamiento que describe una situación de TC, queda reflejado en el índice de imagen X de un usuario p sobre otro p' y en el índice de imagen U de una terna ordenada de usuarios (p, p', p''). De este modo, los promedios de las funciones de regresión monótonas desconocidas $\Phi(X)$ y $\Psi(U)$ en las expresiones de los índices de imagen de los perfiles que administran dichos usuarios (2.8) y (2.9), asumen el mismo valor, digamos $C_1 \in \mathbb{R}$ para todos los pares de usuarios distintos y $C_2 \in \mathbb{R}$ para todas las ternas ordenadas de usuarios distintos dos a dos respectivamente. Por consiguiente, los índices de imagen dados por las fórmulas (2.8) y (2.9) mencionadas se reducen a $C_1 + \varepsilon_t(f, g)$ y $C_2 + \eta_t(f, g, h)$.

Hemos visto en la Sección anterior que las probabilidades de transición en un paso del estado de amistad $\mathbb{A}$ al estado $\mathbb{B}$ de la cadena de Markov $\{\mathcal{A}_t\}$ pueden calcularse y están dadas por la expresión (2.15). A continuación veremos, suponiendo un contexto de TC,

que dichas probabilidades no dependen del tiempo t, es decir, probaremos el siguiente resultado.

Teorema 2.3.1 ***(Homogeneidad)***
En el contexto de TC, la cadena de Markov $\{\mathcal{A}_t\}$ *es homogénea en el tiempo.*

Demostración

Supongamos un contexto de TC en la red social Facebook. Para probar la homogeneidad en el tiempo de la cadena $\{\mathcal{A}_t\}$ basta ver que las probabilidades involucradas en (2.15), $P(D_t^i(f))$, $i = 1, 2, 3, 4$, no dependen de t. Para ello, introducimos algunas notaciones útiles.

Para el ruido blanco correspondiente a la fórmula (2.8), $\varepsilon_t(f, g)$, llamaremos F a la función de distribución de ε_t que, por hipótesis de TC, no dependerá de t ni de los perfiles f y g. Esto es, $P(\varepsilon_t(f, g) \leq \alpha) = F(\alpha)$, para todo $t = 0, 1, \ldots$, y para todo par de perfiles $f, g \in \mathcal{F}_\infty$, con F continua.

De igual manera, para el ruido blanco correspondiente a la fórmula (2.9), $\eta_t(f, g, h)$, llamaremos G a la función de distribución de η_t que, por hipótesis de TC, no dependerá de t ni de los perfiles f, g y h. Es decir, $P(\eta_t(f, g, h) \leq \alpha) = G(\alpha)$, para todo $t = 0, 1, \ldots$, y para toda terna ordenada de perfiles $f, g, h \in \mathcal{F}_\infty$, con G continua.

Además, notaremos con $c^{i,j}_{\mathbb{A},\mathbb{B}}(f)$ y $d^{i,j}_{\mathbb{A},\mathbb{B}}(f, g)$ a los cardinales de los conjuntos:

$$c^{i,j}_{\mathbb{A},\mathbb{B}}(f) = card\,\{g \prec f : \mathbb{A}_{f,g} = i, \mathbb{B}_{f,g} = j; \quad i, j \in \{0, 1\}\} \quad \text{y}$$
$$d^{1,j}_{\mathbb{A},\mathbb{B}}(f, g) = card\,\{h \prec f : \mathbb{A}_{f,h} = 1, \mathbb{B}_{h,g} = j; \quad j \in \{0, 1\}\}\,.$$

En consecuencia, bajo un contexto de TC, tenemos que:

$$\begin{aligned} P(D_t^1(f)) &= \prod_{\substack{g \prec f: \\ \mathbb{A}_{f,g}=1, \\ \mathbb{B}_{f,g}=0}} 1 - P\Big(\{C_1 + \varepsilon_t(f, g) < -\delta_B\}\Big)\, P\Big(\{C_1 + \varepsilon_t(g, f) < -\delta_B\}\Big) \\ &= \prod_{\substack{g \prec f: \\ \mathbb{A}_{f,g}=1, \\ \mathbb{B}_{f,g}=0}} 1 - [1 - F(-\delta_B - C_1)]^2 \\ &= \Big[F(-\delta_B - C_1)(2 - F(-\delta_B - C_1))\Big]^{c^{1,0}_{\mathbb{A},\mathbb{B}}(f)}, \end{aligned}$$

$$\begin{aligned} P(D_t^2(f)) &= \prod_{\substack{g \prec f: \\ \mathbb{A}_{f,g}=1, \\ \mathbb{B}_{f,g}=1}} \Big[1 - P\Big(\{C_1 + \varepsilon_t(f, g) < -\delta_B\}\Big)\Big]\Big[1 - P\Big(\{C_1 + \varepsilon_t(g, f) < -\delta_B\}\Big)\Big] \\ &= \Big[1 - F(-\delta_B - C_1)\Big]^{2\, c^{1,1}_{\mathbb{A},\mathbb{B}}(f)}, \end{aligned}$$

$$P(D_t^3(f)) = 1 - \prod_{\substack{g \prec f: \\ \mathbb{A}_{f,g}=0, \\ \mathbb{B}_{f,g}=1}} \left[1 - P\Big(\{C_1 + \varepsilon_t(f,g) > \delta_R\}\Big)\, P\Big(\{C_1 + \varepsilon_t(g,f) > \delta_A\}\Big)\right] \times$$
$$\times \prod_{\substack{g \prec f: \\ \mathbb{A}_{f,g}=0, \\ \mathbb{B}_{f,g}=0}} \left[1 - P\Big(\{C_1 + \varepsilon_t(f,g) > \delta_R\}\Big)\, P\Big(\{C_1 + \varepsilon_t(g,f) \leq \delta_A\}\Big)\right]$$
$$= 1 - \Big[F(-\delta_R - C_1) + F(-\delta_A - C_1) - F(-\delta_R - C_1)F(-\delta_A - C_1)\Big]^{c_{\mathbb{A},\mathbb{B}}^{0,1}(f)} \times$$
$$\times \Big[1 - F(-\delta_A - C_1) + F(-\delta_R - C_1)F(-\delta_A - C_1)\Big]^{c_{\mathbb{A},\mathbb{B}}^{0,0}(f)},$$

$$P(D_t^4(f)) = \prod_{\substack{g \prec f: \\ \mathbb{A}_{f,g}=1}} \left[1 - \prod_{\substack{h \prec f: \\ \mathbb{A}_{f,h}=1, \\ \mathbb{B}_{g,h}=1}} P\Big(\{C_2 + \eta_t(f,g,h) \leq \delta_I\}\Big) \prod_{\substack{h \prec f: \\ \mathbb{A}_{f,h}=1, \\ \mathbb{B}_{g,h}=0}} P\Big(\{C_2 + \eta_t(f,g,h) > \delta_I\}\Big)\right]$$
$$= \Big[1 - G(\delta_I - C_2)^{d_{\mathbb{A},\mathbb{B}}^{1,1}(f,g)}[1 - G(\delta_I - C_2)]^{d_{\mathbb{A},\mathbb{B}}^{1,0}(f,g)}\Big]^{c_{\mathbb{A},\mathbb{B}}^{1,1}(f) + c_{\mathbb{A},\mathbb{B}}^{1,0}(f)}.$$

Luego, como las probabilidades $P(D_t^i(f))$, $i = 1, 2, 3, 4$, no dependen de t, las probabilidades de transición en un paso para la cadena $\{\mathcal{A}_t\}$, $p_{t,t+1}^{\mathbb{A},\mathbb{B}}$, no dependen de t por lo que $\{\mathcal{A}_t\}$ es *homogénea* en el tiempo, esto es, la probabilidad de pasar de un estado a otro sólo depende del estado de salida y no del tiempo en que ocurre la transición. □

2.3.2. Ergodicidad de $\{\mathcal{A}_t\}$

Sea $S = \mathcal{M}_{M\times M}$, $M < \infty$, el espacio de estados de la cadena de Markov $\{\mathcal{A}_t\}$ representada por la matriz de "amistad" $\mathcal{A}_t$, donde $\mathcal{M}_{M\times M}$ es el espacio de matrices binarias simétricas de orden M definido anteriormente.

Supongamos un contexto de TC en Facebook y por lo tanto homogeneidad en el tiempo de la cadena $\{\mathcal{A}_t\}$. Queremos estudiar la dinámica a largo plazo de la red social cuando está en estado de madurez (descripción del "steady state" o "sistema en régimen") suponiendo que en esta instancia todos los perfiles han sido creados.

Notaremos con $\mathcal{C}$ al conjunto de estados pertenecientes a S representados por las matrices binarias de orden M con todos unos en la diagonal, es decir,

$$\mathcal{C} = \{\mathbb{A} \in S : \mathbb{A}_{i,i} = 1, \text{ para todo } i, i = 1, 2, \ldots, M\}.$$

Proposición 2.35 *$\mathcal{C}$ es una clase de comunicación cerrada.*

Demostración

Sea $\mathbb{A}' \in S$ tal que $\mathbb{A}' \notin \mathcal{C}$. $\mathbb{A}'$ es una matriz perteneciente a S con al menos un cero en la diagonal, es decir, $\mathbb{A}'_{i,i} = 0$, para algún i, $i = 1, 2, \ldots, M$. Esto significa que en el instante t, el estado de amistad $\mathbb{A}'$ está indicando que $\alpha_t(f_i, f_i) = 0$, para algún i, $i = 1, 2, \ldots, M$, es decir, el perfil f_i a tiempo t no existe en Facebook. Como hemos supuesto que una vez que un perfil ingresa a la red no puede darse de baja, dado $\mathbb{A} \in \mathcal{C}$, la probabilidad de transición de $\mathbb{A}$ a $\mathbb{A}'$ es cero, esto es, un estado fuera de $\mathcal{C}$ no es accesible desde un estado dentro de $\mathcal{C}$. Luego, $\mathcal{C}$ es una clase de comunicación cerrada. □

Observación 2.36 *Estamos suponiendo que la red está en estado de madurez y, en estas condiciones, todos los perfiles de la misma ya han sido creados. Luego, si $\mathbb{A}' \notin \mathcal{C}$ en el instante t, en una cantidad finita de pasos el estado $\mathbb{A}'$ será atraído por $\mathcal{C}$. Por lo tanto, $\{\mathcal{A}_t\}$ es tal que en el largo plazo todos los estados del espacio de estados S son atraídos en una cantidad finita de pasos por la clase de comunicación cerrada $\mathcal{C}$.*

Proposición 2.37 *$\mathcal{C}$ es una clase de comunicación irreducible y aperiódica.*

Demostración

Sean $\mathbb{A}, \mathbb{B} \in \mathcal{C}$. Ambos estados de amistad son matrices binarias simétricas de orden M con todos unos en la diagonal, Fuera de la diagonal, un cero puede pasar a un uno o un uno a un cero en un sólo paso y en consecuencia $\mathbb{A}$ y $\mathbb{B}$ están comunicados y son aperiódicos. Por lo tanto $\mathcal{C}$ es irreducible y aperiódica. □

Teorema 2.38 ***(Ergodicidad)***
En el contexto de TC la cadena de Markov $\{\mathcal{A}_t\}$ es ergódica y, bajo la distribución ergódica $\vec{\pi}(\infty)$, la indicatriz de amistad entre cualquier par de perfiles f y g, con $g \prec f$, denotada por $\alpha_\infty(f, g)$, es una variable aleatoria con distribución Bernoulli de parámetro p, $0 < p < 1$, con la misma distribución e independiente de la indicatriz de otro par de perfiles distinto.

Demostración

Hemos probado que $\mathcal{C}$ es una clase de comunicación cerrada, irreducible y aperiódica y, por ser $\mathcal{C}$ un conjunto finito, todos sus estados son recurrentes positivos. Entonces, restringiendo la cadena $\{\mathcal{A}_t\}$ al espacio de estados $\mathcal{C}$, la misma es irreducible, aperiódica y con todos sus estados recurrentes positivos. Luego, por la Definición 2.12, $\{\mathcal{A}_t\}$ es ergódica y, por lo tanto, existe la distribución límite que comprende el único vector de probabilidades estacionarias $\vec{\pi}$. En particular, si $\mathbb{A} \in \mathcal{C}$,

$$\begin{aligned}
\pi_{\mathbb{A}}(\infty) &= \lim_{t\to\infty} P(\mathcal{A}_t = \mathbb{A}) \\
&= \lim_{t\to\infty} P\Big[\bigcap_{f\in\mathcal{F}_\infty} \Big(\{\alpha_t(f,f) = 1\} \cap \bigcap_{\substack{g\prec f \\ g\in\mathcal{F}_\infty}} \{\alpha_t(f,g) = \mathbb{A}_{f,g}\} \Big) \Big].
\end{aligned}$$

Por otra parte, las funciones de amistad entre pares de perfiles f y g de $\mathcal{F}_t$, α_t, fueron definidas como variables dicotómicas que tomaban valores uno o cero de acuerdo a si estos perfiles eran amigos o no en Facebook en el instante t. Entonces, bajo la distribución ergódica, a tiempo infinito, las variables aleatorias $\alpha_\infty(f,g)$, con $f \prec g$, tienen distribución Bernoulli.

En el contexto de TC, cada perfil de $\mathcal{F}_\infty$ entabla amistad con cualquier otro con la misma probabilidad, digamos p, por lo que las variables aleatorias $\alpha_\infty(f,g)$, con $f \prec g$, $f,g \in \mathcal{F}_\infty$ son además idénticamente distribuidas Bernoulli de parámetro p.

Además, estas variables aleatorias guardan relación directa con los "índices de imagen" entre pares de perfiles, Y_t, y hemos visto que estos índices bajo TC quedaban reducidos a una constante más un ruido blanco. Estos ruidos blancos son variables aleatorias independientes para todo $f,g \in \mathcal{F}_\infty$, con $f \prec g$, y por esto los índices de imagen entre pares de perfiles distintos son independientes. En consecuencia, $\alpha_\infty(f,g)$, con $f \prec g$, $f,g \in \mathcal{F}_\infty$ son independientes. $\square$

Por lo expuesto en el teorema anterior podemos concluir que, para $\mathbb{A} \in \mathcal{C}$, la distribución ergódica bajo TC es

$$\pi_{\mathbb{A}}(\infty) \cong \prod_{f \in \mathcal{F}_\infty} \prod_{\substack{g \prec f \\ g \in \mathcal{F}_\infty}} p^{\mathbb{A}_{f,g}}.(1-p)^{1-\mathbb{A}_{f,g}},\ 0 < p < 1.$$

Capítulo 3

Estimación y Test de Hipótesis

En este Capítulo presentamos estimadores que involucran a dos perfiles de Facebook que luego utilizaremos para probar la hipótesis de TC en la red.

Para realizar los test de hipótesis necesitamos conocer la distribución asintótica de dichos estimadores, para lo cual utilizaremos teoremas como los de *Lindeberg* y *Lyapunov* para arreglos triangulares. Además, daremos a conocer una familia de estadísticos llamados *U-Estadísticos*, introducida por Hoeffding en 1948, quien se basó en un trabajo de Halmos realizado en 1946. Veremos que esta familia admite un Teorema Central del Límite canónico y resuelve el problema de encontrar la distribución límite de varios estadísticos no lineales de gran utilidad.

Por último, introducimos el concepto de Transversalidad Segmentada y proponemos un test para probar TC entre los segmentos para luego definir un índice de performance de utilidad para medir la calidad en la segmentación.

3.1. U-Estadísticos

3.1.1. Definición y ejemplos

Supongamos que $h : \mathbb{R}^r \to \mathbb{R}$ es una función a valores reales de r argumentos $x_1, \ldots, x_r$, y que los argumentos pueden ser números reales o vectores. Sean $X_1, \ldots, X_N$, N observaciones independientes idénticamente distribuidas con función de distribución F. Dado $r \geq 1$, queremos estimar o hacer inferencias sobre el parámetro

$$\theta = \theta(F) = E\Big(h(X_1, \ldots, X_r)\Big) = \int \ldots \int h(x_1, \ldots, x_r)\, dF(x_1) \ldots dF(x_r),$$

suponiendo que $N \geq r$.

Trivialmente, un estimador insesgado de θ es $h(X_1, \ldots, X_r)$. Si $N > r$, como $h(X_1, \ldots, X_r)$ no involucra a todas las observaciones de la muestra, este estimador de θ

puede mejorarse utilizando el Teorema de Rao-Blackwell mediante el cálculo de la esperanza condicionada a un estadístico suficiente, obteniéndose así un estimador de menor varianza. Por ejemplo, si las variables aleatorias X_i toman valores reales, el conjunto de estadísticos de orden $X_{(1)}, \ldots, X_{(N)}$ es siempre suficiente y $E\Big(h(X_1, \ldots, X_r)|X_{(1)}, \ldots, X_{(N)}\Big)$ es un mejor estimador insesgado de θ que $h(X_1, \ldots, X_r)$ y, en este caso

$$E\Big(h(X_1, \ldots, X_r)|X_{(1)}, \ldots, X_{(N)}\Big) = \frac{1}{C_N^r} \sum_{1 \leq i_1 < i_2 < \ldots < i_r \leq N} h(X_{i_1}, \ldots, X_{i_r}),$$

con $(i_1, \ldots, i_r)$ una de las $C_N^r = \frac{N!}{r!(N-r)!}$ colecciones de r enteros distintos del conjunto $\{1, \ldots, N\}$.

Luego, tenemos la siguiente definición:

Definición 3.1 *Sean $X_1, \ldots, X_N$, N observaciones independientes idénticamente distribuidas, un U-Estadístico de orden r, $N \geq r$, con núcleo h se define por:*

$$U = U_N = \frac{1}{C_N^r} \sum_{1 \leq i_1 < i_2 < \ldots < i_r \leq N} h(X_{i_1}, \ldots, X_{i_r}).$$

Claramente, U_N es un estimador insesgado de θ.

Observación 3.2 *Suponemos sin pérdida de generalidad que h es simétrica para que U también tenga esa propiedad. Es decir, suponemos que h no cambia su valor cuando permuta sus argumentos.*

Ejemplo 3.3 *Sea $\theta(F) = \mu(F) = \int x \, dF(x)$. Para el núcleo $h(x) = x$, el correspondiente U-Estadístico es:*

$$U(X_1, \ldots, X_N) = \frac{1}{N} \sum_{i=1}^{N} h(X_i) = \frac{\sum_{i=1}^{N} X_i}{N} = \bar{X},$$

es decir, la media muestral.

Ejemplo 3.4 *Sea $\theta(F) = \mu^2(F) = (\int x \, dF(x))^2$. Para el núcleo $h(x_1, x_2) = x_1 x_2$, el correspondiente U-Estadístico es:*

$$U(X_1, \ldots, X_N) = \frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} X_i \, X_j.$$

Ejemplo 3.5 *Sea $\theta(F) = Var(F) = \sigma^2(F) = \int (x - \mu)^2 \, dF(x)$. Para el núcleo $h(x_1, x_2) = \frac{(x_1 - x_2)^2}{2}$, el correspondiente U-Estadístico es:*

$$U(X_1, \ldots, X_N) = \frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} h(X_i, X_j) = \frac{1}{N-1} \Big(\sum_{i=1}^{N} X_i^2 - N\bar{X}^2 \Big) = s^2,$$

es decir, la varianza muestral.

