Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Efectos de la topología de redes complejas en el aprendizaje social

Tesis presentada para optar al título de
Licenciado en Ciencias de la Computación

Gustavo Juan Cairo

Director: Esteban Eduardo Mocskos

Codirector: Gustavo Andrés Landfried

Buenos Aires, Julio 2021

Resumen

La experiencia individual es un factor fundamental para el aprendizaje. Sin embargo, el aprendizaje humano es también un fenómeno poblacional que depende de las interacciones entre los individuos: la cultura es información que una comunidad de agentes desarrolla y acumula en el tiempo. Una de las preguntas abiertas intenta develar la influencia que la estructura de la red de interacciones entre agentes tiene sobre el flujo de la información cultural y el aprendizaje de las personas. Ninguna de las metodologías utilizadas hasta ahora (desde los experimentos de laboratorio, los trabajos de campo antropológicos, o las simulaciones numéricas) analizan la evolución de la estructura de la red en el tiempo y su relación con el aprendizaje. En este contexto nos preguntamos, ¿podemos detectar la influencia que la posición de un individuo en la red tiene sobre su aprendizaje?

Para avanzar en la resolución de esta pregunta tomamos como caso de estudio una comunidad de jugadores en línea del juego de mesa Go. En primer lugar, estudiamos las curvas de aprendizaje individuales utilizando el estimador de habilidad estado del arte en la industria de los videojuegos. Luego, creamos la red de interacciones en base a las partidas (una arista entre dos jugadores cuando se enfrentan en una partida) y analizamos la evolución de las centralidades individuales usando ventanas temporales deslizantes. El objetivo es explorar el impacto de las medidas de centralidad sobre el aprendizaje que no puede ser explicado por simple experiencia individual.

Nuestra hipótesis inicial esperaba observar una relación positiva entre la conectividad y el aprendizaje. Lo que encontramos fue evidencia en el sentido contrario: una centralidad relativamente baja está asociada a mejores curvas de aprendizaje. Este resultado, sin embargo, está en línea con las conclusiones a la que llegaron los primeros trabajos que abordan el efecto de la estructura sobre la acumulación cultural. En un extremo, redes demasiado interconectadas pueden ser contraproducentes para el aprendizaje, por ejemplo debido a la sobre-información. En el otro extremo, niveles muy bajos de conectividad incrementan las chances de pérdida cultural. Son las redes de conectividad moderada las que exhiben mejores procesos de aprendizaje.

Keywords: Habilidad, Go, Topología, Centralidad

Índice general

1. Introducción **5**

1.1. Objetivo . . . 6

1.2. Antecedentes . . . 6

1.3. Preliminares . . . 8

1.3.1. ¿Qué es el Go y cuáles son sus reglas? . . . 8

1.3.2. Sobre OGS, el sistema de partidas online . . . 9

1.3.3. Curvas de aprendizaje poblacional . . . 9

2. Modelos de estimación de habilidad **11**

2.1. Medida *Elo* . . . 11

2.2. Medida *Glicko* (y otras medidas basadas en Inferencia Bayesiana) . . . 12

2.3. Medida *TrueSkill* . . . 13

2.4. Medida *TrueSkill Through Time (TTT)* . . . 15

3. Topología y análisis de su evolución **17**

3.1. Grafos dinámicos y su análisis . . . 17

3.1.1. Medidas que consideran el tiempo . . . 18

3.1.2. Análisis estático con ventana deslizante o *snapshots* . . . 18

3.2. Medidas de centralidad . . . 19

3.2.1. Degree Centrality . . . 19

3.2.2. Closeness . . . 19

3.2.3. Harmonic Centrality . . . 21

3.2.4. Betweenness . . . 22

3.2.5. Communicability Betweenness . . . 24

3.2.6. Load Centrality . . . 25

3.2.7. Current-Flow Betweenness . . . 27

3.2.8. Information Centrality . . . 30

3.2.9. Eigenvector Centrality . . . 33

4. Datos **35**

4.1. Selección del tipo de base de datos . . . 35

4.2. Scraping inicial del sitio web OGS . . . 35

4.3. Creación de la base de datos 36
4.3.1. Esquema de la base de datos 36
4.4. Limpieza y descripción del dataset 37
4.4.1. Bots, el primer imprevisto 37
4.4.2. Fusión con otro sitio web, el segundo imprevisto 37
4.4.3. Estimaciones erróneas, el tercer imprevisto 38
4.5. Procesamiento de los datos 38

5. Resultados **40**
5.1. Propiedades generales de la red 40
5.2. Curvas de aprendizaje poblacional 44
5.3. Relación entre el aprendizaje y la topología 46

6. Conclusiones y trabajo futuro **53**

Capítulo 1

Introducción

Los humanos pertenecemos a una especie bastante especial. Hemos llegado a todos los rincones de nuestros planeta partiendo desde África hace apenas cien mil años. Entre nuestros ancestros y nosotros no hay casi ninguna diferencia a nivel biológico. Pero entonces, ¿qué es lo que cambió en este tiempo para que hoy estemos rodeados de toda esta tecnología, en un estado de modernización sin precedentes?

A diferencia de otros animales, nosotros nacemos sin demasiadas capacidades innatas. Sin embargo tenemos una capacidad extraordinaria para imitar que nos permite aprender eficientemente de otros. Desde que nacemos, aprendemos copiando a los que nos rodean: la forma en que nos saludamos, la comida que cocinamos, la manera de vestirnos... Constantemente estamos tomando de los demás lo que parece funcionar mejor, lo que nos parece más interesante, o simplemente lo que hace la mayoría de la gente o nuestro círculo cercano.