3.1.2. Varianza de un U-Estadístico:

Consideremos un núcleo simétrico $h(x_1, \ldots, x_r)$ que satisface $E\Big(h^2(X_1, \ldots, X_r)\Big) < \infty$. Definimos las funciones asociadas

$$h_c(x_1, \ldots, x_c) = E\Big(h(x_1, \ldots, x_c, X_{c+1}, \ldots, X_r)\Big),$$

para cada $c = 1, \ldots, r-1$ y $h_r \equiv h$.

Como $\int_A h_c(x_1, \ldots, x_c)\ dF(x_1) \ldots dF(x_c) = \int_{A \times \mathbb{R}^{r-c}} h(x_1, \ldots, x_r)\ dF(x_1) \ldots dF(x_r)$, para todo Boreliano $A \in \mathbb{R}^c$, h_c es la esperanza condicional de $h(X_1, \ldots, X_r)$ dado $X_1, \ldots, X_c$, es decir,

$$h_c(x_1, \ldots, x_c) = E\Big(h(X_1, \ldots, X_r) | X_1 = x_1, \ldots, X_c = x_c\Big).$$

Además, notemos que para $1 \leq c \leq r-1$

$$h_c(x_1, \ldots, x_c) = E\Big(h_{c+1}(x_1, \ldots, x_c, X_{c+1})\Big).$$

Sean $\tilde{h} = h - \theta$ y $\tilde{h}_c = h_c - \theta$, donde $\theta = \theta(F) = E\Big(h(X_1, \ldots, X_r)\Big)$, con $1 \leq c \leq r$.

Observemos que $E\Big(\tilde{h}_c(X_1, \ldots, X_c)\Big) = 0$, para $1 \leq c \leq r$.

Definición 3.1.1 *Definimos* $\zeta_0 = 0$ *y, para* $1 \leq c \leq r$,

$$\zeta_c \ = \ Var\Big(h_c(X_1, \ldots, X_c)\Big) = E\Big(\tilde{h_c}^2(X_1, \ldots, X_c)\Big).$$

Puede probarse que $0 = \zeta_0 \leq \zeta_1 \leq \cdots \leq \zeta_r = Var_F(h) < \infty$, [42].

Sean $\{a_1, \ldots, a_r\}$ y $\{b_1, \ldots, b_r\}$ dos subconjuntos de $\{1, \ldots, N\}$ con r enteros distintos y sea c el número de enteros comunes a ambos conjuntos. Por simetría de $\tilde{h}$, ya que h es simétrica, y por independencia de $\{X_1, \ldots, X_N\}$ resulta

$$E\Big(\tilde{h}(X_{a_1}, \ldots, X_{a_r})\ \tilde{h}(X_{b_1}, \ldots, X_{b_r})\Big) = \zeta_c.$$

La cantidad de elecciones posibles de estos subconjuntos con c elementos en común es $C_N^r\ C_r^c\ C_{N-r}^{r-c}$.

Luego, escribiendo $U_N - \theta = \frac{1}{C_N^r} \sum_c \tilde{h}(X_{i_1}, \ldots, X_{i_r})$, tenemos:

$$\begin{aligned} Var(U_N) &= E_F[(U_N - \theta)^2] \\ &= \frac{1}{(C_N^r)^2} \sum_c \sum_c E\Big(\tilde{h}(X_{a_1}, \ldots, X_{a_r})\ \tilde{h}(X_{b_1}, \ldots, X_{b_r})\Big) \\ &= \frac{1}{(C_N^r)^2} \sum_{c=0}^{N} C_N^r\ C_r^c\ C_{N-r}^{r-c}\ \zeta_c. \end{aligned}$$

Este resultado y otras relaciones útiles probadas por Hoeffding en 1948, pueden resumirse en el lema que daremos a continuación.

Lema 3.1.2 *La varianza de U_N está dada por*

$$Var(U_N) = \frac{1}{C_N^r} \sum_{c=1}^{r} C_r^c \, C_{N-r}^{r-c} \, \zeta_c,$$

y satisface que

$$Var(U_N) = \frac{r^2 \, \zeta_1}{N} + O\left(\frac{1}{N^2}\right),$$

cuando $N \to \infty$.

3.1.3. Distribución asintótica de un U-Estadístico

Los sumandos en la definición de un U-Estadístico no son independientes. Por lo tanto, ni la distribución exacta ni la asintótica se deducen directamente. Sin embargo, puede realizarse una proyección del estadístico U sobre la familia de estadísticos lineales de la forma $\frac{1}{N} \sum_{i=1}^{N} h(X_i)$, resultando esta proyección la parte dominante que determina la distribución límite de U. Esta es la llamada "Proyección de Hájek". Los teoremas principales relacionados con estas ideas pueden verse en Serfling R. [42], Lee A. [30] y Lehmann E. [31] y son enunciados a continuación.

La Proyección de Hájek

Definición 3.1.3 *Supongamos que $E\,|h| < \infty$, la proyección de Hájek del U-Estadístico U_N se define como*

$$\hat{U}_N = \sum_{i=1}^{N} E(U_N | X_i) - (N-1)\,\theta. \tag{3.1}$$

Observemos que la proyección de U_N es exactamente una suma de variables aleatorias independientes e idénticamente distribuidas. En términos de la función $\tilde{h}_1 = h_1 - \theta$, tenemos que

$$\tilde{h}_1(x) = h_1(x) - \theta = E\Big(h(X_1, \ldots, X_r) | X_1 = x\Big) - \theta,$$

y por lo tanto, la expresión (3.1) puede escribirse como

$$\hat{U}_N - \theta = \sum_{i=1}^{N} E(U_N - \theta | X_i) = \frac{r}{N} \sum_{i=1}^{N} \tilde{h}_1(x).$$

A continuación veremos la proyección de U_N para el caso general $\zeta_0 = \cdots = \zeta_{c-1} = 0 < \zeta_c$. Suponemos que $E_F(h^2) < \infty$ y como $\zeta_d = 0$, para $d < c$, la fórmula de la varianza para U-Estadísticos del Lema 3.1.2 deriva en:

$$Var(U_N) = \frac{c!(C_r^c)^2\,\zeta_c}{N^c} + O\left(\frac{1}{N^{c+1}}\right),$$

y

$$Var_F\left((\sqrt{N})^c(U_N - \theta)\right) \to c!\,(C_r^c)^2\,\zeta_c,$$

cuando $N \to \infty$. Esto sugiere que la variable aleatoria $(\sqrt{N})^c(U_N - \theta)$ converge en distribución a una variable no degenerada.

Para $c = 1$, Hoeffding demuestra en 1948 que la variable $\sqrt{n}(U_N - \theta)$ converge en distribución a una variable gaussiana. A continuación enunciamos el teorema correspondiente cuya demostración puede verse en Serfling R. [42].

Teorema 3.1.4 *Supongamos que $E(h^2(X_1, \ldots, X_r)) < \infty$. Sean $X_1, X_2, \ldots, X_r, Y_2, \ldots, Y_r$ observaciones independientes idénticamente distribuidas con función de distribución F, y sea*

$$\zeta_1 = Var(h_1(X_1)) = Cov\left(h(X_1, X_2, \ldots, X_r), h(X_1, Y_2, \ldots, Y_r)\right).$$

Supongamos que $0 < \zeta_1 < \infty$. Entonces

$$\sqrt{N}(U_N - \theta - \hat{U}_N) \xrightarrow{\mathcal{P}} 0, \quad \textit{cuando } N \to \infty, \tag{3.2}$$

y

$$\sqrt{N}(U_N - \theta) \xRightarrow{w} \mathcal{N}(0, r^2\,\zeta_1), \quad \textit{cuando } N \to \infty. \tag{3.3}$$

Observación 3.6 *Existen expresiones exactas para la $Var(U_N)$ aunque alcanza con una aproximación y ya hemos visto que si $\zeta_1 > 0$, entonces $Var(U_N) = \frac{r^2\,\zeta_1}{N} + O(\frac{1}{N^2})$.*

3.2. Estimación y test

En esta Sección nuestro objetivo es estudiar la comunicación entre dos perfiles de Facebook. Para ello nos valdremos de algunos estadísticos que resultarán de utilidad para probar la hipótesis de TC. Intentaremos hallar la distribución asintótica de dichos estadísticos bajo el supuesto de que se cumple la hipótesis mencionada.

Recordemos que nuestro interés se centra en la dinámica de la red a largo plazo cuando está en estado de madurez y que en estas condiciones hemos probado que bajo TC y bajo la distribución ergódica $\vec{\pi}(\infty)$, las indicatrices de si existe amistad entre los perfiles

$f, g \in \mathcal{F}_\infty$, con $g \prec f$ y denotadas por $\alpha_\infty(f,g)$, son variables aleatorias independientes con distribución Bernoulli de parámetro p, $0 < p < 1$.

Los estadísticos que enunciaremos a continuación se basarán en muestras tomadas al azar de N perfiles de $\mathcal{F}_\infty$, $f_1, \ldots, f_N$.

3.2.1. Promedio de la comunicación en la muestra

Sea

$$E_N = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{j=1}^{N}\alpha_\infty(f_i,f_j)\right), \tag{3.4}$$

este estadístico promedia la proporción de amigos o comunicaciones que tienen los perfiles en la muestra y por consiguiente mide la "comunicación" de la muestra en promedio.

Si suponemos que en la red social Facebook se verifica la hipótesis de TC, como las variables aleatorias $(\alpha_\infty(f_i,f_j))_{i<j}$ son Bernoulli independientes de parámetro p, $E(\alpha_\infty(f_i,f_j)) = p$ y $Var(\alpha_\infty(f_i,f_j)) = p(1-p)$. En este contexto, es sencillo calcular $E(E_N)$ y $Var(E_N)$. En efecto, teniendo en cuenta la simetría de las variables $\alpha_\infty(f_i,f_j)$ y que en el largo plazo suponemos que todos los perfiles fueron creados y por lo tanto $\alpha_\infty(f_i,f_i) = 1$, escribimos (3.4) de la siguiente manera:

$$E_N = \frac{1}{N^2}\left(N + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\alpha_\infty(f_i,f_j)\right)$$

entonces

$$\begin{aligned} E(E_N) &= \frac{1}{N} + \frac{2}{N^2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}E(\alpha_\infty(f_i,f_j)) \\ &= \frac{1}{N} + \frac{2}{N^2}\frac{N(N-1)}{2}\,p \\ &= \frac{1}{N} + \frac{N-1}{N}\,p \to p, \quad \text{cuando } N \to \infty, \end{aligned}$$

y

$$\begin{aligned} Var(E_N) &= \frac{4}{N^4}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}Var(\alpha_\infty(f_i,f_j)) \\ &= \frac{4}{N^4}\frac{N(N-1)}{2}p(1-p) \\ &= \frac{2(N-1)}{N^3}\,p(1-p). \end{aligned} \tag{3.5}$$

Distribución asintótica de $\mathbf{E_N}$

Para hallar la distribución asintótica de E_N estudiaremos la distribución asintótica del estadístico centrado

$$\begin{aligned} E_N - p &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{j=1}^{N}(\alpha_\infty(f_i,f_j)-p)\right) \\ &= \frac{1}{N^2}\left[N(1-p)+2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(\alpha_\infty(f_i,f_j)-p)\right] \\ &= \frac{(1-p)}{N}+\frac{2}{N^2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(\alpha_\infty(f_i,f_j)-p). \end{aligned} \tag{3.6}$$

Notando $Y_N^i = \frac{1}{N^2}\sum\limits_{i=i+1}^{N}(\alpha_\infty(f_i,f_j)-p)$, la expresión (3.6) resulta

$$E_N - p = 2\sum_{i=1}^{N-1} Y_N^i + \frac{1-p}{N}, \tag{3.7}$$

donde $\{Y_N^i\}_{1\le i\le N-1}$ forma un arreglo triangular.

Teorema 3.2.1 *Si el arreglo triangular $\{Y_N^i : i = 1,2,\ldots,N-1\}$ verifica las siguientes condiciones*

i) Para cada $N \in \mathbb{N}$, $\{Y_N^i : i = 1,2,\ldots,N-1\}$ son independientes;

ii) $E(Y_N^i) \to 0$, cuando $N \to \infty$;

iii) $s_N^2 = \sum\limits_{i=1}^{N-1} Var(Y_N^i) < \infty$;

iv) Existe $\delta > 0$ tal que

$$E[(Y_N^i)^{2+\delta}] < \infty, \text{ para todo } N \text{ y para todo } i,$$

y se verifica la condición de Lyapunov, es decir,

$$L(N,\delta) = \frac{1}{s_N^{2+\delta}}\sum_{i=1}^{N-1} E[(Y_N^i)^{2+\delta}] \to 0, \text{ cuando } N \to \infty, \tag{3.8}$$

entonces

$$\frac{1}{s_N}\sum_{i=1}^{N-1} Y_N^i \overset{w}{\Longrightarrow} \mathcal{N}(0,1), \text{ cuando } N \to \infty. \tag{3.9}$$

Demostración

La hipótesis *i)* se verifica trivialmente, ya que para $i \neq i'$ y N fijo, $\sum_{j=i+1}^{N} (\alpha_\infty(f_i, f_j) - p)$ es independiente de $\sum_{j=i'+1}^{N} (\alpha_\infty(f_{i'}, f_j) - p)$.

Además,

$$\begin{aligned} E(Y_N^i) &= \frac{1}{N^2} E\Big(\sum_{j=i+1}^{N} (\alpha_\infty(f_i, f_j) - p) \Big) \\ &= \frac{1}{N^2} \sum_{j=i+1}^{N} E(\alpha_\infty(f_i, f_j) - p) = 0 \end{aligned}$$

ya que $E(\alpha_\infty(f_i, f_j) - p) = 0$. Por otro lado,

$$\begin{aligned} s_N^2 &= \sum_{i=1}^{N-1} Var(Y_N^i) = \sum_{i=1}^{N-1} Var\Big(\frac{1}{N^2} \sum_{j=i+1}^{N} (\alpha_\infty(f_i, f_j) - p) \Big) \\ &= \frac{1}{N^4} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} Var(\alpha_\infty(f_i, f_j) - p) \\ &= \frac{1}{N^4} \frac{N(N-1)}{2} p(1-p) < \infty. \end{aligned} \tag{3.10}$$

Entonces se cumplen las hipótesis *ii)* y *iii)*.

Sea $\delta = 2$. Veamos que se verifica la hipótesis *iv)*. Para esto, calcularemos

$$\begin{aligned} E[(Y_N^i)^4] &= \frac{1}{N^8} E\left[\Big(\sum_{j=i+1}^{N} (\alpha_\infty(f_i, f_j) - p) \Big)^4 \right] \qquad (3.11) \\ &= \frac{1}{N^8} \left[\sum_{j=i+1}^{N} E[(\alpha_\infty(f_i, f_j) - p)^4] + C_4^2 \sum_{i<j<k\leq N} E[(\alpha_\infty(f_i, f_j) - p)^2]\, E[(\alpha_\infty(f_i, f_k) - p)^2] \right], \end{aligned}$$

pues $E(\alpha_\infty(f_i, f_j) - p) = 0$ y por esto, los términos $(\alpha_\infty(f_i, f_j) - p)^3\, (\alpha_\infty(f_i, f_k) - p)$, los términos $(\alpha_\infty(f_i, f_j) - p)^2\, (\alpha_\infty(f_i, f_k) - p)\, (\alpha_\infty(f_i, f_l) - p)$ y los términos $(\alpha_\infty(f_i, f_j) - p)\, (\alpha_\infty(f_i, f_k) - p)\, (\alpha_\infty(f_i, f_l) - p)\, (\alpha_\infty(f_i, f_m) - p)$ en la expresión (3.11) tienen esperanza cero, si $j \neq k \neq l \neq m$. Por lo tanto,

$$\begin{aligned} E[(Y_N^i)^4] &= \frac{1}{N^8} \Big[(N-i)(-3p^4 + 6p^3 - 4p^2 + p) + 3(N-i)[(N-i)-1]p^2(1-p)^2 \Big] \\ &= \frac{1}{N^8} \Big[(N-i)\mathbb{C}_p + (N-i)[(N-i)-1)\tilde{\mathbb{C}}_p] \\ &= \frac{1}{N^8} \Big[N^2\tilde{\mathbb{C}}_p - N[(2i+1)\tilde{\mathbb{C}}_p - \mathbb{C}_p] + [(i^2+i)\tilde{\mathbb{C}}_p - i\mathbb{C}_p] \\ &\leq \frac{N^2}{N^8} \tilde{\mathbb{C}}_p < \infty, \text{ para todo } N \text{ y para todo } i, \end{aligned} \tag{3.12}$$

con $\mathbb{C}_p = -3p^4 + 6p^3 - 4p^2 + p$ y $\tilde{\mathbb{C}}_p = 3p^2(1-p)^2$.

Entonces vemos que se verifica la condición de Lyapunov dada por (3.8) para $\delta = 2$:

$$\begin{aligned} L(N,2) &= \frac{1}{s_N^4} \sum_{i=1}^{N-1} E[(Y_N^i)^4] \\ &= \sum_{i=1}^{N-1} \left(\frac{2\ N^3}{(N-1)\ p(1-p)} \right)^2 E[(Y_N^i)^4] \\ &\leq \sum_{i=1}^{N-1} \frac{4\ N^4}{p^2(1-p)^2} \frac{N^2}{N^8} \tilde{\mathbb{C}}_p \\ &< \frac{4}{p^2(1-p)^2} \frac{\tilde{\mathbb{C}}_p}{N} \downarrow 0^+, \text{ cuando } N \to \infty. \end{aligned}$$

En consecuencia, como se verifican las hipótesis *i)* a *iv)* concluimos que

$$\frac{1}{s_N} \sum_{i=1}^{N-1} Y_N^i \overset{w}{\Longrightarrow} \mathcal{N}(0,1), \text{ cuando } N \to \infty.$$

□

Volviendo a la expresión del estadístico centrado $E_N - p$ en función de $\sum_{i=1}^{N-1} Y_N^i$ dada por (3.7) y a la expresión para s_N dada por (3.10), tenemos que para N suficientemente grande $E_N - p \cong 2 \sum_{i=1}^{N-1} Y_N^i$ y que $\frac{1}{s_N} = \frac{\sqrt{2}N}{\sqrt{\frac{N-1}{N} p(1-p)}}$. Luego,

$$\frac{1}{s_N} \sum_{i=1}^{N-1} Y_N^i \cong \frac{N(E_N - p)}{\sqrt{2\frac{N-1}{N} p(1-p)}} \overset{w}{\Longrightarrow} \mathcal{N}(0,1), \text{ cuando } N \to \infty,$$

es decir,

$$N(E_N - p) \overset{w}{\Longrightarrow} \mathcal{N}(0, 2p(1-p)), \text{ cuando } N \to \infty. \tag{3.13}$$

Prueba de hipótesis para la *TC*

E_N es un estimador de la probabilidad de comunicación de los perfiles y, en el caso de TC, hemos supuesto que esta probabilidad es la misma para todos los perfiles. Luego, si tomamos distintas muestras de perfiles, bajo el supuesto de TC, no deberíamos detectar diferencias entre las estimaciones de p basadas en muestras distintas.

Para estudiar este fenómeno proponemos una prueba de hipótesis para comparar la comunicación promedio que existe dentro de poblaciones independientes, suponiendo TC.

Considerando dos muestras independientes de N perfiles de $\mathcal{F}_\infty$, $f_1, \ldots, f_N$ y $g_1, \ldots, g_N$, con $\{f_1, \ldots, f_N\} \cap \{g_1, \ldots, g_N\} = \emptyset$, construimos los estadísticos

$$E_N = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{N} \sum_{j=1}^{N} \alpha_\infty(f_i, f_j) \right), \tag{3.14}$$

donde $N(E_N - p) \overset{w}{\Longrightarrow} \mathcal{N}(0, 2p(1-p)),$ cuando $N \to \infty$, y

$$E_N^* = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{N} \sum_{j=1}^{N} \alpha_\infty(g_i, g_j) \right), \tag{3.15}$$

donde $N(E_N^* - p) \overset{w}{\Longrightarrow} \mathcal{N}(0, 2p(1-p)),$ cuando $N \to \infty$.

En general cuando la media μ y la varianza σ^2 de una distribución están ligadas, es decir, $\sigma = \sigma(\mu)$ y además $\sigma(\mu)$ es derivable, existe una generalización del Teorema Central del Límite canónico que consiste en tomar una función $g(\mu)$ tal que $g'(\mu) = \frac{1}{\sigma(\mu)}$, es decir, una primitiva de $\frac{1}{\sigma(\mu)}$. Luego, el Teorema Central del Límite canónico

$$\sqrt{n}(\bar{X}_n - \mu) \overset{w}{\Longrightarrow} \mathcal{N}(0, \sigma^2), \text{ cuando} N \to \infty,$$

implica que

$$\sqrt{n}\left(g(\bar{X}_n) - g(\mu)\right) \overset{w}{\Longrightarrow} \mathcal{N}\left(0, \sigma^2(\mu) g'(\mu)^2\right) = \mathcal{N}(0,1), \text{ cuando } N \to \infty.$$

Si queremos entonces, realizar un test aproximado para comparación de medias de dos poblaciones X e Y independientes entre sí con el mismo tipo de distribución y de modo que al tomar muestras de tamaño m y n de las poblaciones, con m y n suficientemente grandes, la función $\sigma = \sigma(\mu)$ es la misma para ambas, la región crítica a nivel α resulta

$$R_\alpha = \{|E| \geq z_{\alpha/2}\},$$

con $E = \sqrt{\frac{m\,n}{m+n}}\left(g(\bar{X}_m) - g(\bar{Y}_n)\right)$, siendo $g(x) = \int_{\mu_0}^{x} \frac{1}{\sigma(t)} dt$.

Volviendo a nuestro caso, como

$$\mu = E(E_N) = E(E_N^*) = p \in (0,1),$$

y

$$\sigma^2 = Var(E_N) = Var(E_N^*) = 2p(1-p),$$

vemos que $\sigma = \sigma(\mu)$ y $\sigma(\mu)$ es derivable. Entonces, tomando

$$g(\mu) = \sqrt{2}\ \operatorname{arc\,sen}(\sqrt{\mu}),$$

$g(\mu)$ verifica que $g'(\mu) = \frac{1}{\sigma(\mu)}$ y por lo tanto

$$N(g(E_N) - g(p)) \overset{w}{\Longrightarrow} \mathcal{N}(0, \sigma^2(p)\ g'(p)^2) = \mathcal{N}(0,1), \text{ cuando } N \to \infty, \tag{3.16}$$

y

$$N(g(E_N^*) - g(p)) \overset{w}{\Longrightarrow} \mathcal{N}(0, \sigma^2(p)\ g'(p)^2) = \mathcal{N}(0,1), \text{ cuando } N \to \infty. \tag{3.17}$$

En consecuencia, para realizar el test aproximado de comparación de las comunicaciones promedio, el estadístico de la prueba suponiendo que se cumple la hipótesis de TC resulta $N(g(E_N) - g(p)) - N(g(E_N^*) - g(p)) = N(g(E_N) - g(E_N^*))$ y, según (3.16) y (3.17),

$$N(g(E_N) - g(E_N^*)) \overset{w}{\Longrightarrow} \mathcal{N}(0,2), \text{ cuando } N \to \infty. \tag{3.18}$$

Luego, para un nivel de significación α, contamos con la siguiente región crítica

$$R_\alpha = \left\{ \frac{N}{\sqrt{2}} \left| g(E_N) - g(E_N^*) \right| \geq z_{\alpha/2} \right\}.$$

Se realizó un experimento sobre una red de amigos de un perfil real particular, tomando dos muestras independientes y disjuntas de tamaño $N = 75$ y se obtuvo que $E_N = 0{,}69$ y $E_N^* = 0{,}57$ y el valor del estadístico fue $\frac{N}{\sqrt{2}} \left| g(E_N) - g(E_N^*) \right| = 2{,}1$. Por consiguiente a un nivel de significación $\alpha = 0{,}05$ el test rechaza la hipótesis de TC.

Este resultado estaría indicando que la red social Facebook es una plataforma en la que la comunicación entre las personas o grupos de personas no es transversal.

3.2.2. Desviación media cuadrática de la comunicación entre perfiles

Presentamos ahora un estadístico que mide la desviación promedio cuadrática de la comunicación entre perfiles con respecto a su media.

$$\begin{aligned} T_N &= \frac{1}{C_N^2} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left[\alpha_\infty(f_i, f_j) - \frac{1}{C_N^2} \sum_{k=1}^{N-1} \sum_{l=k+1}^{N} \alpha_\infty(f_k, f_l) \right]^2 \\ &= \frac{1}{C_N^2} \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \alpha_\infty(f_i, f_j)^2 - \frac{1}{C_N^2} \left(\sum_{k=1}^{N-1} \sum_{l=k+1}^{N} \alpha_\infty(f_k, f_l) \right)^2 \right] \\ &= \frac{1}{(C_N^2)^2} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \sum_{k=1}^{N-1} \sum_{l=k+1}^{N} \frac{(\alpha_\infty(f_i, f_j) - \alpha_\infty(f_k, f_l))^2}{2}. \end{aligned}$$

Distribución asintótica de $\mathbf{T_N}$

Podemos ver que T_N es un U-estadístico de orden 2 y núcleo

$$h(v_\infty(f_i), v_\infty(f_k)) = \frac{(\alpha_\infty(f_i, f_j) - \alpha_\infty(f_k, f_l))^2}{2},$$

con $v_\infty(f_i) = \Big(\alpha_\infty(f_i, f_h)\Big)_{h\in\mathcal{F}_\infty, i<h}$ y $v_\infty(f_k) = \Big(\alpha_\infty(f_k, f_h)\Big)_{h\in\mathcal{F}_\infty, k<h}$.

Además, bajo el supuesto de TC,

$$\begin{aligned}
\theta &= E[h(v_\infty(f_i), v_\infty(f_k))] = \frac{1}{2}\, E\Big[(\alpha_\infty(f_i, f_j) - \alpha_\infty(f_k, f_l))^2\Big] \\
&= \frac{1}{2}\, E[\alpha_\infty(f_i, f_j)^2 - 2\alpha_\infty(f_i, f_j)\alpha_\infty(f_k, f_l) + \alpha_\infty(f_k, f_l)^2] \\
&= \frac{1}{2}\,(p - 2p^2 + p) = p(1-p),
\end{aligned}$$

y

$$\begin{aligned}
\zeta_1 &= Cov\Big(h(v_\infty(f_1), v_\infty(f_3)); h(v_\infty(f_1), v_\infty(f_5))\Big) \\
&= Cov\left(\frac{(\alpha_\infty(f_1, f_2) - \alpha_\infty(f_3, f_4))^2}{2}; \frac{(\alpha_\infty(f_1, f_2) - \alpha_\infty(f_5, f_4))^2}{2}\right) \\
&= E\left(\frac{(\alpha_\infty(f_1, f_2) - \alpha_\infty(f_3, f_4))^2}{2} \times \frac{(\alpha_\infty(f_1, f_2) - \alpha_\infty(f_5, f_4))^2}{2}\right) - \theta^2 \\
&= \frac{1}{4}p(1-p) - (p(1-p))^2.
\end{aligned}$$

Observemos que $\zeta_1 = 0$ si y sólo si $p = 0$, $p = 1$ o $p = 0{,}5$. Sin embargo, a los efectos del modelo de Facebook con el que estamos tratando $p = P(\alpha_\infty(f_i, f_j) = 1)$ es la probabilidad de que dos perfiles cualesquiera sean amigos en el largo plazo, y este es un número entre 0 y 1, probablemente menor que 0,5 debido al gran tamaño de la red completa.

Luego, como $0 < \zeta_1 < \infty$, por el Teorema 3.1.4,

$$\sqrt{N}(T_N - p(1-p)) \overset{w}{\Longrightarrow} \mathcal{N}\Big(0, p(1-p) - 4\,(p(1-p))^2\Big), \text{ cuando } N \to \infty. \quad (3.19)$$

Si notamos

$\mu = p(1-p)$ y $\sigma = \sqrt{p(1-p) - 4(p(1-p))^2}$, tenemos que $\sigma = \sigma(\mu)$ y $\sigma(\mu)$ es derivable. Luego, tomando

$$g(\mu) = \operatorname{arc\,sen}(2\sqrt{\mu}),$$

resulta que $g(\mu)$ verifica que $g'(\mu) = \frac{1}{\sigma(\mu)}$ y en consecuencia la expresión del límite en (3.19) equivale a

$$\sqrt{N}\Big(g(T_N) - g(p(1-p))\Big) \overset{w}{\Longrightarrow} \mathcal{N}(0, \sigma^2(\mu)\ g'(\mu)^2) = \mathcal{N}(0,1), \text{ cuando } N \to \infty.$$

Prueba de hipótesis para la *TC*

Podemos también, realizar el test de TC comparando las desviaciones medias cuadráticas para dos poblaciones independientes de perfiles, de manera similar a la comparación de las comunicaciones promedio realizada en la subsección anterior. Es decir, podemos tomar una muestra de N perfiles de $\mathcal{F}_\infty$, $g_1, \ldots, g_N$, independiente de la anterior $f_1, \ldots, f_N$, tal que $\{f_1, ..., f_N\} \cap \{g_1, ..., g_N\} = \emptyset$, construir el estadístico

$$T_N^* = \frac{1}{C_N^2} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left[\alpha_\infty(g_i, g_j) - \frac{1}{C_N^2} \sum_{k=1}^{N-1} \sum_{l=k+1}^{N} \alpha_\infty(g_k, g_l)\right]^2$$

y concluir que

$$\sqrt{N}\Big(g(T_N^*) - g(p(1-p))\Big) \overset{w}{\Longrightarrow} \mathcal{N}(0, \sigma(\mu)^2\ g'(\mu)^2) = \mathcal{N}(0,1), \text{ cuando } N \to \infty.$$

Luego, el estadístico para la comparación de las desviaciones medias cuadráticas

$$\sqrt{N}(g(T_N) - g(T_N^*)) \overset{w}{\Longrightarrow} \mathcal{N}(0,2), \text{ cuando } N \to \infty.$$

y la región crítica a nivel de significación α para el test de TC utilizando este estadístico es

$$R_\alpha = \left\{ \sqrt{\frac{N}{2}}\, |g(T_N) - g(T_N^*)| \geq z_{\alpha/2} \right\}.$$

Del mismo modo que expusimos para el test de comparación de la proporción media de la comunicación entre perfiles, hemos tomando el mismo perfil particular real y calculado el estadístico y la región crítica llegando a la conclusión de que se rechaza la hipótesis de TC.

3.3. Transversalidad Segmentada

Hemos utilizado el supuesto de TC en Facebook para el modelado de la red social y la obtención de la distribución asintótica de los estadísticos que presentamos en la sección anterior propuestos para testear dicha hipótesis. Sin embargo, un contexto comunicacional como el de TC en una red social con más de mil millones de usuarios está muy alejado de la realidad, como lo demuestran las conclusiones de los dos test realizados. Es razonable pensar que los perfiles tienden a agruparse en segmentos de acuerdo a distintos

criterios sociales como ideologías políticas, intereses económicos, gustos musicales, edades, deportes, etc, y que estos segmentos también se relacionen entre sí.

En esta Sección, introducimos el concepto de Transversalidad Segmentada, (TS), es decir, TC entre segmentos. En un contexto de estas características, lo que dejará de cumplirse es la homogeneidad del parámetro p de las variables aleatorias Bernoulli $(\alpha_\infty(f,g))_{g\prec f}$ que representan la comunicación entre perfiles pero se mantiene la independencia entre dichas variables.

Podemos realizar una segmentación "a priori" en la red considerando $S_1,\ldots,S_k$, k segmentos disjuntos formando una partición de $\mathcal{F}_\infty$ y, reordenando los elementos de $\mathcal{F}_\infty$, reescribir a la matriz de amistad $\mathcal{A}_t$ como una matriz de bloques, donde el primer bloque corresponda a los perfiles que integran el segmento S_1, el segundo bloque a los que pertenezcan a S_2, y así sucesivamente. Luego de la segmentación, podemos realizar un test de TC dentro de cada segmento como los que hemos visto en la sección anterior y también podemos construir un estadístico que represente la comunicación entre pares de segmentos con el fin de probar la hipótesis de TC entre los perfiles de dichos segmentos. Además, podemos estar interesados en medir la calidad en la segmentación mediante índices de performance adecuados. A continuación detallamos estas ideas.

3.3.1. TC entre segmentos

Sea $S_1,\ldots,S_k$, una partición en segmentos de $\mathcal{F}_\infty$ y sea $a_i=\frac{card\{S_i\}}{card\{\mathcal{F}_\infty\}}$, $i=1,2,\ldots,k$. Es claro que cada $a_i>0$, $\sum_{i=1}^{k} a_i=1$.

Realizamos un muestreo estratificado al azar por segmentos, es decir, los perfiles $f_1,\ldots,f_{[a_1N]}$ son elegidos al azar dentro del segmento S_1, los perfiles $f_{[a_1N]+1},\ldots f_{[(a_1+a_2)N]}$ son elegidos al azar dentro del segmento S_2, y así sucesivamente hasta que los perfiles $f_{\left[\left(\sum_{i=1}^{k-1} a_i\right)N\right]+1},\ldots,f_{\left(\sum_{i=1}^{k} a_i\right)N}$ son elegidos al azar dentro de S_k, donde $[x]$ es la parte entera de x, es decir $[x]=\text{máx}\{k\in\mathbb{N}: k\leq x\}$, para $x>0$.

Sean los conjuntos:

$I_1=\{1,\ldots,[a_1N]\}$, con $card\{I_1\}=[a_1N]$,

$I_2=\{[a_1N]+1,\ldots,[(a_1+a_2)N]\}$, con $card\{I_2\}=[(a_1+a_2)N]-[a_1N]$,

$$\vdots$$

$$I_k=\left\{\left[\left(\sum_{i=1}^{k-1} a_i\right)N\right]+1,\ldots,\left(\sum_{i=1}^{k} a_i\right)N\right\},\ \text{con } card\{I_k\}=\left(\sum_{i=1}^{k} a_i\right)N-\left[\left(\sum_{i=1}^{k-1} a_i\right)N\right].$$

Dados S_r y S_t, con $r,t\in\{1,\ldots,k\}$, dos segmentos de la partición en k segmentos de $\mathcal{F}_\infty$, para $f_i\in S_r$ y $f_j\in S_t$, con $i\in I_r$, $j\in I_t$. Notaremos con q_{ij} a la probabilidad de que los perfiles f_i y f_j sean amigos, es decir, $P(\alpha_\infty(f_i,f_j)=1)=q_{ij}$.

Observemos que las funciones aleatorias de amistad $\alpha_\infty(f_i,f_j)$ siguen siendo variables

Bernoulli aunque el parámetro de su distribución ahora depende de la intensidad con que se relacione el par de perfiles considerado. Sin embargo, puede ocurrir que dado un par de segmentos distintos haya homogeneidad en cuanto a la comunicación entre los perfiles de un segmento y el otro. En este sentido, formulamos la siguiente definición.

Definición 3.7 *Dados dos segmentos S_r y S_t, decimos que existe TC entre ellos si dados $f_i \in S_r$ y $f_j \in S_t$, con $i \in I_r$ y $j \in I_t$, $\alpha_\infty(f_i, f_j)$ tiene distribución Bernoulli de parámetro q_{rt}, para todo $f_i \in S_r$ y para todo $f_j \in S_t$.*

Es decir, TC entre segmentos indica un comportamiento comunicacional homogéneo distintivo entre los mismos.

A continuación presentamos un estadístico que promedia la proporción de amigos del segmento S_t que tienen los perfiles del segmento S_r:

$$E(r,t) = \frac{1}{card\{I_r\}} \sum_{i \in I_r} \frac{1}{card\{I_t\}} \sum_{j \in I_t} \alpha_\infty(f_i, f_j), \tag{3.20}$$

luego, si suponemos un contexto de TC entre S_r y S_t,

$$E(E(r,t)) = q_{rt},$$

y

$$Var(E(r,t)) = \frac{q_{rt}(1-q_{rt})}{card\{I_r\}\, card\{I_t\}},$$

y el estadístico (3.20) centrado queda

$$E(r,t) - q_{rt} = \sum_{i \in I_r} \frac{1}{card\{I_r\}} \sum_{j \in I_t} \frac{\alpha_\infty(f_i, f_j) - q_{rt}}{card\{I_t\}}. \tag{3.21}$$

Por lo tanto, notando

$$Y^i_{card\{I_r\}} = \frac{1}{card\{I_r\}} \sum_{j \in I_t} \frac{\alpha_\infty(f_i, f_j) - q_{rt}}{card\{I_t\}}$$

resulta que $Y^1_{card\{I_r\}}, \ldots, Y^{card\{I_r\}}_{card\{I_r\}}$ es un arreglo triangular que verifica las hipótesis del Teorema 3.2.1. En consecuencia,

$$\frac{1}{s_{card\{I_r\}}} \sum_{i \in I_r} Y^i_{card\{I_r\}} \overset{w}{\Longrightarrow} \mathcal{N}(0,1), \text{ cuando } N \to \infty,$$

con $s^2_{card\{I_r\}} = \sum_{i \in I_r} Var(Y^i_{card\{I_r\}}) = \frac{q_{rt}(1-q_{rt})}{card\{I_r\}\, card\{I_t\}}$. Es decir,

$$\sqrt{card\{I_r\}\, card\{I_t\}}\, (E(r,t) - q_{rt}) \overset{w}{\Longrightarrow} \mathcal{N}(0, q_{rt}(1-q_{rt})), \text{ cuando } N \to \infty. \tag{3.22}$$

Luego, podemos realizar una prueba de hipótesis para probar si existe TC entre un par de segmentos de una partición de $\mathcal{F}_\infty$. El procedimiento para obtener la región crítica es análogo al que utilizamos en la Sección anterior, esto es, se realiza otro muestreo estratificado al azar por segmentos, independiente del anterior en el cual los perfiles $g_1, \ldots, g_{[a_1 N]}$ son elegidos al azar dentro del segmento S_1 con $\{g_1, \ldots, g_{[a_1 N]}\} \cap \{f_1, \ldots, f_{[a_1 N]}\} = \emptyset$, los perfiles $g_{[a_1 N]+1}, \ldots g_{[(a_1+a_2)N]}$ son elegidos al azar dentro S_2 con $\{f_{[a_1 N]+1}, \ldots f_{[(a_1+a_2)N]}\} \cap \{g_{[a_1 N]+1}, \ldots g_{[(a_1+a_2)N]}\} = \emptyset$, y así sucesivamente hasta que los perfiles $g_{\left[\left(\sum_{i=1}^{k-1} a_i\right)N\right]+1}, \ldots, g_{\left(\sum_{i=1}^{k} a_i\right)N}$ son elegidos al azar dentro de S_k con $\left\{f_{\left[\left(\sum_{i=1}^{k-1} a_i\right)N\right]+1}, \ldots, f_{\left(\sum_{i=1}^{k} a_i\right)N}\right\} \cap \left\{g_{\left[\left(\sum_{i=1}^{k-1} a_i\right)N\right]+1}, \ldots, g_{\left(\sum_{i=1}^{k} a_i\right)N}\right\} = \emptyset$.