Lo que cambió radicalmente en estos más de cien mil años no fue nuestra información genética, sino nuestra información cultural. Cada uno de nuestros antepasados imitó, modificó y transmitió parte de su conocimiento a la siguiente generación, llegando a nosotros sedimentado en forma de herramientas, creencias y rutinas. El conocimiento acumulado a través de las generaciones, lo que llamamos **cultura**, es lo que nos hace especiales como animales: esa capacidad que tenemos de aprender de nuestros pares y antepasados, de generar nuevo conocimiento construyendo juntos sobre esas bases aprendidas, y de transmitirlo a nuestros pares y descendientes, en un ciclo que se repite y potencia. Gracias a la imitación, las sociedades se comportan como *sistemas distribuidos de procesamiento de información.* Lo que nos hace especialmente "inteligentes" no radica sólo en nuestras capacidades cognitivas, sino en la información cultural heredada. *La inteligencia humana emerge como una propiedad poblacional, no individual.*

Entendemos a la habilidad de una persona como la probabilidad de llevar a cabo correctamente una determinada actividad, trabajo u oficio. La habilidad generalmente aumenta con la experiencia: cuando una persona realiza una misma tarea muchas veces, mejor se vuelve en ella. Siguiendo la sabiduría popular, *"la práctica hace al maestro"*. Pero acceder a la información cultural es lo que nos permite a los humanos desarrollar habilidades complejas que jamás podríamos descubrir por simple experiencia individual. Tener bien caracterizada la evolución de la habilidad en la realización de una tarea nos permitirá estudiar factores que puedan impactar en algunas de las dimensiones del desarrollo de una persona. Mientras más precisa sea esta medición, más útil será la información que se pueda extraer.

Entonces, ¿cómo medir numéricamente la habilidad? ¿Es posible describirla en forma cuantitativa? A lo largo de los años se crearon diferentes métodos aplicados a diferentes actividades [1–3]. En general, todos los modelos consideran que un jugador tiene un *desempeño* al momento de ejecutar la tarea, es decir, qué tan bien una persona la ejecutó en una instancia particular. El rendimiento está relacionado con su habilidad para realizar la tarea, pero afectada por diversos factores (por ejemplo, el estado de ánimo particular de ese día). Todos los modelos consideran también la cantidad de instancias o veces que el agente realizó la tarea (i.e. su experiencia). Además, en el caso de que la habilidad esté relacionada a un juego, mientras más información tengamos de un jugador (en este caso, mayor número de partidas),

entonces mayor certeza tendremos acerca de su habilidad.

Gracias a la revolución de internet y a la creciente facilidad de acceso a datos, la cantidad de información disponible está dejando de ser un problema en muchos dominios. En particular, los juegos en línea resultan una de las más interesantes fuentes de información en este aspecto, ya que para algunos de ellos tendremos enormes bases de datos donde la cantidad de información será la menor de las preocupaciones. En esta tesis nos enfocaremos particularmente en el análisis del aprendizaje o evolución de la habilidad de un juego de mesa particular: el Go.

1.1. Objetivo

Esta tesis de grado tiene como objetivo principal estudiar los posibles efectos que la estructura y propiedades de la red de interacción entre agentes tienen sobre el aprendizaje individual.

Si el aprendizaje depende de la información heredada o imitada, entonces la topología de la red será un factor fundamental para el aprendizaje, debido a que afecta cómo se distribuye la cultura entre los agentes. La interacción social puede ser representada como una red, en la cual los distintos individuos se interconectan con otros cada vez que intercambian, heredan información cultural, o aprenden del otro mediante imitación.

Como objeto de estudio, usaremos una red de jugadores de Go extraída de un sitio web que permite jugar partidas online. La resolución de este problema involucra trabajar con un gran volumen de datos, verificar su correcto procesamiento, realizar el proceso de curado y, principalmente, **hacer un análisis temporal sobre la red** para observar la evolución del aprendizaje de los jugadores.

1.2. Antecedentes

La especie humana posee la peculiar capacidad de desarrollar innovaciones culturales a lo largo de múltiples generaciones, un fenómeno denominado *evolución cultural acumulada* [4] (o **CCE** por sus siglas en inglés). El modelo causal básico de la literatura de evolución cultural afirma que la adaptación humana depende, además de la experiencia individual y los cambios en el ambiente, de la información cultural a la que tienen acceso los agentes.

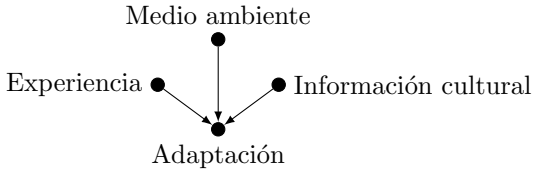


Figura 1.1: Representación gráfica de un modelo causal para la adaptación humana.

Todas las tecnologías, sean éstas viejas o nuevas, surgen como resultado de siglos de esfuerzo por parte de grupos de individuos. Basada en los principios básicos de la teoría evolutiva (descendencia, modificación y selección), la antropología ha desarrollado un marco teórico para explicar cómo cambia la cultura en el tiempo. Este marco ha motivado el desarrollo de los primeros modelos matemáticos para explicar distintas dinámicas poblacionales de producción, transmisión y mantenimiento de información cultural [5, 6].

Una idea central en estos trabajos es que el aprendizaje social (por imitación, observación, etc.) no es al azar, sino que los individuos eligen de quién aprender. Por ejemplo, una estrategia simple consiste en copiar lo que hace la mayoría (*frequency-based strategy*). Otra estrategia típica toma como modelo de referencia a las personas más exitosas o prestigiosas (*payoff-based strategy*). Las diferentes estrategias

de aprendizaje social que se han propuesto [7] pueden ser clasificadas entre aquellas que identifican de quiénes aprenden los individuos, y aquellas que especifican las circunstancias bajo las cuales los individuos copian a otros. Por ejemplo, mantener compañeros de equipo estable está relacionado con un mayor aprendizaje [8].

Uno de los problemas actuales es determinar las condiciones que la *demografía* impone sobre los procesos de acumulación cultural: en particular, el tamaño y estructura (o conectividad) de la población.