Luego, para el par de segmentos S_r y S_t, con $r, t \in \{1, \ldots, k\}$, el estadístico

$$E^*(r,t) = \frac{1}{card\{I_r\}} \sum_{i \in I_r} \frac{1}{card\{I_t\}} \sum_{j \in I_t} \alpha_\infty(g_i, g_j), \tag{3.23}$$

bajo TC entre S_r y S_t, tiene $E(E^*(r,t)) = q_{rt}$ y $Var(E^*(r,t)) = \frac{q_{rt}(1-q_{rt})}{card\{I_r\}\ card\{I_t\}}$ y al igual que para el estadístico centrado de la expresión (3.21) obtenemos la distribución asintótica de $E^*(r,t) - q_{rt}$, cuando $N \to \infty$:

$$\sqrt{card\{I_r\}\ card\{I_t\}}\ (E^*(r,t) - q_{rt}) \overset{w}{\Longrightarrow} \mathcal{N}(0, q_{rt}(1-q_{rt})), \text{ cuando } N \to \infty. \tag{3.24}$$

Como los estadísticos $E(r,t)$ y $E^*(r,t)$ verifican que la varianza σ es función de la esperanza μ, es decir, $\sigma(\mu) = \mu(1-\mu)$, con $\mu = q_{rt}$, tomando $g(\mu) = 2\ \text{arc sen}(\sqrt{\mu})$, se cumple que $g'(\mu) = \frac{1}{\sigma(\mu)}$ y, procediendo en forma análoga que en la Sección anterior, podemos concluir que

$$\sqrt{card\{I_r\}\ card\{I_t\}}\ (g(E(r,t) - g(E^*(r,t)))) \overset{w}{\Longrightarrow} \mathcal{N}(0,2), \text{ cuando } N \to \infty,$$

y la región crítica para probar TC a nivel de significación α entre los perfiles involucrados en ambos segmentos es

$$R_\alpha = \left\{ \sqrt{\frac{card\{I_r\}\ card\{I_t\}}{2}}\ |g(E(r,t) - g(E^*(r,t)))| \geq z_{\alpha/2} \right\}.$$

Luego, si el test nos conduce al rechazo de la hipótesis de TC entre los segmentos S_r y S_t, podemos concluir con una probabilidad de error α que dichos segmentos no tienen un comportamiento comunicacional homogéneo distintivo.

3.3.2. Calidad en la segmentación

Contando con el test de TC entre segmentos que acabamos de exponer, si hemos dividido a la red en k segmentos disjuntos, podemos tomar todos los posibles pares de

segmentos de esos k segmentos, C_k^2, y realizar una cantidad C_k^2 de test, uno para cada par y probar si los segmentos que componen el par tienen un comportamiento comunicacional homogéneo distintivo o no. Estos C_k^2 test pueden representarse en una matriz binaria simétrica de orden k, $\mathcal{S}$, en la que cada elemento $\mathcal{S}_{ij}$ es cero si los segmentos S_i y S_j no tuvieron homogeneidad en cuanto a la comunicación, esto es, si el test correspondiente rechaza la hipótesis de TC entre S_i y S_j y, en caso contrario, indicamos con un uno al elemento $\mathcal{S}_{ij}$.

Luego, si notamos al cardinal del conjunto de "unos" en la subdiagonal de $\mathcal{S}$ como

$$g(d) = card\{1's \text{ en } SD(\mathcal{S})\},$$

podemos definir el siguiente índice de performance de utilidad para medir la Calidad en la Segmentación:

$$C_p = \frac{g(d)}{C_k^2}\, 100\,\%. \tag{3.25}$$

Repitiendo m veces este procedimiento, es decir, realizando m muestreos estratificados al azar por segmento en los k segmentos, con m un número suficientemente grande, podemos calcular cada vez el índice de performance definido en (3.25), C_p^i, con $i = 1, \ldots, m$, y visualizar el histograma que representa la distribución de la Calidad en la Segmentación. Si para la mayoría de las repeticiones esta medida resultó, por ejemplo, mayor que la media de las observaciones, continuamos segmentando según el criterio que se venía utilizando, caso contrario es conveniente modificar el criterio de segmentación.

Capítulo 4

Conectividad y Teoría de Mundo Pequeño

En este Capítulo tratamos el "fenómeno de mundo pequeño" y su viabilidad en los distintos contextos comunicacionales que se pueden presentar en la red social Facebook, es decir, en los contextos de TC y distintos tipos de segmentación. En la primera Sección, proporcionamos una síntesis de los estudios y experimentos más relevantes que se realizaron para abordar el problema. Con el objetivo de analizar el fenómeno mencionado, en la segunda Sección introducimos un estadístico de utilidad en la medición de la conectividad entre los usuarios y estudiamos su distribución asintótica en el contexto particular de TC, cuando el tamaño de la muestra es suficientemente grande. Otra forma de evaluar la conectividad consiste en el cálculo de las distancias entre pares de perfiles que componen la red y para ello, en la tercera Sección damos una definición formal del grado medio de separación entre dos perfiles y, según el contexto, podremos calcularlo y realizar el análisis. La última Sección, se refiere a conclusiones y planteos futuros sobre el tema.

4.1. Redes mundo pequeño

4.1.1. El fenómeno de mundo pequeño

El estudio de las redes de mundo pequeño, de amplia repercusión tanto en el campo de las ingenierías telemática e informática como en otras áreas de las ciencias sociales y la naturaleza, ha permitido entre otros logros esclarecer la dinámica de propagación de virus (biológicos o virtuales) como así también la reducción del planeta en materia de vínculos o la posibilidad de disminuir el diámetro de las grandes redes telemáticas.

Una red social presenta el "fenómeno de mundo pequeño" si, en términos generales, dos individuos cualesquiera en la red pueden ser conectados a través de una cadena corta de intermediarios.

La medición de este fenómeno en una red es sencilla, consiste en hallar las distancias entre pares de vértices que la componen y calcular su distancia media. Uno de los métodos más conocidos es el de la "búsqueda de anchura". En una red social, una distancia de un grado es la que conecta a un individuo con un amigo o conocido, la relación es directa y sin intermediarios. Una distancia de dos grados, conecta a un individuo con un amigo de su amigo; una de tres, a un individuo con el amigo del amigo del amigo, y así sucesivamente.

En particular, la teoría de los "Seis Grados de Separación" conjetura que cualquier persona puede estar conectada con cualquier otra, a través de una cadena de no más de cinco intermediarios en promedio que se conocen mutuamente dos a dos, conectando a ambas personas con sólo seis enlaces o saltos.

4.1.2. Origen y experimentos

La idea del fenómeno de mundo pequeño fue planteada por primera vez en los años 20 por el escritor húngaro Frigyes Karinthy y convertido en un problema de investigación en los años 50 por el matemático austríaco Manfred Kochen (1928-1989), el sociólogo estadounidense Ithiel de Sola Pool (1917-1984) y el psicólogo social estadounidense Stanley Milgram (1933-1984).

A principios de los 50, Kochen M. y Pool, I. [39] escriben un manuscrito que circuló por más de 20 años entre los investigadores antes de ser publicado en 1978. El trabajo articula formalmente la mecánica de las redes sociales, y explora sus consecuencias matemáticas (incluyendo sus grados de conectividad), además de dejar el interrogante sobre el número de grados de separación en las redes sociales reales.

En 1967, Stanley Milgram inició un experimento en la Universidad de Harvard en Cambridge, Massachussets, orientado a profundizar en los planteos de Kochen y Pool. Su trabajo *Small World Phenomenon* [36] es una de las publicaciones científicas más citadas en el estudio de las redes sociales. Su idea fue conocer la probabilidad de que dos personas al azar estén conectadas a través de conocidos comunes. El experimento consistió en enviar una carta a una muestra al azar de personas de una ciudad de Estados Unidos, solicitándoles a cada una de ellas remitirla a cierta persona objetivo. El destinatario también fue elegido al azar pero residía en otra ciudad, por lo que era de esperar que entre los emisores y receptores no hubiera relación. Los participantes tenían cierta información sobre el destinatario, aunque no podían enviar la carta directamente a la persona objetivo, sino que debían enviarla a un amigo o conocido con mayor probabilidad de conocerla. De este modo, se fueron generando cadenas de intermediarios que permitieron el análisis de ciertas propiedades de la estructura social.

La experiencia arrojó un promedio de cinco intermediarios antes de alcanzar el destino final, lo que significa una distancia final igual a seis, por lo que los investigadores concluyeron que la población de Estados Unidos estaba separada por seis grados o saltos

en promedio. Por otra parte, prácticamente la mitad de las cadenas llegaron a su destino (en el último paso) a través de las mismas tres personas, por lo que pudo demostrarse la presencia de una red de relaciones densa, y que la cadena de intermediarios para contactar a dos personas cualesquiera era relativamente corta.

Posteriormente Milgram reanudó el experimento con una modificación, el destinatario y el emisor fueron elegidos entre miembros de distintas comunidades raciales, con la intuición de que el trayecto se alargaría. No obstante, la distancia media fue otra vez igual a seis dando pie a la hipótesis de los "Seis Grados de Separación". Aunque Milgram nunca utilizó esta frase, la misma se popularizó a raíz de su uso por parte del dramaturgo americano John Guare J. [25], en su libro *Six Degrees of Separation* publicado en 1990.

Los resultados obtenidos por Milgram muestran las propiedades de las "redes de mundo pequeño" que luego son formalizadas por Watts D. y Strogatz S. en 1998 [46], quienes argumentan que se trata de una estructura que puede representarse con un grafo caracterizado por un alto agrupamiento y una corta distancia promedio, justificando que "el universo social estaría constituido por conglomerados de individuos altamente interconectados entre sí, pero que no llegan a conformar islas separadas, en la medida en que se enlazan a través de vértices que disminuyen considerablemente la distancia geodésica promedio". Para analizar el fenómeno de mundo pequeño, partieron de redes regulares donde cada vértice tenía el mismo número de enlaces, como puede observarse en la Figura 4.1, distribuyendo los vértices sobre un círculo y uniéndolos con los más próximos tanto por la derecha como por la izquierda. Estos grafos, presentan un alto grado de agrupamiento y un diámetro elevado. Luego, a estas redes les fueron aplicando alteraciones en forma sucesiva hasta convertirlas en redes aleatorias, con bajo agrupamiento y diámetro reducido como en la Figura 4.3, llegando a la conclusión de que las redes de mundo pequeño, es decir, redes con un diámetro pequeño y un agrupamiento alto ocupaban un lugar en medio del espectro entre las redes regulares y las redes aleatorias, como puede verse en la Figura 4.2.

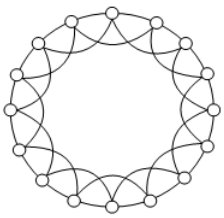


Figura 4.1

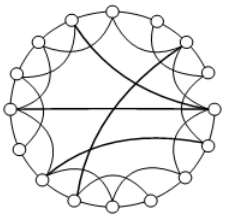


Figura 4.2

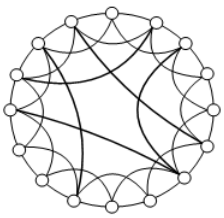


Figura 4.3

Entre los ejemplos de redes de mundo pequeño encontramos, en el universo del cine, el llamado juego de Bacon que consiste en determinar la distancia de cualquier actor a Kevin Bacon. Se sospecha que ningún actor de Hollywood tiene distancia a Bacon mayor a seis.

La comunidad matemática, por ejemplo, cuenta con un epicentro: el matemático húngaro Paul Erdös. La producción de Erdös es tan grande, computada en el número de publicaciones, que en la literatura se habla de la distancia a Erdös. Un matemático tiene distancia a Erdös 1 si ha publicado un artículo en colaboración con él, distancia 2 si ha publicado con un matemático a distancia Erdös 1, y así sucesivamente.

Entre otros ejemplos de redes de mundo pequeño, encontramos el mapa de propagación de epidemias de un área geográfica o la red de suministro de una compañía eléctrica.

Cualquiera de las redes descriptas anteriormente puede ser representada por un grafo simple, es decir un grafo sin bucles ni aristas paralelas. Así, en el llamado grafo de colaboración de Erdös, los vértices designan matemáticos y las aristas, las colaboraciones. Igualmente puede esbozarse un grafo descriptivo de la sociedad de actores, indicando las aristas como intervenciones en un mismo film.

El mundo puede ser representado por un grafo con pautas de regularidad. Así, para un individuo dado puede aproximarse, dentro de su círculo de amistades, un dominio de cien personas. Si cada uno de estos individuos conoce a cien personas y volvemos a multiplicar por cien tendremos, después de una cadena de seis eslabones, un número de personas igual a cien a la sexta, es decir, un billón, número que abarca la superficie entera del planeta. Sin embargo, en la realidad, los amigos de nuestros amigos suelen ser nuestros amigos, o al menos caer en el círculo de nuestros allegados, de tal manera que la cadena no se multiplica cada vez por cien sino por un factor más pequeño. Por otro lado, el círculo puede encerrar saltos a zonas remotas, en el sentido de amistades poco probables.

Existen múltiples detractores de la Teoría de Mundo Pequeño, considerándola únicamente como el reflejo del dicho popular de que "el mundo es un pañuelo". No obstante, ésta sí podría tener plena validez aplicada a poblaciones con ciertos rasgos o patrones básicos comunes: de comportamiento, indicadores sociodemográficos o perfiles descriptivos, acotando la teoría a un escenario posible y plausible.

4.1.3. Mundo pequeño y Facebook

Un estudio realizado en Estados Unidos en el año 2011 por Ugander J., Karrer B., Backstrom L. y Marlow C. [44] analizó el fenómeno de mundo pequeño en Facebook. Por primera vez, la distancia entre individuos de todo el mundo se ha podido medir a partir de datos del mundo real provenientes de una enorme población, ya que al momento del estudio, Facebook contaba con 720 millones de usuarios activos (más del 10 por ciento de la población mundial) y con 69 mil millones de amistades entre ellos.

En dicho estudio, se midió en primer lugar el número de amigos que tenía cada usuario y se encontró que esta distribución difería significativamente de los estudios previos realizados sobre redes sociales a gran escala. Luego se mostró que dos usuarios de Facebook elegidos al azar de cualquier parte del mundo estarían separados por una distancia media

de 4.7 grados o relaciones, poco más de tres individuos. Si se consideraban sólo los usuarios de un mismo país la distancia era menor (3 grados). Este hallazgo fue significativo ya que disminuyó la distancia tradicional de "seis pasos" hallada por Milgram en los años 50, posteriormente un fenómeno universal. Además, pudo observarse que los amigos de un usuario eran más propensos a ser de la misma edad y proceder del mismo país.

En la Figura 4.4 podemos ver la distribución acumulada del grado que muestra en el eje de absisas el número de amigos y en el de ordenadas el porcentaje de personas que tiene menos de un determinado número de amigos.

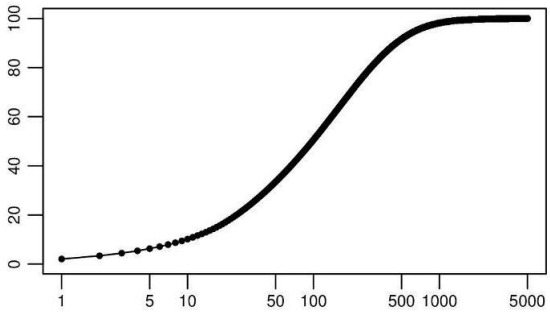


Figura 4.4

Se observó que sólo el 10 por ciento de la población tenía menos de 10 amigos, el 20 por ciento menos de 25, mientras que el 50 por ciento tenía más de 100. Además se obtuvo que la distribución era sesgada y que el número promedio de amigos por usuario era de 190. Un importante hallazgo en este estudio fue sin embargo que la distribución no era tan sesgada como sugerían estudios anteriores sobre redes sociales.

En consecuencia, los autores concluyen que Facebook es a la vez una red social global y local, conecta a gente que está apartada pero también tiene una estructura local densa como vemos en las pequeñas comunidades. Ha crecido rápidamente en estos últimos años, representando una proporción cada vez mayor de la población mundial y por lo tanto el mundo se ha vuelto más conectado.

4.2. Ciclos y conectividad en muestras de perfiles de Facebook de gran tamaño

En esta Sección, presentamos un estimador de utilidad para analizar el nivel de conectividad en la red social Facebook. Hemos visto que cualquier red social puede ser modelada por un grafo donde los vértices son los actores y las relaciones que los conectan están representadas por las aristas. En particular, Facebook puede representarse por un grafo cuyos vértices son los perfiles y las relaciones de amistad, las aristas. Luego, para poder analizar la conectividad entre los perfiles, y por lo tanto el fenómeno de mundo pequeño en

la red, resulta de utilidad estimar la proporción de ciclos de cierta longitud que se presentan en muestras de perfiles, como así también estudiar la distribución asintótica de dicho estimador cuando el tamaño de la muestra tiende a infinito, lo cual implica considerar a todos los perfiles creados en la red.

Sea $\mathbb{A}_N = (\alpha_{ij})_{i,j\in\{1,\dots,N\}}$ la matriz de amistad que corresponde a una muestra de N perfiles de $\mathcal{F}_\infty$, con $\alpha_{ij} := \alpha_\infty(f_i, f_j)$. Vimos que bajo la hipótesis de TC, α_{ij} tiene distribución Bernoulli de parámetro p.

Para $r = 1, 2, \dots, N-1$, sea

$$C = \left\{ \vec{i} = (i_1, \dots, i_r) : \prod_{l=1}^{r} \alpha_{i_l i_{l+1}} = 1, \text{con } i_1, \dots, i_r \in \{1, \dots, N\}, i_{r+1} := i_1 \right\},$$

el conjunto de caminos que forman ciclos de longitud r y, utilizando la siguiente notación para la suma sobre cada uno de los índices $i_1, \dots, i_r \in \{1, \dots, N\}$,

$$\sum_{i_1, i_2, \dots, i_r = 1}^{N} := \sum_{i_1=1}^{N} \cdots \sum_{i_r=1}^{N},$$

sea

$$T_N(r) = \frac{1}{N^r} \sum_{i_1, i_2, \dots, i_r = 1}^{N} \alpha_{i_1 i_2}\, \alpha_{i_2 i_3} \dots \alpha_{i_r i_1}, \tag{4.1}$$

la proporción de ciclos de longitud r en una muestra de perfiles de tamaño N.

Luego, dada una muestra de N perfiles, el cardinal del conjunto C, $card\ \{C\}$, indica el número de ciclos que se producen en esa muestra y $\frac{card\ \{C\}}{N^r}$ su proporción. Por lo tanto, mediante esta cantidad podremos inferir acerca del nivel de conectividad existente en la muestra de perfiles. Por ejemplo, si para un valor de r grande en relación a N obtenemos un valor pequeño de $\frac{card\ \{C\}}{N^r}$, este resultado estaría indicando un nivel de conectividad bajo en la red.