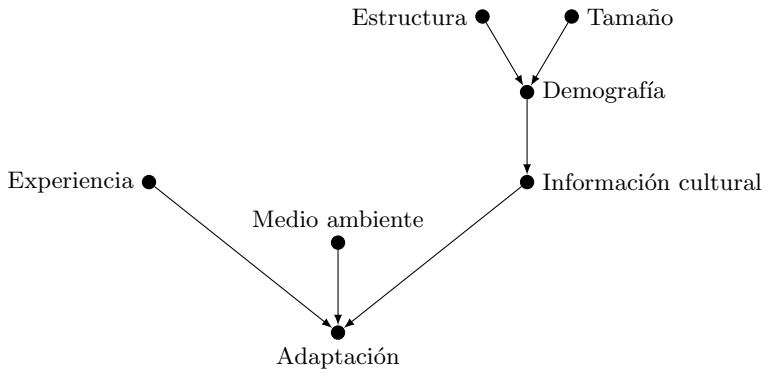


Figura 1.2: Modelo causal que incuye el efecto de la demografía sobre la información cultural

No parece haber dudas de que el tamaño de las poblaciones naturales afecta la información cultural que estará disponible a las subsiguientes generaciones. Dado que la acumulación cultural es un fenómeno que ocurre cuando existe transmisión de información inter-generacional [9, 10], cuanto mayor sea la población, menor será el riesgo de que esa información no pueda ser transmitida a lo largo del tiempo.

Una de las primeras hipótesis de las teorías que incluyen a la demografía en sus modelos [11, 12], es que el *tamaño efectivo de la población* (que dependerá no solo del tamaño de la misma sino también de su conectividad) es capaz de compensar el riesgo de perder información cultural. Por el contrario, una caída muy grande en el tamaño efectivo de la población (por ejemplo por aislamiento, plagas o guerras) puede causar pérdida másiva de información culturales. El ejemplo paradigmático de ésto último es la "involución" cultural que sufre la población de la isla de Tasmania cuando queda separada del continente australiano por la subida de las aguas de los océanos [12].

Las metodologías desarrolladas para analizar este problema van desde los experimentos y trabajos de campo, hasta las simulaciones numéricas. En los experimentos, en general, se les pide a grupos de distintos tamaños que trabajen para mejorar alguna pieza tecnológica [13–16]. Sus limitaciones se relacionan con el reducido tamaño y la estructura simple de las "poblaciones" creadas artificialmente en comparación a la observada en sociedades reales. Los trabajos de campo, por su parte, se enfocan en detectar correlaciones entre la cantidad de herramientas creadas y el tamaño de la población que las creó, utilizando datos etimológicos y arqueológicos. Las limitaciones de los estudios arqueológicos y entográficos tienen que ver con la imposibilidad o dificultad para medir en detalle la estructura de la red de interacciones y analizarla temporalmente. La tercera alternativa, consiste en estudiar las propiedades de los sistemas culturales mediante simulaciones numéricas [17]. Sus limitaciones tienen que ver con que los modelos pueden no ser representativos del problema real o tener suposiciones implícitas o explícitas que impacten o introduzcan sesgos en los resultados.

Una línea de investigación activa en la actualidad intenta develar de qué forma el flujo de información dentro de las poblaciones puede promover o dificultar la acumulación gradual de innovaciones culturales. **El desafío actual es estudiar los efectos de la estructura de las poblaciones sobre la información cultural o la adaptabilidad [4].** Las causas estructurales parecen ser el principal factor para entender los procesos de evolución cultural. Entender el efecto de la estructura sobre la acumulación cultural implica distinguir las *barreras estructurales* (falta de contacto entre individuos) de las *barreras conductuales* (resistirse a compartir conocimiento cultural con otros individuos) [18] que, en conjunto, se

combinan para dar como resultado la *estructura efectiva de la población.*

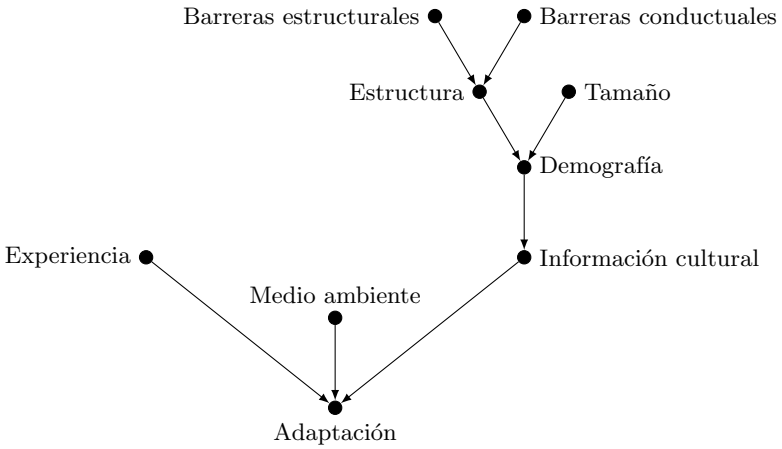


Figura 1.3: Modelo causal aumentado con los distintos tipos de barreras que van a determinar la estructura poblacional

Una hipótesis ingenua sobre el efecto de la estructura sobre la red propone que la alta conectividad favorecería el proceso de acumulación cultural. Sin embargo, una de las conclusiones más importantes de estos primeros estudios afirma que estructuras demasiado interconectadas pueden disminuir las oportunidades para innovar por homogeneización de los comportamientos culturales [19–21]. Niveles muy altos de conectividad hacen que los individuos converjan a soluciones similares, lo cual en consecuencia resulta en niveles bajos de diversidad cultural y una velocidad menor de innovación en comparación a grupos menos conectados. En el otro extremo, niveles muy bajos de conectividad incrementan las chances de pérdida cultural.

1.3. Preliminares

1.3.1. ¿Qué es el Go y cuáles son sus reglas?

El Go es un juego estratégico de mesa para dos jugadores creado en China. Se estima que tiene entre 3 y 4 milenios de antigüedad y, a pesar de tener reglas relativamente sencillas, presenta una enorme complejidad (comparativamente, es muchísimo mas complejo que el ajedrez, por tener millones de veces la cantidad de configuraciones de tablero).

En pocas palabras, el juego consiste en colocar piezas (blancas y negras, un color para cada jugador) sobre una grilla cuadriculada (generalmente de 19×19 posiciones), con el objetivo de conseguir ocupar la mayor superficie posible del tablero con piezas del propio color. Para lograrlo, cada jugador intenta rodear otras piezas con las de su color: al hacer esto, todas las piezas del contrincante son eliminadas del tablero. Al final de la partida, gana el jugador con mayor territorio y fichas capturadas.

El jugador que usa las fichas negras es quien comienza el juego, poniendo la primera ficha en el tablero. Como iniciar la partida permite estar siempre una jugada por delante que el jugador con fichas blancas, existe una regla llamada **Komi**, la cual le asigna al jugador blanco un número de puntos adicionales en el conteo final para contrarrestar esta ventaja. El puntaje extra otorgado por Komi no tiene un valor aceptado a nivel mundial, pero para evitar la posibilidad de empate, se suele usar un valor no entero (generalmente 6,5).

Cuando la diferencia de rango (el "nivel" de experiencia como jugador de Go) entre los dos jugadores es muy grande, existe otra regla adicional llamada **handicap**, mediante la cual el jugador menos hábil comienza con una cierta cantidad de fichas en el tablero a modo de ventaja. La cantidad de fichas

dependerá de la diferencia entre ambos rangos, y las posiciones de las mismas en el tablero son fijas: no son elegidas por el jugador.

1.3.2. Sobre OGS, el sistema de partidas online

OGS (`www.online-go.com`) es un sitio web que permite a jugadores de todo el mundo participar en partidas de Go de manera online. Fue creado en 2005 y actualmente sigue en funcionamiento.

El sitio permite jugar contra otros humanos y contra robots entrenados a distintos niveles de dificultad. Cada jugador cuenta con un nivel de habilidad que se utiliza para compararse contra otros jugadores. Se pueden personalizar diversos aspectos de cada partida, tal como elegir el tamaño del tablero, o si se quiere jugar "en tiempo real" o de manera asincrónica, permitiendo al contrincante tomarse varias horas o hasta días para jugar su turno. En el segundo caso, se permite que un jugador gane a causa de abandono o *timeout* por parte del oponente. Otras opciones incluyen definir valores de *handicap* y *komi*, si se quiere jugar contra un jugador en particular, o si se desea utilizar el sistema de *matchmaking*.

Sistemas de ranking y matchmaking

Cuando un jugador quiere comenzar una partida, puede elegir qué tan alejado de su propio rango desea que se encuentre su oponente como máximo (por ejemplo, si el rango del jugador es X, puede elegir que su contrincante tenga un rango entre $X-3$ y $X+3$). Una vez elegido esto y cualquiera de las otras opciones mencionadas, el sitio buscará a otros jugadores que también estén buscando jugar una partida y que se encuentren dentro de los parámetros configurados. También es posible decidir si se quiere que la partida actualice el rango, o si se quiere mantener como una partida privada: en este último caso, la partida será simplemente "amistosa" y no generará actualizaciones en el ranking de los jugadores.

En los comienzos, el *matchmaking* utilizaba el sistema **Elo** para categorizar a sus usuarios. Este sistema es uno de los primeros y más simples creados con este propósito, y sigue siendo utilizado hoy en día por la asociación internacional de ajedrez para evaluar a los jugadores profesionales. Sin embargo, cuenta con algunas deficiencias, como la lentitud con la cual el resultado de una nueva partida afecta al nuevo ranking de cada jugador, o el hecho de que no maneje la incertidumbre en las estimaciones. Hablaremos de esto en mas detalle en el capítulo 2.

A partir de 2017, OGS decidió comenzar a utilizar el sistema *Glicko-2*. Este modelo mejora la velocidad con la cual se estiman las habilidades de los jugadores e incorpora el concepto de incertidumbre de las estimaciones. Hacemos esta breve distinción ya que cuando los datos con los que trabajamos fueron originalmente obtenidos, todavía el sitio se basaba en Elo para su ranking. Sin embargo, más tarde tuvimos que obtener los nuevos valores basados en Glicko e incorporarlos a la base de datos analizada.

Por último, vale la pena mencionar que el Go tiene su propio sistema de ranking tradicional, que se divide en dos grupos: rankings *kyu* y rankings *dan*. Los rankings kyu son considerados niveles de estudiantes. El nivel más básico es el 30avo kyu. A medida que un jugador mejora, su nivel de kyu disminuye hasta llegar a primer kyu. Pasado este nivel, el jugador llega a los niveles de maestro, conocidos como rankings dan. Éstos van del primer dan al séptimo (de manera similar a los sistemas de artes marciales). Es posible transformar de manera aproximada este sistema de ranking con kyu y dan a los otros sistemas de ranking mencionados anteriormente. OGS utiliza funciones de correspondencia para mostrar el ranking tradicional en los perfiles, pero utiliza Elo/Glicko para realizar el *matchmaking* entre jugadores.

1.3.3. Curvas de aprendizaje poblacional

Teniendo medidas para cuantificar la habilidad, se ha mostrado que las curvas de aprendizaje poblacionales pueden representarse como variantes de la *ley de potencias* o *power law* [22]. Por ejemplo:

$$Habilidad_t = Habilidad_0 \cdot Experiencia^{-t} \tag{1.1}$$

donde $Habilidad_0$ es el primer valor de habilidad conocido (constante) utilizando alguna medida de habilidad, y *Experiencia* es el número de partida jugada (notar que $Experiencia \in \mathbb{N}$). Algunas de las medidas de habilidad utilizadas incluyen tiempo de respuesta, tiempo de juego, poder de cómputo utilizado, etc [22].

Se ha observado empíricamente que estas curvas son rectas al graficarlas en escala logarítmica para ambos ejes. En nuestra base de datos analizamos si esta propiedad se cumplía, lo que veremos en detalle más adelante.

Otras veces las curvas de aprendizaje pueden tomar la forma de curvas sigmoideas, en las cuales es posible diferenciar cuatro áreas distintivas, tal como se ejemplifica con la curva de aprendizaje idealizada en la figura 1.4:

0. **Estado previo al aprendizaje:** esta primera parte de la curva puede estar presente o no. Si lo está, describe el período previo a comenzar el aprendizaje, o los primeros pasos, en donde el individuo aún no absorbió nada de conocimiento.

1. **Comienzos del proceso de aprendizaje:** esta etapa se caracteriza por un período de lento aprendizaje hasta que el individuo finalmente "pega el salto" y comienza a aprender. Este momento es el de mayor aprendizaje.

2. **Zona de *expertise*:** el individuo ya no está aprendiendo al ritmo inicial. Podría decirse que "le queda poco por aprender".