Hallaremos, suponiendo un conexto de TC, la distribución asintótica de la proporción de ciclos de longitud r en muestras de perfiles suficientemente grandes. Para ello, utilizaremos elementos de la teoría de matrices aleatorias y grafos aleatorios basándonos en Domínguez Molina, J. y Rocha Arteaga, A. [18] y Diestel, R. [17], como así también los resultados obtenidos por Wigner E. P. ([52], [51]), Girko V. L. [23] y Sinai Ya. y Soshnikov A.[43] sobre la traza de potencias de matrices aleatorias.

4.2.1. Elementos de la teoría de grafos

En el primer Capítulo hemos presentado algunos conceptos sobre la teoría de grafos que ahora ampliaremos debido a su utilidad como herramienta en las demostraciones que siguen.

Representamos a un grafo G como un par ordenado $(V(G), E(G))$, con $V(G)$ el conjunto de vértices y $E(G)$ el conjunto de aristas, siendo $card\{V(G)\}$ y $card\{E(G)\}$ los cardinales de los respectivos conjuntos.

Enunciamos el siguiente lema cuya demostración puede verse en el trabajo de Bondy Z. y Murty U. [8].

Lema 4.2.1 *Si $G = (V(G), E(G))$ es un grafo conexo entonces*

$$card\{V(G)\} \leq card\{E(G)\} + 1.$$

Además, $card\{V(G)\} = card\{E(G)\} + 1$ si y sólo si G es árbol.

Es decir, los grafos conexos no pueden tener más vértices que aristas y, si el grafo conexo es árbol, el número de vértices supera en uno al número de aristas.

Un *árbol con raíz* es un árbol al que se le ha especificado un vértice al que llamaremos *raíz* y un *árbol orientado con raíz* es un árbol con raíz inmerso en un plano, cuyo contorno se recorre sobre el plano siempre en un mismo sentido previamente convenido, de la siguiente manera: el recorrido inicia y termina en la raíz, al término del recorrido se recorre la misma cantidad de aristas de ida que de vuelta y en cada paso del recorrido no se tienen más aristas recorridas de vuelta que de ida.

La siguiente proposición, cuya demostración puede verse en el Capítulo 6 de Koshy, T. [29], es de especial relevancia para la demostración de uno de los teoremas más importantes utilizados en este Capítulo.

Proposición 4.1 *El número de árboles orientados con raíz que se pueden formar con k aristas coincide con el k-ésimo número de Catalan*

$$\mathcal{C}_k = \frac{1}{k+1}\, C_{2k}^k, \quad k = 1, 2, ...,$$

con $C_{2k}^k = \frac{(2k)!}{k!\,k!}$, el número de subconjuntos con k elementos tomados de un conjunto con $2k$ elementos.

Es decir, los números de Catalan cuentan el total de árboles orientados con raíz que se pueden formar con un número fijo de aristas.

A cada grafo dirigido G se le asocia un grafo con el mismo conjunto de vértices simplemente reemplazando cada arista dirigida por una arista con los mismos extremos. A este grafo se lo denomina *grafo subyacente* del grafo dirigido G. Un grafo dirigido es conexo si y sólo si su grafo subyacente es conexo.

El *esqueleto* del grafo dirigido G es el grafo que se obtiene del grafo subyacente de G al eliminar las aristas paralelas.

4.2.2. Distribución asintótica de la traza de potencias de matrices de Wigner

Uno de los principales objetivos de la teoría de matrices aleatorias es el estudio de la convergencia de la distribución espectral empírica de ciertas matrices aleatorias cuadradas cuando su dimensión tiende a infinito.

Sea A_N una matriz aleatoria simétrica de orden N y las variables aleatorias reales $\lambda_1, \ldots, \lambda_N$, sus autovalores. La *función de distribución espectral empírica* de A_N,

$$F_{A_N}(\lambda) = \frac{1}{N}\ card\ \{\lambda_i \leq \lambda : i = 1, \ldots, N\},\ \lambda \in \mathbb{R}, \tag{4.2}$$

es una función de distribución aleatoria que representa la proporción de autovalores de A_N menores o iguales que λ.

En 1958, Wigner E. [52] introdujo los clásicos conjuntos de matrices aleatorias $A_N = (A_{ij})_{i,j\in\{1,\ldots,N\}}$, comunmente llamadas *matrices de Wigner*. Las componentes de las matrices reales y simétricas de orden N, $A_{ij} = A_{ji} = \frac{\xi_{ij}}{\sqrt{N}}$, son tales que:

i) $\{\xi_{ij}\}_{1\leq i\leq j\leq N}$ son variables aleatorias independientes;

ii) Las leyes de distribución para ξ_{ij} son simétricas;

iii) Cada momento $E[\xi_{ij}^r]$ existe y $E\left|\xi_{ij}^r\right| \leq \mathbb{C}_r$, con $\mathbb{C}_r$ una constante que depende sólo de r;

iv) Los segundos momentos de ξ_{ij} son iguales a $\frac{1}{4}$ si $i < j$, y están uniformemente acotados si $i = j$.

Observación 4.2 *La condición ii) implica que todos los momentos de orden impar se anulan.*

Estudiando la función de distribución espectral empírica de A_N, $F_{A_N}(\lambda)$, Wigner ([52], [51]) probó convergencia en probabilidad de los momentos de $F_{A_N}(\lambda)$ a los momentos de la función de distribución no aleatoria

$$F(\lambda) = \frac{2}{\pi}\int_{-\infty}^{\lambda} \sqrt{1-x^2}\ \mathbb{1}_{[-1,1]}(x)\ dx, \tag{4.3}$$

siendo $F(\lambda)$ la *función de distribución del semicírculo*. Su trabajo estuvo motivado en la mecánica cuántica, donde logró explicar el comportamiento estadístico de los niveles de energía de un sistema físico en términos de los autovalores cuando la dimensión de las matrices es grande.

Por un lado, los momentos de la distribución espectral empírica verifican que

$$\begin{aligned}\int_{-\infty}^{\infty} \lambda^r \, dF_{A_N}(\lambda) &= \sum_{j=1}^{N} (\lambda_j)^r [F_{A_N}(\lambda_j) - F_{A_N}(\lambda_j -)] \\ &= \frac{1}{N} \sum_{j=1}^{N} (\lambda_j)^r \\ &= \frac{1}{N} \, tr(A_N^r),\end{aligned} \tag{4.4}$$

para $r = 0, 1, 2, \ldots$, donde $tr(A_N^r)$ es la traza de la potencia r-ésima de la matriz A_N.

Por otro lado la función de distribución del semicírculo $F(\lambda)$ tiene soporte acotado y está únicamente determinada por sus momentos

$$\int_{-\infty}^{\infty} \lambda^r \, dF(\lambda) = \begin{cases} \int_{-\infty}^{\infty} \lambda^r \, \frac{2}{\pi} \sqrt{1-\lambda^2} \, \mathbb{I}_{[-1,1]}(\lambda) \, d\lambda, & \text{si } r = 2s, \\ 0, & \text{si } r = 2s+1, \end{cases} \tag{4.5}$$

con $s = 0, 1, 2, \ldots$, y $\int_{-\infty}^{\infty} \lambda^r \, \frac{2}{\pi} \sqrt{1-\lambda^2} \, \mathbb{I}_{[-1,1]}(\lambda) \, d\lambda = \mathcal{C}_s$, siendo $\mathcal{C}_s = \frac{1}{s+1} \, C_{2s}^s$ los números de Catalan.

Observemos que los momentos impares de $F(\lambda)$ son cero por la simetría de la distribución.

Lo que Wigner ([52], [51]) prueba entonces es que

$$\frac{1}{N} \, tr(A_N^r) \xrightarrow{Pr} \mu_r := \begin{cases} \mathcal{C}_s \, (\frac{1}{4})^s, & \text{si } r = 2s, \\ 0, & \text{si } r = 2s+1, \end{cases} \tag{4.6}$$

cuando $N \to \infty$.

Más tarde, bajo condiciones más generales Arnold L. y Girko V. L. ([2], [23]) prueban convergencia casi segura en (4.6), es decir,

$$\frac{1}{N} \, tr(A_N^r) \xrightarrow{cs} \mu_r, \tag{4.7}$$

cuando $N \to \infty$ y a esto se lo llamó la *Ley del Semicírculo de Wigner*.

Existen varias versiones de este resultado, dependiendo de las condiciones que se consideren sobre las matrices aleatorias.

Dominguez Molina J. y Rocha Arteaga A. [18] dan una demostración detallada de (4.7) probando para esto que

$$E\left(\frac{1}{N} tr(A_N^r)\right) \longrightarrow \mu_r, \tag{4.8}$$

cuando $N \to \infty$, utilizando el enfoque combinatorio de grafos. Argumentan que la clave es que la traza presenta una estructura cíclica que permite asociarle grafos conexos, de manera que al tomar límite en (4.8), quedarán sólo aquellos grafos conexos correspondientes a los árboles orientados con raíz que se pueden formar con r aristas, y esta cantidad coincide con el r-ésimo número de Catalan $\mathcal{C}_r$.

Según Girko V. L [23] debido a la fuerte correlación entre autovalores, las fluctuaciones $\frac{1}{N}\sum_{i=1}^{N}\lambda_i^r - \mu_r$ son del orden de $\frac{1}{N}$ y

$$tr(A_N^r) - E(tr(A_N^r)) \overset{w}{\Rightarrow} \mathcal{N}\Big(0, Var(tr(A_N^r))\Big), \tag{4.9}$$

cuando $N \to \infty$, con r fijo, pudiéndose hallar una expresión explícita para la $Var(tr(A_N^r))$ que dependerá del segundo y el cuarto momento de ξ_{ij}.

En 1998, Sinai Ya. y Soshnikov A. [43] demuestran el siguiente teorema que extiende los resultados anteriores al caso de potencias r-ésimas de A_N, que crecen con N.

Teorema 4.2.2 *Consideremos los conjuntos de matrices aleatorias simétricas de Wigner que verifican las condiciones i) a iv) antes enunciadas, con el supuesto adicional de que*

$$E(\xi_{ij}^{2k}) \leq (\mathbb{C}.k)^k,$$

uniformemente en i,j y k, con $k = 1, 2, \ldots$ y $\mathbb{C}$ una constante positiva, significando esto que los momentos de ξ_{ij} crecen pero no más rápido que los gaussianos. Entonces si $1 << r << \sqrt{N}$,

$$E(tr(A_N^r)) = \begin{cases} \frac{1}{\sqrt{\pi}}\ \frac{N}{s^{3/2}}\ (1 + o(1)), & si\ r = 2s, \\ 0, & si\ r = 2s+1 \end{cases} \tag{4.10}$$

y

$$tr(A_N^r) - E(tr(A_N^r)) \overset{w}{\Rightarrow} \mathcal{N}\left(0, \frac{1}{\pi}\right), \tag{4.11}$$

cuando $N \to \infty$.

Observación 4.3 *Sinai Ya. y Soshnikov A. [43] remarcan que la técnica que utilizan para la demostración de este resultado puede ser modificada para extender el mismo al caso de variables aleatorias ξ_{ij}, $i \leq j$, no necesariamente simétricamente distribuidas, imponiendo la siguiente condición:*

$$\left|E(\xi_{ij}^k)\right| \leq (\mathbb{C}.k)^k, \tag{4.12}$$

con $k = 1, 2, \ldots$ y $\mathbb{C}$ una constante positiva.

4.2.3. Convergencia débil del estadístico $T_N(r)$

Nuestro objetivo ahora consiste en obtener la distribución asintótica de

$$T_N(r) = \frac{1}{N^r} \sum_{i_1,i_2,\ldots,i_r=1}^{N} \alpha_{i_1 i_2}\, \alpha_{i_2 i_3} \ldots \alpha_{i_r i_1}, \tag{4.13}$$

en el contexto de TC. Para ello, nos serán de utilidad los resultados presentados sobre la distribución asintótica de la traza de potencias de matrices de Wigner.

Considerando la regla del producto de dos matrices N-compatibles C y D,

$$(CD)_{ij} = \sum_{k=1}^{N} C_{ik}\, D_{kj},$$

podemos expresar a la traza de la r-ésima potencia de una matriz de orden N cualquiera A_N como:

$$tr(A_N^r) \;=\; \sum_{i_1=1}^{N} A_{i_1 i_1}^r = \sum_{i_1,i_2=1}^{N} A_{i_1 i_2} A_{i_2 i_1}^{r-1} = \cdots = \sum_{i_1,\ldots,i_r=1}^{N} A_{i_1 i_2} A_{i_2 i_3} \ldots A_{i_r i_1}.$$

Sea el estadístico

$$T'_N(r) = \frac{1}{N^r} \sum_{i_1,i_2,\ldots,i_r=1}^{N} \beta_{i_1 i_2}\, \beta_{i_2 i_3} \ldots \beta_{i_r i_1}, \tag{4.14}$$

con β_{ij} las funciones aleatorias de amistad centradas, es decir $\beta_{ij} := \alpha_{ij} - p$.

Luego, tomando $A_{ij} = \frac{\xi_{ij}}{\sqrt{N}}$, con $\xi_{ij} = \frac{\beta_{ij}}{2\sqrt{p(1-p)}}$, podemos hallar la distribución asintótica de $T'_N(r)$.

Teorema 4.2.3 *Supongamos que ξ_{ij} verifica las condiciones i), iii) y iv) enunciadas en la subsección anterior, con el supuesto adicional de que*

$$\left|E(\xi_{ij}^r)\right| \leq (\mathbb{C}.r)^r, \tag{4.15}$$

para $r = 1, 2, \ldots, N-1$ y $\mathbb{C}$ una constante positiva. Entonces, si $1 << r << \sqrt{N}$,

$$T'_N(r) \overset{w}{\Rightarrow} \mathcal{N}\left(\left(2\sqrt{\frac{p(1-p)}{N}}\right)^r E(tr(A_N^r)), \left(4\,\frac{p(1-p)}{N}\right)^r Var(tr(A_N^r))\right), \tag{4.16}$$

cuando $N \to \infty$, con $E(tr(A_N^r)) \to \begin{cases} \frac{1}{\sqrt{\pi}}\, \frac{N}{(\frac{r}{2})^{3/2}}\, (1+o(1)), & r \text{ par}, \\ 0, & r \text{ impar}, \end{cases}$ cuando $N \to \infty$ y $Var(tr(A_N^r)) \to \frac{1}{\pi}$, cuando $N \to \infty$.

Demostración

i) $\{\xi_{ij}\}_{1\leq i\leq j\leq N}$ son variables aleatorias independientes, ya que las variables $\{\alpha_{ij}\}_{1\leq i\leq j\leq N}$ son independientes;

iii) Cada momento $E(\xi_{i,j}^r) = \frac{(1-p)(-p)^r+p(1-p)^r}{2^r\left(\sqrt{p(1-p)}\right)^r}$ existe y $E(\left|\xi_{ij}^r\right|) \leq \frac{1}{2^r\left(\sqrt{p(1-p)}\right)^r} = \mathbb{C}_r$, con $\mathbb{C}_r$ una constante que depende sólo de r;

iv) Los segundos momentos de ξ_{ij} son iguales a $\frac{1}{4}$ si $i<j$, y si $i=j$, $E(\xi_{ii}^2) = \frac{1}{4}(\frac{1}{p}-1)$.

Además, para $r = 1, 2, \ldots, N-1$,

$$\begin{aligned}
\left|E(\xi_{ij}^r)\right| &\leq \frac{|(-1)^r p^r(1-p)|}{2^r\left(\sqrt{p(1-p)}\right)^r} + \frac{|p(1-p)^r|}{2^r\left(\sqrt{p(1-p)}\right)^r} \\
&\leq \frac{1}{2}\left(\frac{1}{2\sqrt{p(1-p)}}\right)^r \\
&< \left(\frac{1}{2\sqrt{p(1-p)}}.r\right)^r \\
&= (\mathbb{C}.r)^r,\ \mathbb{C}>0.
\end{aligned} \tag{4.17}$$

Luego, si $1 << r << \sqrt{N}$, por el Teorema 4.2.2

$$tr(A_N^r) \overset{w}{\Rightarrow} \mathcal{N}\Big(E(tr(A_N^r)), Var(tr(A_N^r))\Big), \tag{4.18}$$

cuando $N\to\infty$, siendo

$$\begin{aligned}
tr(A_N^r) &= \frac{1}{(\sqrt{N})^r}\sum_{i_1,i_2,\ldots,i_r=1}^{N}\xi_{i_1i_2}\xi_{i_2i_3}\ldots\xi_{i_ri_1} \\
&= \frac{1}{(\sqrt{N})^r}\frac{1}{2^r\left(\sqrt{p(1-p)}\right)^r}\sum_{i_1,i_2,\ldots,i_r=1}^{N}\beta_{i_1i_2}\beta_{i_2i_3}\ldots\beta_{i_ri_1} \\
&= \frac{1}{(\sqrt{N})^r}\frac{1}{2^r\left(\sqrt{p(1-p)}\right)^r}N^r\ T_N'(r).
\end{aligned} \tag{4.19}$$

Observemos que (4.17) es la condición necesaria para extender la demostración que realiza Sinai Ya. y Soshnikov A. [43] al caso de variables aleatorias cuyas leyes no son necesariamente simétricas como es el caso de ξ_{ij}, según indica la fórmula (4.12).

Por lo tanto,

$$T_N'(r) \overset{w}{\Rightarrow} \mathcal{N}\left(\left(2\sqrt{\frac{p(1-p)}{N}}\right)^r E(tr(A_N^r)), \left(4\,\frac{p(1-p)}{N}\right)^r Var(tr(A_N^r))\right),$$

cuando $N \to \infty$, ya que $T'_N(r) = \left(2\sqrt{\frac{p(1-p)}{N}}\right)^r tr(A_N^r)$.

Para estudiar el comportamiento de $E(tr(A_N^r))$ y $Var(tr(A_N^r))$ cuando $N \to \infty$ daremos la idea de la demostración que puede verse en detalle en el trabajo de Domínguez Molina J. y Rocha Arteaga A. [18].

En efecto, sea $\vec{i} := (i_1, \ldots, i_r) \in \{1, \ldots, N\}^r$, $i_{r+1} := i_1$ y $Q(\vec{i}) := E(\beta_{i_1 i_2}\beta_{i_2 i_3} \ldots \beta_{i_r i_1})$. Luego,

$$\frac{1}{N}\, E(tr(A_N^r)) = \frac{1}{N^{\frac{r}{2}+1}}\, \frac{1}{2^r \left(\sqrt{p(1-p)}\right)^r} \sum_{\vec{i}} Q(\vec{i}). \tag{4.20}$$

Identificamos cada índice $\vec{i} := (i_1, \ldots, i_r)$ de la suma anterior con un grafo conexo dirigido

$$G(\vec{i}) = (V(\vec{i}), E(\vec{i})),$$

con $V(\vec{i}) = \{i_1, \ldots, i_r\}$ el conjunto de vértices y $E(\vec{i}) = \{(i_1, i_2), \ldots, (i_r, i_1)\}$ el conjunto de aristas dirigidas.

Luego, a cada grafo $G(\vec{i})$ le asociamos su esqueleto

$$\widetilde{G}(\vec{i}) = (\widetilde{V}(\vec{i}), \widetilde{E}(\vec{i})),$$

que corresponde al grafo subyacente de $G(\vec{i})$ sin aristas paralelas.