3. **Fin o allanamiento del proceso de aprendizaje:** a esta altura, el individuo ya no tiene nada, o muy poco, por aprender. Se puede decir que "ha llegado a su techo".

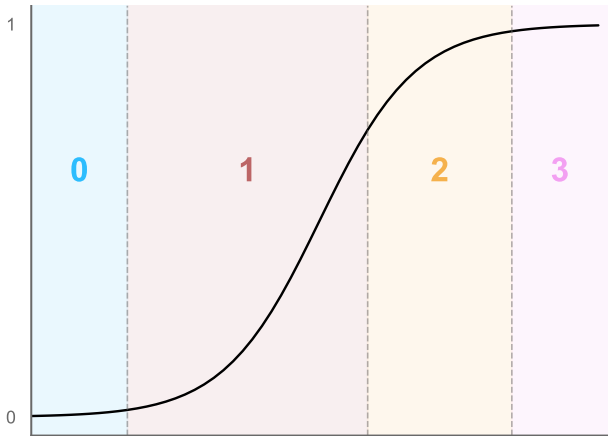


Figura 1.4: Curva sigmoidea mostrando las distintas etapas del aprendizaje: el comienzo (donde mayor es el conocimiento obtenido), el momento donde el aprendizaje comienza a desacelerarse, y por último el fin del proceso de aprendizaje cuando ya se obtuvo todo el conocimiento disponible.

Capítulo 2

Modelos de estimación de habilidad

Existen diversas formas para estimar la habilidad de un jugador. En esta sección exploraremos y compararemos algunas de las más conocidas y aquellas utilizadas en este trabajo. No es el objetivo explicarlas en detalle, sino comunicar sus fundamentos y aspectos más sobresalientes.

2.1. Medida *Elo*

El modelo de ranking *Elo* fue uno de los primeros en proponerse. Fue desarrollado por Arpad Elo en 1959 [23]. Originalmente desarrollado como un sistema de ranking para ajedrez, es capaz de estimar la probabilidad de que un jugador gane frente a otro en base a las habilidades de ambos. Este método fue adoptado en 1970 por la Federación mundial de ajedrez (FIDE).

La idea central de *Elo* es modelar la probabilidad del resultado de un juego a través de las estimaciones de la habilidad de los jugadores. Llamaremos a las estimaciones de habilidad como s_i, dado que i identifica a un jugador cualquiera.

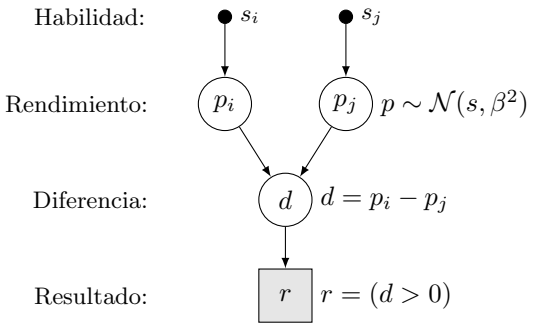


Figura 2.1: Modelo generativo en el que las habilidades causan los resultados observables a través de la diferencia de rendimientos, $d = p_i - p_j$, ambas variables aleatorias centradas en la verdadera habilidad, $p \sim \mathcal{N}(s, \beta^2)$. Gana quien haya obtenido mayor rendimiento, $r = (d > 0)$. Las variables observables se pintan de gris, las ocultas de blanco, y las constantes se muestran como puntos negros.

La figura 2.1 ofrece un modelo causal en el que las habilidades generan el resultado observable. A pesar de que la habilidad de un jugador es una cantidad incierta, la propuesta de *Elo* fue asignarle un valor inicial arbitrario a las estimaciones de habilidad y actualizarlas en base al resultado de las partidas: aumentarán para el jugador que gane, y disminuirán para el que pierda. La pregunta es, *¿cuánto hay que ajustar estos valores?*

La solución de Elo fue usar la **sorpresa** como magnitud de actualización:

- Si el ganador era mucho mejor que su contrincante, la victoria entonces podría suponerse como esperada (en otras palabras, la sorpresa de la victoria sería pequeña), y no habría motivos para pensar que las estimaciones previas eran incorrectas, ni tampoco que debían ser ajustadas.
- Por el contrario, si el ganador tenía una habilidad mucho menor que la del perdedor al momento de la partida, la sorpresa sobre el resultado sería grande. Esto nos señala que nuestras estimaciones estaban desactualizadas, y que deberán ser ajustadas.

Matemáticamente, podemos entonces pensar a la *sorpresa* (que llamaremos Δ) como el complemento de la probabilidad de que se haya dado un resultado r en la partida dadas las habilidades de los jugadores:

$$\Delta = (1 - P(r|s_i, s_j)) \tag{2.1}$$

Y las habilidades de los jugadores luego de la partida t se actualizarán de acuerdo a la siguiente expresión:

$$s_{\text{ganador}_{t+1}} = s_{\text{ganador}_t} + \Delta \qquad s_{\text{perdedor}_{t+1}} = s_{\text{perdedor}_t} - \Delta \tag{2.2}$$

Esta solución logra estimar bastante bien la habilidad relativa entre jugadores partiendo de valores iniciales arbitrarios. Sin embargo, tiene un par de debilidades importantes:

- La regla de actualización expresada en la ecuación (2.2), es simétrica, es decir lo que gana un jugador en habilidad, el otro lo pierde en igual cantidad.
- Debido a que a los jugadores nuevos comienzan con estimaciones arbitrarias (el mismo valor inicial para cualquier individuo), los resultados observados en jugadores nuevos tienden a generan una alta sorpresa y a modificar bruscamente estimaciones que ya habían convergido. Esta debilidad está presente porque no se tiene en cuenta la incertidumbre sobre las estimaciones iniciales de los agentes.

Para resolver este último problema, se propuso una solución *ad-hoc* (adoptada luego por la FIDE): reducir el impacto de la sorpresa en función de la cantidad de veces que el agente ha participado previamente. De esta manera, se buscar que las partidas iniciales de un jugador no generen una alta sorpresa. A medida que el sistema lo va conociendo, entonces puede confiar más en su predicción de habilidad y comenzar a aplicar la regla de actualización más fuertemente.

2.2. Medida *Glicko* (y otras medidas basadas en Inferencia Bayesiana)

En lugar de seleccionar un único valor como estimación inicial de habilidad (como en Elo), una alternativa superadora consiste en distribuir la certidumbre entre todas las posibles hipótesis de habilidad. Este enfoque, conocido como inferencia Bayesiana, ha demostrado ser exitoso en la práctica [24] y garantiza un razonamiento consistente en el manejo de la incertidumbre (o creencias plausibles) [25, 26].

El sistema de calificación *Glicko* [1] (y su sucesor, *Glicko-2* [27]) es una de las primeras extensiones Bayesianas del sistema *Elo*. El sistema Glicko usa un modelo causal similar al de la figura 2.1, pero en vez de evaluar una única hipótesis de habilidad, evalúa todo el espacio de hipótesis en base a una distribución de creencias *a priori*.

Para representar las distribuciones de creencias a priori, Glicko utiliza distribuciones Gaussianas modelando la incertidumbre resecto de la habilidad del agente i como $s_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. La regla práctica para seleccionar distribuciones a priori es que sean lo suficientemente anchas para que no dejen por

fuera ninguna posible hipótesis (en otras palabras, que sean "no informativas"): que todos los jugadores comiencen con el mismo μ y con un σ muy elevado.

La distribución de creencias se irá modificado a partir de los resultados de las partidas utilizando las dos reglas de la probabilidad: la regla de la suma y la regla del producto.

$$P(x) = \sum_y P(x, y) \qquad \text{ó} \qquad p(x) = \int p(x, y)\, dy \qquad \text{(regla de la suma)}$$

donde $P(\cdot)$ y $p(\cdot)$ representan distribuciones de probabilidad discretas y continuas, respectivamente.

Por otro lado, la regla del producto señala que cualquier distribución conjunta puede ser expresada como el producto de distribuciones condicionales uni-dimensionles.

$$p(x, y) = p(x|y)p(y) \qquad \text{(regla del producto)}$$

De estas reglas, se deduce el Teorema de Bayes:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \qquad \text{(Teorema de Bayes)}$$

El Teorema de Bayes permite actualizar de forma óptima las creencias sobre las hipótesis dado un modelo y los datos. En nuestro caso, para cuantificar la certidumbre relativa de las hipótesis de habilidades, utilizaremos la información que nos ofrecen el resultado observado y el modelo causal descripto para resolver:

$$\underbrace{p(\overbrace{\text{Habilidad}_i}^{\text{Oculta}} \,|\, \overbrace{\text{Resultado}}^{\text{Observado}}, \text{Modelo})}_{\text{Posterior}} = \frac{\overbrace{P(\text{Resultado} \,|\, \text{Habilidad}_i\,, \text{Modelo})}^{\text{Verosimilitud}} \overbrace{p(\text{Habilidad}_i)}^{\text{Prior}}}{\underbrace{P(\text{Resultado} \,|\, \text{Modelo})}_{\text{Evidencia o predicción a priori}}} \tag{2.3}$$

donde la única variable libre es la hipótesis de habilidad del agente i.

El *prior* cuantifica la incertidumbre sobre la habilidad, y el *posterior* contiene la incertidumbre que queda luego de ver el nuevo dato (es decir, el resultado del evento). Tanto la *verosimilitud* como la *evidencia* son probabilidades del resultado observado, por lo que pueden ser vistas como predicciones. Notar que, como los resultados son variables discretas, las probabilidades también lo serán y se las nota con P mayúscula, como se menciona en la expresión (regla de la suma). Debido a que la evidencia es la misma para todas las hipótesis, el único factor que efectivamente actualiza nuestras creencias es la verosimilitud.

2.3. Medida *TrueSkill*

El método de *TrueSkill* fue introducido en 2007 por Herbrich et al. [3]. Al igual que *Glicko*, utiliza el mismo modelo causal que *Elo*. El éxito de la solución TrueSkill se basa en la aplicación de un método eficiente para calcular la distribución Gaussiana que mejor aproxima al posterior exacto.

A modo de ejemplo, consideremos un caso ganador ($p_i > p_j$) usando priors Gaussianos $\mathcal{N}(\, s \,|\, \mu, \sigma^2)$. La diferencia de rendimientos, $d = p_i - p_j$, también se puede expresar como una Gaussiana centrada en la diferencia de las medias de las estimaciones a priori ($\mu_i - \mu_j$), con una varianza que incorpora la incertidumbre de ambas estimaciones (σ) y la varianza de ambos rendimientos (β): $\mathcal{N}(\, d \,|\, \mu_i - \mu_j\,,\ 2\beta^2 + \sigma_i^2 + \sigma_j^2\,)$.

Como observamos que el agente i ganó, sabemos por el modelo causal que la diferencia de rendimientos oculta fue en efecto positiva. Por lo tanto, la predicción a priori del resultado observado (o *evidencia*) es la densidad acumulada (Φ) de todos los valores positivos de la diferencia de rendimientos:

$$\overbrace{P(r)}^{\text{Evidencia}} = 1 - \Phi(0 \,|\, \overbrace{\mu_i - \mu_j}^{\substack{\text{Diferencia}\\ \text{esperada}}}, \overbrace{2\beta^2 + \sigma_i^2 + \sigma_j^2}^{\substack{\text{Incertidumbre}\\ \text{total}}}) \tag{2.4}$$

La evidencia es una predicción hecha con todas las hipótesis a priori. Como la evidencia es constante, la incertidumbre a posteriori de cada hipótesis es proporcional al producto de su incertidumbre a priori y su verosimilitud, como se muestra en la siguiente expresión:

$$\underbrace{p(\, s_i \,|\, r\,)}_{\text{Posterior}} \propto \underbrace{1 - \Phi(0 \,|\, s_i - \mu_j,\, 2\beta^2 + \sigma_j^2)}_{\text{Verosimilitud } P(r|s_i)} \underbrace{\mathcal{N}(s_i \,|\, \mu_i,\, \sigma_i^2)}_{\text{Prior } p(s_i)} \tag{2.5}$$

donde el posterior normalizado se obtiene dividiendo el lado derecho por la evidencia $P(r)$ como está definida en la ecuación (2.4).