Descomponemos (4.20) según el número de vértices que tienen los esqueletos $\widetilde{G}(\vec{i})$ correspondientes a los grafos $G(\vec{i})$

$$\frac{1}{N^{\frac{r}{2}+1}} \sum_{\vec{i}} \frac{Q(\vec{i})}{2^r \left(\sqrt{p(1-p)}\right)^r} = \tag{4.21}$$

$$= \frac{1}{N^{\frac{r}{2}+1}} \left(\sum_{\substack{G(\vec{i}) \\ 1 \le card\{\widetilde{V}(\vec{i})\} \le \left[\frac{r}{2}\right]+1}} \frac{Q(\vec{i})}{2^r \left(\sqrt{p(1-p)}\right)^r} + \sum_{\substack{G(\vec{i}) \\ card\{\widetilde{V}(\vec{i})\} \ge \left[\frac{r}{2}\right]+2}} \frac{Q(\vec{i})}{2^r \left(\sqrt{p(1-p)}\right)^r} \right).$$

El segundo sumando es igual a cero, ya que la suma se realiza sobre los índices de grafos que tienen aristas que no son paralelas a ninguna otra y por lo tanto

$$Q(\vec{i}) = E(\beta_{i_l i_{l+1}}) E(\beta_{i_1 i_2} \ldots \beta_{i_{l-1} i_l} \beta_{i_{l+1} i_{l+2}} \ldots \beta_{i_r i_1}) = 0.$$

Entonces la expresión (4.20) se reduce a

$$\frac{1}{N}\, E(tr(A_N^r)) = \frac{1}{N^{\frac{r}{2}+1}} \sum_{\substack{G(\vec{i}) \\ 1 \le card\{\widetilde{V}(\vec{i})\} \le \left[\frac{r}{2}\right]+1}} \frac{Q(\vec{i})}{2^r \left(\sqrt{p(1-p)}\right)^r}$$

y en consecuencia,

$$\left|\frac{1}{N}E(tr(A_N^r))\right| \leq \frac{1}{N^{\frac{r}{2}+1}} \sum_{\substack{G(\vec{i}) \\ 1\leq card\{\widetilde{V}(\vec{i})\}\leq\left[\frac{r}{2}\right]+1}} \frac{|Q(\vec{i})|}{2^r\left(\sqrt{p(1-p)}\right)^r}. \tag{4.22}$$

Por hipótesis, de la desigualdad (4.17) tenemos que $|E(\xi_{ij}^r)| \leq \left(\frac{1}{2\sqrt{p(1-p)}}.r\right)^r$, es decir, $\frac{|E(\beta_{ij}^r)|}{2^r\left(\sqrt{p(1-p)}\right)^r} \leq \left(\frac{1}{2\sqrt{p(1-p)}}.r\right)^r$, por lo que reemplazando en (4.22)

$$\begin{aligned}
\left|\frac{1}{N}E(tr(A_N^r))\right| &\leq \frac{1}{N^{\frac{r}{2}+1}} \sum_{\substack{G(\vec{i}) \\ 1\leq card\{\widetilde{V}(\vec{i})\}\leq\left[\frac{r}{2}\right]+1}} \left(\frac{1}{2\sqrt{p(1-p)}}.r\right)^r \\
&\leq \frac{1}{N^{\frac{r}{2}+1}} \sum_{j=1}^{\left[\frac{r}{2}\right]+1} \sum_{\substack{G(\vec{i}) \\ card\{\widetilde{V}(\vec{i})\}=j}} \left(\frac{1}{2\sqrt{p(1-p)}}.r\right)^r \\
&\leq \frac{1}{N^{\frac{r}{2}+1}} \left(\left[\frac{r}{2}\right]+1\right) N^{\left[\frac{r}{2}\right]+1} \left(\frac{1}{2\sqrt{p(1-p)}}.r\right)^r.
\end{aligned}$$

Luego, si r es impar, $\left[\frac{r}{2}\right] = \frac{r}{2} - \frac{1}{2}$ y

$$\left|\frac{1}{N}E(tr(A_N^r))\right| \leq \frac{(\frac{r}{2}+\frac{1}{2})}{\sqrt{N}} \left(\frac{1}{2\sqrt{p(1-p)}}.r\right)^r \to 0,$$

cuando $N \to \infty$.
Por otro lado, cuando r es par, $\left[\frac{r}{2}\right] = \frac{r}{2}$ y $\left|\frac{1}{N}E(tr(A_N^r))\right| \leq (\frac{r}{2}+1)\left(\frac{1}{2\sqrt{p(1-p)}}.r\right)^r$. Si descomponemos $\frac{1}{N}E(tr(A_N^r))$ de la siguiente manera

$$\frac{1}{N}E(tr(A_N^r)) = S_1 + S_2,$$

con $S_1 = \frac{1}{N^{\frac{r}{2}+1}} \sum_{\substack{G(\vec{i}) \\ 1\leq card\{\widetilde{V}(\vec{i})\}\leq\frac{r}{2}}} \frac{Q(\vec{i})}{2^r(\sqrt{p(1-p)})^r}$ y $S_2 = \frac{1}{N^{\frac{r}{2}+1}} \sum_{\substack{G(\vec{i}) \\ card\{\widetilde{V}(\vec{i})\}=\frac{r}{2}+1}} \frac{Q(\vec{i})}{2^r(\sqrt{p(1-p)})^r}$,

podemos deducir que $|S_1| \leq \frac{r}{2}\frac{N^{\frac{r}{2}}}{N^{\frac{r}{2}+1}}\left(\frac{1}{2\sqrt{p(1-p)}}.r\right)^r \to 0$, cuando $N \to \infty$ y estudiar el comportamiento de S_2.

En efecto, si $card\{\widetilde{V}(\vec{i})\} = \frac{r}{2}+1$, por el Lema 4.2.1 $card\{\widetilde{E}(\vec{i})\} \geq \frac{r}{2}$. Por un lado, si $card\ \{\widetilde{E}(\vec{i})\} > \frac{r}{2}$, $Q(\vec{i})$ es cero, ya que en este caso $E(\vec{i})$ tendría por lo menos una arista

que no es paralela a ninguna otra. Por otro lado, si $card\ \{\widetilde{E}(\vec{i})\} = \frac{r}{2}$, por el Lema 4.2.1, $\widetilde{G}(\vec{i})$ es árbol y a cada arista de $\widetilde{E}(\vec{i})$ le corresponde una arista paralela en $E(\vec{i})$, siendo $\frac{Q(\vec{i})}{2^r(\sqrt{p(1-p)})^r} = \left[E\left(\frac{\beta_{ij}^2}{2^2(\sqrt{p(1-p)})^2}\right)\right]^{\frac{r}{2}} = \left[E(\xi_{ij}^2)\right]^{\frac{r}{2}} = \left(\frac{1}{4}\right)^{\frac{r}{2}}$.

Además, Domínguez Molina J. y Rocha Arteaga A. [18], prueban que el número de árboles orientados con raíz que pueden formarse con una cantidad cualquiera r de aristas coincide con el r-ésimo número de Catalan, $\mathcal{C}_r = \frac{1}{r+1}\ C_{2r}^r$, con $C_{2r}^r = \frac{(2r)!}{r!r!}$. Por lo expuesto, obtienen que

$$S_2 \to \left(\frac{1}{4}\right)^{\frac{r}{2}}\ \mathcal{C}_{\frac{r}{2}}, \quad \text{cuando } N \to \infty,$$

es decir,

$$\frac{1}{N}E(tr(A_N^r)) \to \begin{cases} \mathcal{C}_s\ \left(\frac{1}{4}\right)^s, & \text{si } r = 2s, \\ 0, & \text{si } r = 2s+1, \end{cases}$$

cuando $N \to \infty$.
La misma conclusión obtienen Sinai Ya. y Soshnikov A. [43], señalando además que $\mathcal{C}_s\ \left(\frac{1}{4}\right)^s = \frac{1}{\sqrt{\pi}}\ \frac{1}{s^{3/2}}(1+o(1))$ y por lo tanto,

$$E(tr(A_N^r)) \to \begin{cases} \frac{N}{\sqrt{\pi}}\ \frac{1}{s^{3/2}}(1+o(1)), & \text{si } r = 2s, \\ 0, & \text{si } r = 2s+1, \end{cases} \tag{4.23}$$

cuando $N \to \infty$.

La $Var(tr(A_N^r))$ está dada por la siguiente expresión:

$$\begin{aligned} Var(tr(A_N^r)) &= E\left(tr(A_N^r)^2\right) - \left(E(tr(A_N^r))\right)^2 \\ &= E\left[\left(\frac{1}{N^{\frac{r}{2}}}\frac{1}{2^r(\sqrt{p(1-p)})^r}\sum_{\vec{i}}\beta_{i_1i_2}\ldots\beta_{i_ri_1}\right)^2\right] - \\ &- \left(\frac{1}{N^{\frac{r}{2}}}\frac{1}{2^r(\sqrt{p(1-p)})^r}\sum_{\vec{i}}E(\beta_{i_1i_2}\ldots\beta_{i_ri_1})\right)^2 \\ &= \frac{1}{N^r\ 2^{2r}\ (\sqrt{p(1-p)})^{2r}}\sum_{\vec{i},\vec{i'}}[Q(\vec{i},\vec{i'}) - Q(\vec{i})Q(\vec{i'})], \end{aligned}$$

siendo $\vec{i} = (i_1, \ldots, i_r)$, $\vec{i'} = (i'_1, \ldots, i'_r)$, $Q(\vec{i},\vec{i'}) = E(\beta_{i_1i_2}\ldots\beta_{i_ri_1}\beta_{i'_1i'_2}\ldots\beta_{i'_ri'_1})$ e identificando a $\vec{i}$, $\vec{i'}$ con el grafo $G(\vec{i},\vec{i'}) = (V(\vec{i},\vec{i'}), E(\vec{i},\vec{i'}))$ de $2r$ vértices y $2r$ aristas.

Para el análisis de los términos de la suma anterior, se los agrupa según el número de vértices de los esqueletos $\widetilde{G}(\vec{i},\vec{i'})$ de los grafos $G(\vec{i},\vec{i'})$ y se llega a la conclusión de que el único aporte a la $Var(tr(A_N^r))$ se da en el caso que $card\{\widetilde{V}(\vec{i},\vec{i'})\} \leq r$. Por lo tanto,

$$\begin{aligned}
Var(tr(A_N^r)) &= \frac{1}{N^r\ 2^{2r}\ (\sqrt{p(1-p)})^{2r}} \sum_{\substack{G(\vec{i},\vec{i'}) \\ 1\leq card\{\widetilde{V}(\vec{i},\vec{i'})\}\leq r}} [Q(\vec{i},\vec{i'}) - Q(\vec{i})Q(\vec{i'})] \\
&\leq \frac{1}{N^r\ 2^{2r}\ (\sqrt{p(1-p)})^{2r}} \sum_{\substack{G(\vec{i},\vec{i'}) \\ 1\leq card\{\widetilde{V}(\vec{i},\vec{i'})\}\leq r}} [|Q(\vec{i},\vec{i'})| + |Q(\vec{i})||Q(\vec{i'})|] \\
&\leq \frac{1}{N^r} \sum_{j=1}^{r} \sum_{\substack{G(\vec{i}) \\ card\{\widetilde{V}(\vec{i})\}=j}} \frac{[|Q(\vec{i},\vec{i'})| + |Q(\vec{i})||Q(\vec{i'})|]}{2^{2r}\ (\sqrt{p(1-p)})^{2r}} \\
&\leq \frac{r}{2^r\ (\sqrt{p(1-p)})^r} \left(\frac{1}{2\sqrt{p(1-p)}}.r\right)^r + r\ \left(\frac{1}{2\sqrt{p(1-p)}}.r\right)^{2r},
\end{aligned}$$

para todo N, en virtud de (4.17).

Del mismo modo Sinai Ya. y Soshnikov A. [43], señalan que si $1 << r << \sqrt{N}$, el aporte a la $Var(tr(A_N^r))$ viene dado por el número de índices en los cuales en la unión de $\vec{i}$ con $\vec{i'}$ cada arista aparece dos veces y concluyen que ese número es $\frac{1}{\pi}\ N^r\ 2^{2r}(1+o(1))$ y que

$$\lim_{N\to\infty} Var(tr(A_N^r)) = \frac{1}{\pi}.$$

□

Finalmente, bajo TC, podemos obtener la distribución asintótica de la proporción de ciclos de longitud r en muestras de N perfiles, representada por el estadístico T_N(r) dado por la expresión (4.13), cuando $r \to \infty$, $N \to \infty$ y $\frac{r}{\sqrt{N}} \to 0$. Para esto, desarrollamos el estadístico $T'_N(r)$ dado por (4.14) en términos de las funciones de amistad α_{ij} de la siguiente manera:

$$\begin{aligned}
T'_N(r) &= \frac{1}{N^r} \sum_{i_1,\ldots,i_r=1}^{N} \beta_{i_1 i_2} \ldots \beta_{i_r i_1} \\
&= \frac{1}{N^r} \sum_{i_1,\ldots,i_r=1}^{N} (\alpha_{i_1 i_2} - p) \ldots (\alpha_{i_r i_1} - p) \qquad (4.24) \\
&= T_N(r) + R_N(r),
\end{aligned}$$

con

$$\begin{aligned} R_N(r) &= -\frac{1}{N^r}\, C_r^{r-1}\, p \sum_{i_1,\ldots,i_r=1}^{N} \alpha_{i_1 i_2}\alpha_{i_2 i_3}\ldots\alpha_{i_{r-1} i_r} + \\ &+ \frac{1}{N^r}\, p^2\Bigg[C_r^{r-1} \sum_{i_1,\ldots,i_{r-1}=1}^{N} \alpha_{i_1 i_2}\alpha_{i_2 i_3}\ldots\alpha_{i_{r-2} i_{r-1}} + (C_r^{r-2} - C_r^{r-1}) \sum_{i_1,\ldots,i_r=1}^{N} \alpha_{i_1 i_2}\alpha_{i_3 i_4}\ldots\alpha_{i_{r-1} i_r}\Bigg] - \\ &- \cdots + (-1)^{r-1}\, \frac{1}{N^r}\, C_r^1\, p^{r-1} \sum_{i_1,i_2=1}^{N} \alpha_{i_1 i_2} + (-1)^r p^r. \end{aligned}$$

Además, utilizaremos los resultados obtenidos por Arnold L. y Girko V.L. ([2], [23]) dados por las expresiones (4.7) y (4.9) respectivamente. Es decir, utilizaremos que

$$E(T'_N(r)) \to \begin{cases} \left(2\sqrt{\frac{p(1-p)}{N}}\right)^r N\, \mathcal{C}_{\frac{r}{2}} (\frac{1}{4})^{\frac{r}{2}}, & \text{si } r \text{ es par,} \\ 0, & \text{si } r \text{ es impar,} \end{cases} \tag{4.25}$$

cuando $N \to \infty$, con $\mathcal{C}_r = \frac{1}{r+1} C_{2r}^r$, los números de Catalan, y que para r fijo,

$$T'_N(r) - E(T'_N(r))) \overset{w}{\Rightarrow} \mathcal{N}\Big(0, Var(T'_N(r))\Big), \tag{4.26}$$

cuando $N \to \infty$.

Comenzando con el caso $r = 2$ tenemos

$$T'_N(2) = T_N(2) + R_N(2),$$

con $T_N(2) = \frac{1}{N^2} \sum_{i_1,i_2=1}^{N} \alpha_{i_1 i_2}\alpha_{i_2 i_1}$ la proporción de ciclos de longitud 2 en muestras de N perfiles y $R_N(2) = -\frac{2}{N^2}\, p \sum_{i_1,i_2=1}^{N} \alpha_{i_1 i_2} + p^2 = -\frac{2}{N^2}\, p\Big(\sum_{\substack{i_1,i_2=1 \\ i_1 \neq i_2}}^{N} \alpha_{i_1 i_2} + N\Big) + p^2$.

Por la simetría de las funciones aleatorias de amistad $\alpha_{i_1 i_2}$, escribimos

$$\sum_{\substack{i_1,i_2=1 \\ i_1 \neq i_2}}^{N} \alpha_{i_1 i_2} = 2 \sum_{1 \leq i_1 < i_2 \leq N} \alpha_{i_1 i_2},$$

siendo $\sum_{1 \leq i_1 < i_2 \leq N} \alpha_{i_1 i_2}$ una suma de $\frac{N(N-1)}{2}$ variables Bernoulli independientes con parámetro p, es decir, $\sum_{1 \leq i_1 < i_2 \leq N} \alpha_{i_1 i_2}$ es Binomial de parámetros $\frac{N(N-1)}{2}$ y p. Luego, para N suficientemente grande, $\sum_{\substack{i_1,i_2=1 \\ i_1 \neq i_2}}^{N} \alpha_{i_1 i_2} = 2 \sum_{1 \leq i_1 < i_2 \leq N} \alpha_{i_1 i_2}$ se aproxima a una variable Normal con

$E\left(\sum_{\substack{i_1,i_2=1\\ i_1\neq i_2}}^{N}\alpha_{i_1i_2}\right)=N(N-1)p$ y $Var\left(\sum_{\substack{i_1,i_2=1\\ i_1\neq i_2}}^{N}\alpha_{i_1i_2}\right)=2N(N-1)p(1-p)$. Por lo tanto, $R_N(2)$ se aproxima a una Normal con

$$\begin{aligned} E(R_N(2)) &= -\frac{2}{N^2}\,p\,(N(N-1)p+N)+p^2 \\ &= -\frac{2}{N}\,p^2\,(N-1)-\frac{2}{N}p+p^2 \\ &= \frac{2}{N}p(1-p)-p^2. \end{aligned}$$

Luego, como $T_N(2)=T'_N(2)-R_N(2)$ y, por (4.26), $T'_N(2)$ es asintóticamente Normal con media dada por (4.25), cuando $N\to\infty$, resulta que $T_N(2)$ converge en distribución a una variable Normal y

$$\begin{aligned} E(T_N(2)) &= E(T'_N(2))-\left[\frac{2}{N}p(1-p)-p^2\right] \\ &\to p(1-p)\mathcal{C}_1-\left[\frac{2}{N}p(1-p)-p^2\right], \end{aligned}$$

cuando $N\to\infty$, es decir,

$$E(T_N(2))\to\left(1-\frac{2}{N}\right)p(1-p)+p^2, \tag{4.27}$$

cuando $N\to\infty$.