Es interesante notar las similitudes y diferencias entre la verosimilitud y la evidencia. La verosimilitud cuantifica la misma densidad acumulada que la evidencia, pero centrada en la diferencia entre la hipótesis que estamos evaluando s_i y la estimación media del oponente μ_j con una varianza que incluye todas las incertidumbres salvo la de la propia hipótesis s_i. En otras palabras, la verosimilitud no es más que la predicción a priori del resultado observado suponiendo verdadera la hipótesis de habilidad que estamos evaluando. La figura 2.2 muestra, en términos gráficos, el procedimiento de actualización que se realiza en la ecuación (2.5).

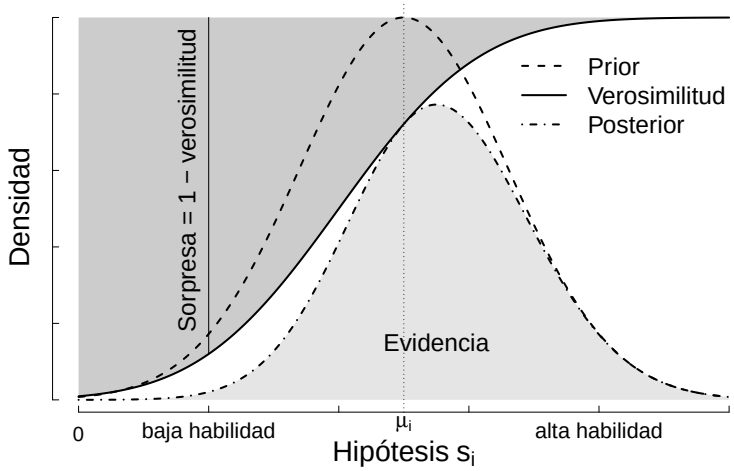


Figura 2.2: Actualización de creencias para el caso ganador. El posterior proporcional se obtiene como el producto de la distribución a priori (distribución gaussiana) y la verosimilitud (distribución gaussiana acumulada). La evidencia es la integral del posterior proporcional. Las distribuciones no están en la misma escala: la distribución a priori integra 1, mientras que la verosimilitud va de 0 a 1.

La sorpresa, definida como el complemento de la verosimilitud, funciona como un filtro para el prior. El posterior puede interpretarse, entonces, como el remanente de la densidad del prior que no fue filtrado por la verosimilitud. En la región de hipótesis de muy alta habilidad, donde el resultado ganador no nos hubiera generado casi ninguna sorpresa ($\lim_{s_i \to \infty} P(r|s_i) = 1$), el posterior recibe casi toda la masa del prior. En cambio, en la región de hipótesis de muy baja habilidad, donde el resultado habría generado mucha sorpresa ($\lim_{s_i \to -\infty} P(r|s_i) = 0$), el posterior no recibe casi nada de la masa del prior.

Es importante remarcar que el posterior, aunque se parezca, no es una distribución gaussiana, lo que nos impide usar la ecuación (2.5) iterativamente. Es por esta razón que se utiliza un método eficiente

para calcular la distribución Gaussiana que mejor aproxima a la posterior exacta, en términos de la divergencia KL, tal como se expresa en la ecuación (2.6):

$$\widehat{p}(s_i|r,s_j) = \underset{\mu,\sigma}{\operatorname{argmin}}\ \mathrm{KL}(\, p(s_i|r,s_j)\,||\,\mathcal{N}(s_i|\mu,\sigma^2)\,) \tag{2.6}$$

Esta expresión se entiende como la combinación de μ y σ que minimicen la divergencia KL entre el posterior (p) y la distribución Gaussiana que mejor lo aproxime ($\mathcal{N}$). De esta manera se propone un mecanismo operativo para poder obtener las estimaciones.

2.4. Medida *TrueSkill Through Time (TTT)*

El enfoque adoptado por TrueSkill para tratar el proceso dinámico usa el posterior como prior del siguiente evento.

$$\widehat{\text{Posterior}}_t \propto \widehat{\text{Verosimilitud}}_t \overbrace{\widehat{\text{Verosimilitud}}_{t-1} \ldots \widehat{\text{Verosimilitud}}_2 \underbrace{\widehat{\text{Verosimilitud}}_1 \text{Prior}_1}_{\widehat{\text{Posterior}}_1 \text{ como Prior}_2}}^{\widehat{\text{Posterior}}_{t-1} \text{ como Prior}_t} \tag{2.7}$$

donde $\widehat{\text{Posterior}}_i$ y $\widehat{\text{Verosimilitud}}_i$ representan las aproximaciones inducidas por la ecuación (2.6) en el i-ésimo evento. Si consideramos la verosimilitud como un filtro del prior, cada posterior puede ser visto como una acumulación de todos los filtros anteriores. Por eso a este enfoque se lo conoce como de "filtrado".

Debido a que el enfoque de filtrado no surge de ningún modelo probabilístico, sufre de una serie de problemas, todos relacionados con el hecho de que la información se propaga en una sola dirección a través del sistema. El más obvio es que el inicio de toda secuencia de estimaciones siempre tiene alta incertidumbre, pero también sufre de desacoplamientos "temporales" y "espaciales": aunque las diferencias relativas entre estimaciones contemporáneas al interior de comunidades bien conectadas sean correctas, las estimaciones separadas en el tiempo y entre comunidades poco conectadas suelen ser incorrectas.

A fin de obtener buenas estimaciones iniciales y garantizar comparabilidad temporal y espacial, necesitamos un modelo causal del proceso temporal que vincule todas las actividades históricas. En la base de datos analizada, esto implica evaluar una red compleja compuesta por miles de eventos. Por esta razón, la figura 2.3 muestra un fragmento del modelo causal (los vecinos más próximos de una secuencia de estimaciones).

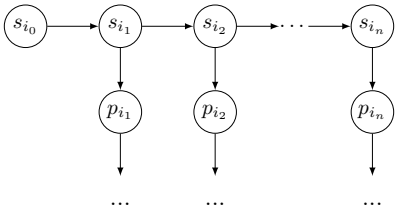


Figura 2.3: Representación esquemática de un único modelo causal para toda la historia de eventos. Solo se muestran los vecinos de una secuencia de estimaciones.