Para el caso $r=3$,

$$T'_N(3)=T_N(3)+R_N(3),$$

con $T_N(3)=\frac{1}{N^3}\sum_{i_1,i_2,i_3=1}^{N}\alpha_{i_1i_2}\alpha_{i_2i_3}\alpha_{i_3i_1}$ la proporción de ciclos de longitud 3 en muestras de N perfiles y

$$\begin{aligned} R_N(3) &= -\frac{3}{N^3}\,p\sum_{i_1,i_2,i_3=1}^{N}\alpha_{i_1i_2}\alpha_{i_2i_3}+\frac{3}{N^3}\,p^2\sum_{i_1,i_2=1}^{N}\alpha_{i_1i_2}-p^3 \\ &= -\frac{3}{N^3}\,p\left[\sum_{\substack{i_1,i_2,i_3=1\\ i_1\neq i_3}}^{N}\alpha_{i_1i_2}\alpha_{i_2i_3}+\sum_{i_1,i_2=1}^{N}\alpha_{i_1i_2}\alpha_{i_2i_1}\right]+\frac{3}{N^3}\,p^2\left[\sum_{\substack{i_1,i_2=1\\ i_1\neq i_2}}^{N}\alpha_{i_1i_2}+N\right]-p^3 \\ &= -\frac{3}{N^3}\,p\left[\sum_{\substack{i_1,i_2,i_3=1\\ i_1\neq i_3}}^{N}\alpha_{i_1i_2}\alpha_{i_2i_3}+N^2\,T_N(2)\right]+\frac{3}{N^3}\,p^2\left[(R_N(2)-p^2)\,\frac{N^2}{-2p}\right]-p^3. \end{aligned}$$

Hemos visto que para $r = 2$ tanto $T_N(2)$ como $R_N(2)$ tienen distribución aproximadamente Normal, cuando $N \to \infty$. Además,

$$\begin{aligned}\sum_{\substack{i_1,i_2,i_3=1\\ i_1\neq i_3}}^{N} \alpha_{i_1i_2}\alpha_{i_2i_3} &= \sum_{\substack{i_1,i_2,i_3=1\\ i_1\neq i_3\\ i_2\neq i_1,i_2\neq i_3}}^{N} \alpha_{i_1i_2}\alpha_{i_2i_3} + \sum_{\substack{i_1,i_2,i_3=1\\ i_1\neq i_3\\ i_2= i_1}}^{N} \alpha_{i_1i_2}\alpha_{i_2i_3} + \sum_{\substack{i_1,i_2,i_3=1\\ i_1\neq i_3\\ i_2= i_3}}^{N} \alpha_{i_1i_2}\alpha_{i_2i_3} \\ &= \sum_{\substack{i_1,i_2,i_3=1\\ i_1\neq i_3\\ i_2\neq i_1,i_2\neq i_3}}^{N} \alpha_{i_1i_2}\alpha_{i_2i_3} + 2\sum_{\substack{i_1,i_3=1\\ i_1\neq i_3}}^{N} \alpha_{i_1i_3}, \qquad (4.28)\end{aligned}$$

donde el primer sumando de (4.28) es una suma de $N(N-1)(N-2)$ variables con distribución Bernoulli de parámetro p^2 y el segundo una suma de $N(N-1)$ variables independientes Bernoulli de parámetro p. Por lo tanto, cuando N es suficientemente grande, ambos sumandos tienen distribución es aproximadamente Normal. En consecuencia, (4.28) es Normal y $E\left(\sum_{\substack{i_1,i_2,i_3=1\\ i_1\neq i_3}}^{N} \alpha_{i_1i_2}\alpha_{i_2i_3}\right) = N(N-1)(N-2)p^2 + 2N(N-1)p$. Concluimos entonces que $R_N(3)$ converge en distribución a una variable Normal con

$$\begin{aligned}E(R_N(3)) \;\to\; & -\frac{3}{N^3}\, p\Big[N(N-1)(N-2)p^2 + 2N(N-1)p + N^2\, E(T_N(2))\Big] + \\ & + \frac{3}{N^3}\, p^2\left[(E(R_N(2)) - p^2)\,\frac{N^2}{-2p}\right] - p^3,\end{aligned}$$

cuando $N \to \infty$.

Como $T_N(3) = T'_N(3) - R_N(3)$, podemos observar que $T_N(3)$ puede escribirse en términos de las variables asintóticamente Normales $T'_N(3)$, $T'_N(2)$, $R_N(2)$ y $\sum_{\substack{i_1,i_2,i_3=1\\ i_1\neq i_3}}^{N} \alpha_{i_1i_2}\alpha_{i_2i_3}$ de la siguiente manera

$$\begin{aligned}T_N(3) &= T'_N(3) + \frac{3}{N^3}\, p\Big[\sum_{\substack{i_1,i_2,i_3=1\\ i_1\neq i_3}}^{N} \alpha_{i_1i_2}\alpha_{i_2i_3} + N^2\,(T'_N(2) - R_N(2))\Big] - \\ &- \frac{3}{N^3}\, p^2\left[(R_N(2) - p^2)\,\frac{N^2}{-2p}\right] + p^3. \qquad (4.29)\end{aligned}$$

Por lo tanto, $T_N(3)$ converge en distribución a una variable Normal, cuando $N \to \infty$, con $E(T_N(3)) \to E(R_N(3))$, ya que $E(T'_N(3)) \to 0$, cuando $N \to \infty$, en virtud de (4.25).

Repitiendo el procedimiento, podemos deducir que el estadístico $T_N(r)$ puede expresarse como combinación lineal de sumas de variables aleatorias Bernoulli y las variables aleatorias $T'_N(r)$, $T'_N(r-1)$, $\ldots$, $T'_N(2)$, $R_N(2)$ de quienes conocemos que su distribución

asintótica es Normal. Luego, $T_N(r)$ converge en distribución a una variable Normal, cuando $N \to \infty$ y pueden hallarse de manera recursiva su esperanza y su varianza. Además, si suponemos que $1 << r << \sqrt{N}$, como $T'_N(r)$ verifica las hipótesis del Teorema 4.2.3, la normalidad de $T_N(r)$ sigue valiendo para r creciendo con N.

4.3. Grado medio de separación entre dos perfiles de Facebook

En esta sección, nos centramos en el análisis de la viabilidad de la teoría de mundo pequeño en Facebook. Daremos una definición del *grado medio de separación* entre dos perfiles $f_i, f_j \in \mathcal{F}_\infty$ y, mediante su cálculo en los contextos de TC y distintos tipos de segmentación, podremos sacar conclusiones sobre la viabilidad de la teoría mencionada.

Definición 4.4 *Sea $\{f_1, \ldots, f_N\}$ una muestra de N perfiles de $\mathcal{F}_\infty$. Dados $i, j \in \{1, \ldots, N\}$, $i \neq j$, decimos que i está conectado con j mediante k enlaces si*

$$\prod_{l=0}^{k-1} \alpha_\infty(f_{\vec{i_l}}, f_{\vec{i}_{l+1}}) = 1,$$

con $\vec{i}_0 = i$, $\vec{i}_k = j$; $\vec{i_l} \neq \vec{i_{l'}}$ si $l \neq l'$.

Luego, el conjunto

$$D_k(i,j) = \{\vec{i} = (\vec{i_0}, \vec{i_1}, \ldots, \vec{i}_{k-1}, \vec{i_k}) : \vec{i_0} = i, \vec{i_k} = j, \vec{i_l} \neq \vec{i_{l'}} \text{ si } l \neq l'\}, \tag{4.30}$$

representa a todos los caminos posibles con k enlaces y $k-1$ intermediarios que conectan al sitio i con el j y, para los perfiles f_i y f_j de la muestra correspondientes a los sitios i y j, tenemos la siguiente definición.

Definición 4.5 *Definimos el grado de separación entre f_i y f_j como*

$$G_N(f_i, f_j) = \begin{cases} \text{mín}\{k : D_k(i,j) \neq \emptyset\}, & si\ i \neq j, \\ 0, & si\ i = j. \end{cases} \tag{4.31}$$

Entonces, $G_N(f_i, f_j) = k$ indica que k es el menor número de saltos necesarios para conectar f_i con f_j, con $k-1$ intermediarios. Observemos además que la cadena más larga que puede existir entre dos perfiles pertenecientes a una muestra de tamaño N tendrá $N-1$ enlaces y $N-2$ intermediarios.

Cuando los perfiles f_i y f_j son inalcanzables entre sí, definiremos el grado de separación como $G_N(f_i, f_j) = N$.

El evento $\{G_N(f_i, f_j) \geq k+1\}$ indica que k pasos no son suficientes para conectar f_i con f_j, es decir, en términos del conjunto $D_k(i,j)$ significa que la suma sobre todos los posibles caminos con k enlaces y $k-1$ intermediarios que conectan a i con j es cero, esto es,

$$\{G_N(f_i, f_j) \geq k+1\} = \left\{ \sum_{\vec{i} \in D_k(i,j)} \prod_{l=0}^{k-1} \alpha_\infty(f_{\vec{i}_l}, f_{\vec{i}_{l+1}}) = 0 \right\}. \tag{4.32}$$

Definición 4.6 *Definimos el grado medio de separación entre f_i y f_j como*

$$\tau_{i,j} = \lim_{N \to \infty} E[G_N(f_i, f_j)]. \tag{4.33}$$

Por lo tanto, utilizando la definición anterior, probaremos el siguiente resultado que nos permitirá analizar el nivel de conectividad entre los perfiles mencionados según el contexto comunicacional en que se relacionen y cuando el tamaño de la muestra tiende a infinito, esto es, cuando consideramos a todos los perfiles que integran la red.

Teorema 4.3.1 ***(Grado medio de separación)***

i) Bajo el supuesto de TC, $\tau_{i,j} = 2 - p$.

ii) Bajo Segmentación Fuerte, es decir, cuando hay al menos dos segmentos no conectados entre sí, $\tau_{i,j} = +\infty$, siendo $f_i \in S_i$ y $f_j \in S_j$, S_i y S_j segmentos disjuntos.

Demostración

Sean $f_i, f_j \in \mathcal{F}_\infty$,

i) Supongamos que se cumple la hipótesis de TC en la red. Como el grado de separación es una variable aleatoria con rango en el conjunto $\{1, \ldots, N-1\}$, podemos calcular su esperanza de la siguiente manera:

$$\begin{aligned} E[G_N(f_i, f_j)] &= \sum_{n=1}^{N-1} P(\{G_N(f_i, f_j) \geq n\}) \qquad (4.34) \\ &= P(G_N(f_i, f_j) = 1) + 2\, P(G_N(f_i, f_j) \geq 2) + \sum_{n=3}^{N-1} P(G_N(f_i, f_j) \geq n). \end{aligned}$$

El primer y segundo sumando de (4.34) son fáciles de deducir. En efecto, la probabilidad de que el grado de separación entre dos perfiles sea uno equivale a que dichos perfiles sean amigos, es decir, a que la función aleatoria de amistad $\alpha_\infty(f_i, f_j)$ con distribución Bernoulli de parámetro p tome el valor uno. Por lo tanto,

$$P(G_N(f_i, f_j) = 1) = P(\alpha_\infty(f_i, f_j) = 1) = p.$$

Además, la probabilidad de que el grado de separación entre dos perfiles sea al menos dos equivale a la probabilidad de que dichos perfiles no sean amigos, es decir, a que $\alpha_\infty(f_i, f_j)$ tome el valor cero. Luego,

$$P(G_N(f_i, f_j) \geq 2) = P(\alpha_\infty(f_i, f_j) = 0) = 1 - p.$$

Para el tercer sumando de (4.34), si tomamos $k+1 = n$ en (4.30) y (4.32) tenemos que:

$$D_{n-1}(i,j) = \{\vec{i} = (\vec{i_0}, \vec{i_1}, \ldots, \vec{i_{n-2}}, \vec{i_{n-1}}) : \vec{i_0} = i, \vec{i_{n-1}} = j, \vec{i_l} \neq \vec{i_{l'}} \text{ si } l \neq l'\}$$

es el conjunto de todos los caminos posibles con $n-1$ enlaces y $n-2$ intermediarios que conectan al sitio i con el j y que

$$\{G_N(f_i, f_j) \geq n\} = \left\{ \sum_{\vec{i} \in D_{n-1}(i,j)} \prod_{l=0}^{n-2} \alpha_\infty(f_{\vec{i_l}}, f_{\vec{i}_{l+1}}) = 0 \right\}$$

es el evento que indica que $n-1$ pasos no son suficientes para conectar f_i con f_j.

Sea $H_{N,n}(f_i, f_j) = \sum_{\vec{i} \in D_{n-1}(i,j)} \prod_{l=0}^{n-2} \alpha_\infty(f_{\vec{i_l}}, f_{\vec{i}_{l+1}})$.

Observemos que los sumandos en $H_{N,n}(f_i, f_j)$ son cero o uno, por lo que $H_{N,n}(f_i, f_j)$ es mayor o igual que cero y por lo tanto $P(H_{N,n}(f_i, f_j) = 0) = P(H_{N,n}(f_i, f_j) \leq 0)$. Luego,

$$\begin{aligned} P(G_N(f_i, f_j) \geq n) &= P(H_{N,n}(f_i, f_j) = 0) = P(H_{N,n}(f_i, f_j) \leq 0) \qquad (4.35) \\ &= P\Big(H_{N,n}(f_i, f_j) - E(H_{N,n}(f_i, f_j)) \leq -E(H_{N,n}(f_i, f_j)) \Big) \\ &\leq P\Big(|H_{N,n}(f_i, f_j) - E(H_N(f_i, f_j))| \geq E(H_{N,n}(f_i, f_j)) \Big). \end{aligned}$$

Aplicando la desigualdad de Chebyshev en (4.35) resulta:

$$P(G_N(f_i, f_j) \geq n) \leq \frac{Var(H_{N,n}(f_i, f_j))}{E(H_{N,n}(f_i, f_j))^2}. \qquad (4.36)$$

Resta calcular $E(H_{N,n}(f_i, f_j))$ y $Var(H_{N,n}(f_i, f_j))$ para completar y estudiar el comportamiento de (4.36), cuando $N \to \infty$.

Como el cardinal del conjunto $D_{n-1}(i,j)$, $card\{D_{n-1}(i,j)\}$, es el número de subconjuntos con $n-2$ elementos que se pueden tomar de un conjunto con $N-2$ elementos sin repetición, es decir, $V_{N-2}^{n-2} = \frac{(N-2)!}{[(N-2)-(n-2)]!}$, es claro que

$$\begin{aligned} E(H_{N,n}(f_i, f_j)) &= V_{N-2}^{n-2}\, p^{n-1} \\ &= (N-2)\,[(N-2)-1] \ldots [(N-2)-(n-3)]\, p^{n-1} \\ &\cong N^{n-2}\, p^{n-1}. \qquad (4.37) \end{aligned}$$

Por otro lado, podemos descomponer a la $Var(H_{N,n}(f_i, f_j))$ de la siguiente manera:

$$\begin{aligned} Var(H_{N,n}(f_i,f_j)) &= \sum_{\vec{i}\in D_{n-1}(i,j)} Var\left(\prod_{l=0}^{n-2}\alpha_\infty(f_{\vec{i}_l}, f_{\vec{i}_{l+1}})\right) + \\ &+ \sum_{\vec{i},\vec{i'}\in D_{n-1}(i,j)} Cov\left(\prod_{l=0}^{n-2}\alpha_\infty(f_{\vec{i}_l}, f_{\vec{i}_{l+1}}), \prod_{l=0}^{n-2}\alpha_\infty(f_{\vec{i'}_l}, f_{\vec{i'}_{l+1}})\right), \end{aligned} \tag{4.38}$$

y, como $card\{D_{n-1}(i,j)\} = V_{N-2}^{n-2}$, resulta

$$\begin{aligned} \sum_{\vec{i}\in D_{n-1}(i,j)} Var\left(\prod_{l=0}^{n-2}\alpha_\infty(f_{\vec{i}_l}, f_{\vec{i}_{l+1}})\right) &= V_{N-2}^{n-2}\, p^{n-1}\,(1-p^{n-1}) \\ &\cong N^{n-2}\, p^{n-1}\,(1-p^{n-1}). \end{aligned} \tag{4.39}$$

Para el segundo sumando de (4.38), notando $\alpha_\infty(f_{\vec{i}_l}, f_{\vec{i}_{l+1}}) := \alpha_{\vec{i}_l\vec{i}_{l+1}}$ tenemos

$$\begin{aligned} &\sum_{\vec{i},\vec{i'}\in D_{n-1}(i,j)} Cov\left(\prod_{l=0}^{n-2}\alpha_\infty(f_{\vec{i}_l}, f_{\vec{i}_{l+1}}), \prod_{l=0}^{n-2}\alpha_\infty(f_{\vec{i'}_l}, f_{\vec{i'}_{l+1}})\right) = \\ &= \sum_{\vec{i},\vec{i'}\in D_{n-1}(i,j)} \left[E\left(\prod_{l=0}^{n-2}\alpha_{\vec{i}_l\vec{i}_{l+1}}\ \prod_{l=0}^{n-2}\alpha_{\vec{i'}_l\vec{i'}_{l+1}}\right) - E\left(\prod_{l=0}^{n-2}\alpha_{\vec{i}_l\vec{i}_{l+1}}\right)\ E\left(\prod_{l=0}^{n-2}\alpha_{\vec{i'}_l\vec{i'}_{l+1}}\right)\right]. \end{aligned}$$

Dejando fijo $\vec{i}$, contabilizamos el número de caminos $\vec{i'}$ de longitud $n-1$ que multiplican al primer camino $\vec{i}$ que no tengan enlaces en común con con $\vec{i}$.

Dichos caminos no aportan a las covarianzas y el número de tales caminos es

$$(N-3)\,(N-4)\ldots(N-n).$$

En consecuencia, el aporte a la suma de las covarianzas vendrá dado por aquellos caminos $\vec{i'}$ que tienen al menos un enlace en común con $\vec{i}$. Podemos contabilizar esta cantidad por diferencia entre el número total de caminos $\vec{i'}$ y el número de caminos $\vec{i'}$ que no aportan a las covarianzas. Es decir,

$$(N-2)\,[(N-2)-1]\ldots[(N-2)-(n-3)] \;-\; (N-3)\,(N-4)\ldots(N-n).$$

Sea la función

$$\psi(x) = (N-x)\,[(N-x)-1]\ \ldots\ [(N-x)-(n-3)], \tag{4.40}$$

con $x > 0$. ψ es continua en el intervalo [2,3] y derivable en el intervalo (2,3), entonces por el Teorema del Valor Medio, existe $z \in (2,3)$ tal que $\psi'(z) = \psi(3) - \psi(2)$. Luego, para $z \in (2,3)$,

$$\begin{aligned} -\psi'(z) &= \psi(2) - \psi(3) \\ &= (N-2)\,(N-3)\ldots[N-(n-1)] \;-\; (N-3)\,(N-4)\ldots(N-n). \end{aligned} \tag{4.41}$$

Por otro lado, calculando la derivada primera de la función ψ en (4.40) y multiplicándola por menos uno tenemos

$$\begin{aligned} -\psi'(x) &= [(N-x)-1]\,[(N-x)-2]\ldots[(N-x)-(n-3)]+ \\ &+ (N-x)\,[(N-x)-2]\ldots[(N-x)-(n-3)]+ \\ &+ \cdots+(N-x)\,[(N-x)-1]\ldots[(N-x)-(n-4)], \end{aligned}$$

por lo que acotando cada uno de los $n-2$ términos de esta suma por $(N-2)^{n-3}$ resulta $-\psi'(x) \leq (N-2)^{n-3}(n-2)$. En particular, para $z \in (2,3)$,

$$-\psi'(z) \leq (n-2)(N-2)^{n-3}. \tag{4.42}$$

De (4.41) y (4.42) podemos concluir que

$$(N-2)(N-3)\ldots[N-(n-1)] \;-\; (N-3)\ldots(N-n) \leq (n-2)(N-2)^{n-3},$$

es decir, el número de caminos $\vec{i'} \in D_{n-1}(i,j)$ que tienen al menos un enlace en común con $\vec{i} \in D_{n-1}(i,j)$ y por lo tanto aportan a la suma de las covarianzas es a lo sumo $(N-2)^{n-3}(n-2)$.

Entonces,

$$\begin{aligned} \sum_{\vec{i},\vec{i'}\in D_{n-1}(i,j)} Cov\left(\prod_{l=0}^{n-2}\alpha_\infty(f_{\vec{i}_l},f_{\vec{i}_{l+1}}),\prod_{l=0}^{n-2}\alpha_\infty(f_{\vec{i'}_l},f_{\vec{i'}_{l+1}})\right) &\leq (n-2)(N-2)^{n-3}\times \\ &\times\; p^{n-1}(1-p^{n-1}). \end{aligned} \tag{4.43}$$

Retomando el cálculo de $Var(H_{N,n}(f_i,f_j)$ en (4.38), por (4.39) y (4.43) resulta

$$\begin{aligned} Var(H_{N,n}(f_i,f_j)) &\leq N^{n-2}\,p^{n-1}(1-p^{n-1})+(n-2)(N-2)^{n-3}p^{n-1}(1-p^{n-1}) \\ &\cong N^{n-2}\,p^{n-1}(1-p^{n-1}). \end{aligned} \tag{4.44}$$

Volviendo a la desigualdad (4.36), con los resultados para $E(H_{N,n}(f_i,f_j))$ y $Var(H_{N,n}(f_i,f_j))$ dados por (4.37) y (4.44), obtenemos que

$$\begin{aligned} P(G_N(f_i,f_j)\geq n) &\leq \frac{Var(H_{N,n}(f_i,f_j))}{E(H_{N,n}(f_i,f_j))^2} \\ &\leq \frac{N^{n-2}p^{n-1}(1-p^{n-1})}{N^{2(n-2)}p^{2(n-1)}} \\ &= \frac{1}{N^{n-2}p^{n-1}}-\frac{1}{N^{n-2}}. \end{aligned} \tag{4.45}$$

Luego,

$$\begin{aligned}\sum_{n=3}^{N-1} P(G_N(f_i,f_j)\geq n) &\leq \sum_{n=3}^{N-1}\frac{1}{N^{n-2}p^{n-1}} - \sum_{n=3}^{N-1}\frac{1}{N^{n-2}} \\ &= \sum_{r=0}^{N-4}\frac{1}{N^{r+1}\,p^{r+2}} - \sum_{r=0}^{N-4}\frac{1}{N^{r+1}} \\ &= \frac{1}{N}\left[\frac{1}{p^2}\sum_{r=0}^{N-4}\left(\frac{1}{Np}\right)^r - \sum_{r=0}^{N-4}\left(\frac{1}{N}\right)^r\right]. \qquad (4.46)\end{aligned}$$

Como las series $\sum_{r=0}^{N-4}\left(\frac{1}{Np}\right)^r$ y $\sum_{r=0}^{N-4}\left(\frac{1}{N}\right)^r$ son convergentes y $\frac{1}{p^2}\sum_{r=0}^{N-4}\left(\frac{1}{Np}\right)^r - \sum_{r=0}^{N-4}\left(\frac{1}{N}\right)^r$ tiende a $\frac{1}{p^2}-1$, cuando $N\to\infty$, (4.46) tiende a cero cuando $N\to\infty$.

Entonces tomando límite para $N\to\infty$ en (4.34) resulta

$$\tau_{i,j} = \lim_{N\to\infty} E(G_N(f_i,f_j)) = p + 2(1-p) = 2-p.$$

ii) Estudiaremos ahora el caso de *Segmentación Fuerte*. Este contexto comunicacional se caracteriza por la presencia de al menos dos segmentos no conectados entre sí en la red.

Sean $S_1, S_2 \subseteq \mathcal{F}_\infty$ dos segmentos disjuntos tales que $S_1 \cup S_2 = \mathcal{F}_\infty$ y sean $a = \frac{card\{S_1\}}{card\{\mathcal{F}_\infty\}}$ y $1-a = \frac{card\{S_2\}}{card\{\mathcal{F}_\infty\}}$. Suponemos que dentro de cada segmento se verifica la hipótesis de TC pero entre ellos no existe interacción alguna y tomamos un muestreo estratificado al azar por segmento con $S_1 = \{f_1,\ldots,f_{[aN]}\}$ y $S_2 = \{f_{[aN]+1},\ldots,f_N\}$. Luego, si $f_i \in S_1$ y $f_j \in S_2$ ó $f_i \in S_2$ y $f_j \in S_1$, los perfiles son inalcanzables entre sí y, por lo tanto, de acuerdo a cómo definimos el grado de separación entre dos perfiles, resulta que $G_N(f_i,f_j) = N$. En consecuencia,

$$\tau_{i,j} = \lim_{N\to\infty} E(G_N(f_i,f_j)) = +\infty.$$

□

Los resultados anteriores nos indican que cuando estamos bajo la hipótesis de TC, donde la probabilidad de que cualquier par de perfiles sean amigos en Facebook es la misma, digamos p, hallamos que a medida que el tamaño de la muestra crece hasta considerar la red completa, el grado medio de separación entre dos perfiles cualesquiera depende de p, más precisamente es $2-p$. Por consiguiente, para el par de perfiles $f_i, f_j \in \mathcal{F}_\infty$, cuanto mayor probabilidad haya de que éstos sean amigos más cercano a uno será el grado medio de separación entre ellos reflejando que con sólo un enlace pueden conectarse. Por otro lado, cuanto menos probable sea la amistad entre los perfiles, son suficientes dos enlaces y

un intermediario para conectarlos. En conclusión, bajo TC hemos verificado que es viable la Teoría de Mundo Pequeño.

En el caso de existir en la red al menos dos segmentos entre los que no haya ninguna interacción, tomando un perfil de cada segmento, el grado medio de separación entre ellos tenderá a infinito cuando el tamaño de la muestra crezca. Por lo tanto, en el contexto comunicacional de Segmentación Fuerte queda invalidada la Teoría de Mundo Pequeño, ya que hay perfiles que son inalcanzables entre sí.

4.3.1. Segmentación débil

En capítulos anteriores presentamos un modelo para estudiar la dinámica a largo plazo de la red social Facebook basado en el hecho de que los "clicks", es decir, las funciones de amistad entre pares de perfiles, α_∞, son variables aleatorias indepedientes con distribución Bernoulli, con parámetro constante (caso de TC) o parámetro dependiendo del esquema de segmentación. Otra característica del modelo es la no restricción en el número de amigos que puede tener un perfil.

En la realidad, podemos observar que los perfiles se conectan con perfiles de características similares e intereses comunes de diversa índole, agrupándose en segmentos y que además existe interacción entre los segmentos. Llamaremos a este contexto comunicacional, el contexto de *Segmentación Débil*.

En este contexto podemos intuir que, dados S_i y S_j dos segmentos de perfiles de $\mathcal{F}_\infty$, $S_i \cap S_j = \emptyset$, con $f_i, f_j \in \mathcal{F}_\infty$ y tales que $f_i \in S_i$, $f_j \in S_j$, el grado medio de separación entre f_i y f_j, $\tau_{i,j} = \lim_{N\to\infty} E(G_N(f_i, f_j))$, podría ser finito o eventualmente infinito, dependiendo de la intensidad con que se relacionen los segmentos a los que pertenecen f_i y f_j.

4.4. Conclusiones y trabajo futuro

Si estudiamos el "fenómeno de mundo pequeño" en cualquier red social que no esté muy agrupada y en la que los individuos tengan relaciones fuera de su círculo íntimo, es probable que en dicha red sea viable la Teoría de Mundo Pequeño y esto dependerá de la fuerza con que se relacionen los perfiles de segmentos distintos.

En este Capítulo, hemos hecho referencia a los trabajos de Watts D. y Strogatz S. [46] y Watts D. [47]. En ellos se muestran dos situaciones extremas:

a) Un mundo social caracterizado por un agrupamiento extremo, en el que las personas sólo se relacionan con aquellas con las que las une un vínculo familiar o de amistad muy cercano. En este contexto, la distancia entre dos individuos del planeta elegidos al azar será muy grande. Por ejemplo, el caso de un individuo que desarrolla toda su vida en una tribu aislada. La densidad relacional que tendrá con los miembros de su tribu, será muy

alta. Sin embargo, le serán inaccesibles otras personas que habiten en otras tribus a miles de kilómetros de la suya.

b) En el otro extremo, en una red social en la que todos sus individuos establecen relaciones aleatorias fuera de su círculo íntimo la distancia media entre ellos será pequeña. Por ejemplo, si un individuo tiene 50 contactos y sus contactos tienen a su vez 50 contactos en promedio, en sólo dos pasos el individuo podría contactarse con 2.500 personas, en tres pasos podría llegar a los 125.000 contactos, en cuatro a 6.250.000 y en cinco a 312.500.000. Si estos contactos se establecieran en forma aleatoria, por ejemplo como cuando entablamos una relación con un desconocido que comparte un viaje con nosotros al ir a trabajar en ómnibus, en muy pocos pasos una persona podría alcanzar a cualquier otra del planeta ya que su información no se agotaría a su círculo íntimo como sucede en las redes muy aglomeradas.

En nuestro análisis, al contexto comunicacional descripto por la situación a) lo hemos denominado el contexto de "Segmentación Fuerte" y pudimos comprobar que dados dos perfiles cualesquiera de distintos segmentos, el grado medio de separación entre ellos, $\tau_{i,j}$ resultó infinito, lo cual hace inviable la Teoría de Mundo Pequeño.

Por otro lado, observamos que el contexto de la situación extrema desarrollada en b), se asemeja al contexto de TC en el que todo induviduo se relaciona con los demás en forma aleatoria y con la misma probabilidad.

Sin embargo, el mundo real es evidente que propone un escenario intermedio en donde los individuos se relacionan con aquellos que presentan características similares, es decir, se agrupan en segmentos, pero que simultáneamente existe interacción entre los segmentos. A este contexto lo hemos denominado el contexto de "Segmentación Débil". En este marco, todo parece indicar que dados dos perfiles de dos segmentos distintos cualesquiera del planeta, el grado medio de separación entre ellos dependerá de la intensidad con que se relacionen los distintos segmentos. La propuesta a futuro es abordar en profundidad este caso a través de la modelización matemática y probabilística aplicada en particular a la red social Facebook, y el análisis del fenómeno de mundo pequeño mediante el cálculo formal del grado medio de separación entre perfiles. Sumamos a esta propuesta el desafío del modelado y estudio de la red restringiendo la cantidad de amigos que pueda tener cada perfil de la misma.

El hecho de Facebook parezca mostrar una distancia corta entre individuos como lo indica el estudio realizado por Ugander J., Karrer B., Backstrom L. y Marlow C. [44], podría deberse a varios factores, como por ejemplo que en general los usuarios tienen un gran número de contactos superficiales con los que ni siquiera mantiene una conversación o nunca vio en persona, aunque aún así, en muchos casos Facebook ha facilitado el encuentro de personas que de otro modo no se hubiera llevado a cabo y mantener contactos. Otro factor que podemos destacar es que no todo el mundo es usuario en Facebook, por ejemplo, un anciano de un pueblo del norte del país que no tiene acceso a internet es prácticamente

improbable que tenga una cuenta de Facebook, aunque si este anciano tuviera un nieto en la capital conectado a 100 individuos, éste podría conectarlo con sólo un paso a la red social.

De todos modos, podemos concluir que en Facebook la teoría de que el mundo es pequeño es de considerable viabilidad. El mundo está cada vez más conectado, debido en gran medida a la tecnología de la información. Las personas se relacionan de modo virtual y real de manera progresiva con personas fuera de su círculo familiar o íntimo, ya sea por cuestiones laborales, de estudio, de recreación, políticas, económicas, compra de artículos, etc. Además, las ciudades han crecido y el mundo esta cada vez más diversificado y urbanizado. La homofilia, es decir, la tendencia de la gente a relacionarse con personas de características similares a las suyas, persistirá siempre ya que está en la naturaleza humana. Sin embargo, por todo lo expuesto, todo parece indicar que el mundo es cada vez más pequeño.

Bibliografía

[1] Airoldi E., Blei D.M., Fienberg S.E., Goldenberg A., Xing E.P. and Zheng A.X. (Eds). (2007). *Statistical Networks Analysis: Models , Issues, and New Directions.* Springer.

[2] Arnold L. (1967). *On the Asymptotic Distribution of the Eigen Values of Random Matrices.* Journal Mathematical Analysis and Applications, Vol. 20, pp. 262-268.

[3] Bandyopadhyay S., Rao A.R. and Sinha B.K. (Eds). (2010). *Models for Social Networks with Statistical Applications.* SAGE.

[4] Bian Y. J. (1997). *Bringing Strong Ties Back in: Indirect Ties, Network Bridges and Job Searches in China.* American Sociological Review, Vol. 62, No. 3, pp. 366-385.

[5] Billignsley P. (1968). *Convergence of Probability Measures.* New York. Wiley and Sons.

[6] Billignsley P. (1979). *Probability and Measures.* New York. Wiley and Sons.

[7] Bolch G., Greiner S., De Meer H. and Trivedi K. (2006). *Queueing Networks and Markov Chains: Modeling and Performance Evaluation With Computer Science Applications.* New Jersey. Wiley and Sons.

[8] Bondy J.A. and Murty U.S.R. (2008). *Graph Theory.* Springer. Graduate Texts in Mathematics, Vol. 224.

[9] Boorman S. and White H. (1976). *Social Structure from Multiple Networks. II. Role Structures.* American Journal of Sociology, Vol. 81, No. 6, pp. 1384-1446.

[10] Breiger R., Carley K. and Pattison P. (Eds). (2003). *Dynamic Social Network Modelling and Analysis: Workshop Summary and Papers.* National Research Council.

[11] Burt R. (1995). *Structural Holes: The Social Structure of Competition.* Harvard University Press.

[12] Carrero F. (2008). http://franciscocarrero.com/2008/06/13/cual-es-el-techo-de-las-redes-sociales-2.

[13] Carrington P., Scott J. and Wasserman S. (2005). *Models and Methods in Social Network Analysis.* Cambridge University Press.

[14] Closa G. (2011). www.elogia.net/blog/autor/ciclo-vital-facebookiano.

[15] DasGupta A. (2008). *Asymptotic Theory of Statistics and Probability.* Springer, New York.

[16] De Ugarte D. (2011). *El poder de las redes.* Manual ilustrado para ciberactivistas. Colección Biblioteca de las Indias Electrónicas. http://lasindias.org.

[17] Diestel R. (2000). *Graph Theory.* Springer.

[18] Dominguez Molina J. y Rocha Arteaga A. (2009). *El Teorema de Wigner para matrices aleatorias.* Comunicación del CIMAT. No. I-09-08/15-10-2009 (PE/CIMAT).

[19] Feller W. (1968). *An Introduction to Probability Theory and its Applications,* Vol. 1. John Wiley, New York, 3rd edition, 15, 51, 63.

[20] Freeman L.C. (2006). *The Development of Social Network Analysis: A Study in The Sociology of Science.* Vancouver: Empirical Press.

[21] Freeman L.C. (2008). *What is Social Network Analysis?* Last Update Friday, Available at: http://www.insna.org/sna/what.html.

[22] Furht B. (Ed). (2010). *Handbook of Social Network Technologies and Applications.* Springer.

[23] Girko V. L. (1988). *Spectral Theory of random matrices.* Nauka. Moscow.

[24] Granovetter M. (1973). *The Strength of Weak Ties.* American Journal of Sociology, Vol. 78, pp. 1360-1380.

[25] Guare J. (1990). *Six Degrees of Separation: A Play.* (First edition ed.). New York: Random House.

[26] Hawe P., Webster C. and Shiell A. (2004). *A Glossary of Terms for Navigating the Field of Social Network Analysis.* Journal of Epidemiology and Community Health, Vol. 58, pp. 971-975.

[27] Hoel P.G., Port S.C. and Stone C.J. (1972). *Introduction to Stochastic Processes.* Houghtoun Mifflin.

[28] Jones G. (2004). *On the Markov Chain Central Limit Theorem.* Prob. Surveys, 1, 299-320.

[29] Koshy T. (2009). *Catalan Numbers with applications.* Oxford University Press, Inc.

[30] Lee A. J. (1990). *U-Statistics: Theory and Practice.* Marcel Dekker, New York.

[31] Lehmann E.L. (1999). *Elements of Large Sample Theory.* Springer, New York.

[32] Lin N. (2002). *Social Capital: A theory of Social Structure and Action.* Cambridge University Press, First Edition.

[33] Marin A. and Wellman B. (2010). *Social Network Analysis: An Introduction.* Carrington P. and Scott J. (Eds). Handbook of Social Network Analysis. London: SAGE.

[34] Meyn S.P. and Tweedie R.L. (1993). *Markov Chains and Stochastic Stability.* Springer, New York.

[35] Mickenberg R. and Dugan J. (1995). *Taxi Driver Wisdom.* San Francisco: Chronicle.

[36] Milgram S. (1967). *The Small-World Problem.* Psychology Today, 1 (1), 61-67.

[37] Plickert G., Coté R. and Wellman B. (2007). *It's Not Who You Know. It's How You Know Them: Who Exchanges What with Whom?* Social Networks, Vol. 29, No. 3, pp. 405-429.

[38] Polanco X. (2006). *Análisis de Redes: Introducción.* Author manuscript, published in "Redes de conocimiento: Construcción, dinámica y gestión", Albornoz M. y Alfaraz C. (Ed). 77-112.

[39] Pool I. and Kochen M. (1978). *Contacts and Influence.* Social Networks, 1, pp.5-51. Elsevie Sequoia S.A., Lausanne.

[40] Rincón L. (2012). *Introducción a los Procesos Estocásticos.* Departamento de Matemáticas, Facultad de Ciencias UNAM, Circuito Exterior de CU, México DF.

[41] Scott J. (1991). *Social Network Analysis.* London: SAGE.

[42] Serfling R. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley and Sons. New York. Chichester. Brisbane. Toronto.

[43] Sinai Y. and Soshnikov A. (1998). *Central Limit Theorem for Traces of Large Random Symmetric Matrices with Independent Matrix Elements.* Bulletin of Brazilian Mathematical Society. Vol. 29, No. 1, pp. 1-24.

[44] Ugander J., Karrer B., Backstrom L. and Marlow C. (2011). *The Anatomy of Facebook Social Graph.* http://arxiv.org/abs/1111.4503.

[45] Wasserman S. and Faust K. (1994). *Social Networks Analysis: Methods and Applications.* Cambridge: Cambridge University Press.

[46] Watts D. and Strogatz S. (1998). *Collective Dynamics of Small-World Networks.* Nature, Vol. 393, No. 6684, pp. 440-442. Nature Publishing Group.

[47] Watts D. (2003). *Six Degrees: The Science of a Connected Age.* W. W. Norton and Company.

[48] Wellman B. and Berkowitz S.D. (Eds). (1988). *Social Structures: A Network Approach.* Cambridge: Cambridge University Press.

[49] Wetherell C., Plakans A. and Wellman B. (1994). *Social Networks, Kinship, and Community in Eastern Europe.* Journal of Interdisciplinary History, Vol. 24, No. 1, pp. 639-663.

[50] White H.D., Wellman B. and Nazer N. (2004). *Does Citation Reflect Social Structure? Longitudinal Evidence from the 'Globenet' Interdisciplinary Research Group.* Journal of the American Society for Information Science and Tecnology, Vol. 55, No. 2, pp. 111-126.

[51] Wigner E. P. (1955). *Characteristic Vectors of Bordered Matrices With Infinte Dimensions.* Annals of Mathematics, Vol. 62, 548-564.

[52] Wigner E. P. (1958). *On the Distributions of the Roots of Certain Symmetric Matrices.* Annals of Mathematics, Vol. 67, No. 2, pp. 325-327.

[53] Zhu J. (2007). *Opportunities and Challenges for Network Analysis of Social and Behavioral Data.* Seminar Series on Chaos, Control and Complex Networks City University of Hong Kong, Poly U University of Hong Kong and IEEE Hong Kong R and A/CS Joint Chapter.