Este enfoque, conocido como *smoothing*, es el implementado por TrueSkill Through Time [28]. Al aplicar las reglas de la probabilidad sobre este modelo temporal, la información histórica se propaga naturalmente hacia todo el sistema, resolviendo los problemas del enfoque de filtro. El prior para un evento intermedio depende así tanto de la información pasada como futura.

$$\text{Prior}_{i_t} = \text{Prior}_{i_0} \underbrace{\prod_{k=1}^{t-1} \text{Verosimilitud}_{i_k}}_{\text{Información pasada}} \underbrace{\prod_{k=t+1}^{T_i} \text{Verosimilitud}_{i_k}}_{\text{Información futura}}$$

donde T_i es la cantidad total de eventos del agente i y Prior_{i_0} es el prior inicial del agente i. Esto produce una mutua dependencia entre verosimilitudes, obligándonos a implementar un algoritmo iterativo para resolver la inferencia. Utilizamos las últimas verosimilitudes disponibles hasta alcanzar la convergencia. Este procedimiento requiere pocas iteraciones lineales sobre los datos.

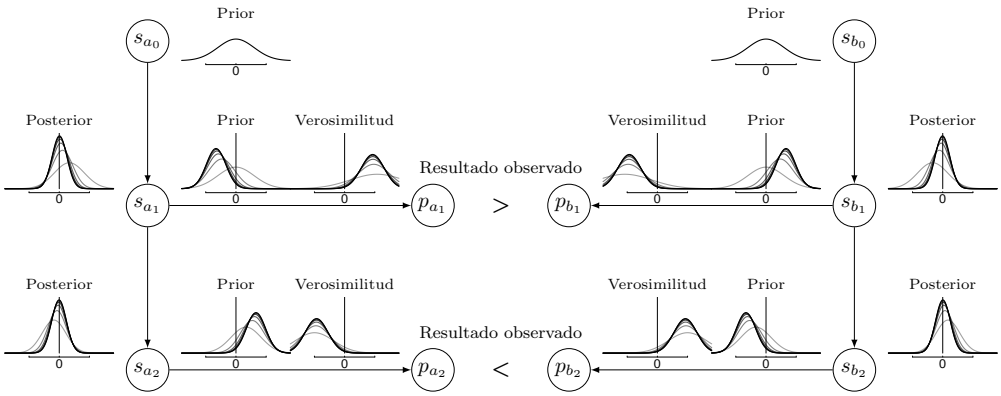


Figura 2.4: Convergencia de TrueSkill Through Time en una historia con sólo dos eventos y dos agentes.

TrueSkill Through Time supera a TrueSkill y a otros modelos de filtrado. En la figura 2.4 se muestra el comportamiento de las estimaciones en un conjunto de datos de dos agentes y dos eventos. La primera partida la gana el jugador a, $p_{a_1} > p_{b_1}$, y la segunda la gana el jugador b, $p_{a_2} < p_{b_2}$. La luminosidad de las curvas indican el orden en la convergencia. La primera (la más clara) se corresponde a la solución TrueSkill. La última (la más oscura) se corresponde a la solución TrueSkill Through Time. Alcanzada la convergencia, el posterior se centra en cero para ambos jugadores, indicando que tienen misma habilidad. TrueSkill Through Time recupera las verdaderas diferencias entre habilidades ocultas incluso al principio de la secuencia.

Capítulo 3

Topología y análisis de su evolución

Una red de jugadores puede ser entendida como un grafo donde los nodos representan a los jugadores, y los ejes, a las partidas jugadas entre ellos. Estos jugadores aprenden durante estas partidas, intercambiando conocimiento, información o cultura, que será difundida por la red, influenciando a su vez el aprendizaje de todos los otros jugadores. El flujo de información, entonces, estará condicionado por la topología (o "forma") de la red generada en base a la interacción entre los jugadores.

Debido a la naturaleza dinámica que tiene la red formada por la interacción de jugadores, estamos obligados a analizar la estructura de una red que evoluciona en el tiempo. Primero analizaremos las técnicas para estudiar grafos dinámicos. Y luego, detallaremos las magnitudes que finalmente utilizamos para analizar la base de datos.

3.1. Grafos dinámicos y su análisis

Hasta ahora hemos hablado de grafos en su definición estándar: un par (V, E), donde V es un conjunto finito de nodos, y E es un conjunto de ejes, es decir, pares $\{u, v\}$, ordenados o no, con $u, v \in V$.

Esta definición de un grafo *estático* claramente es aplicable a innumerables contextos concretos. Sin embargo, no es capaz de describir correctamente todos los escenarios. Algunas aplicaciones, como el caso de ciertos escenarios relacionados a *networking*, en realidad están muy lejos de ser modelos estáticos: el conjunto de nodos y los ejes que los conectan están constantemente sujetos a cambios a medida que pasa el tiempo.

Podremos definir entonces a un *grafo dinámico* como un grafo cuyo conjunto de ejes y/o nodos cambia en función del tiempo [29]. En particular, lo llamaremos *nodo-dinámico* si el conjunto de nodos es el que varía, o *eje-dinámico* si es el conjunto de ejes. Es posible que un grafo sea eje- y nodo-dinámico a la vez. La figura 3.1 esquematiza dos momentos de un grafo dinámico. La estructura del grafo sufre una profunda modificación entre el instante inicial y el final: por un lado, aparece un nuevo nodo (con un correspondiente eje) y, por el otro, solo se conserva uno de los ejes preexistentes.

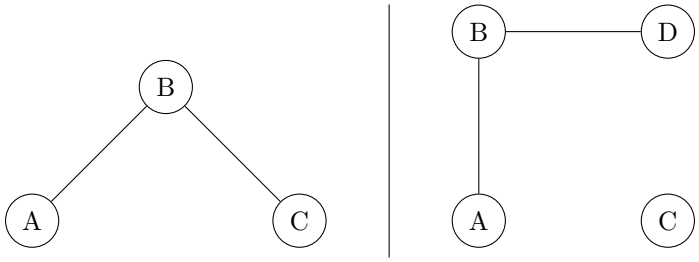


Figura 3.1: Un grafo eje- y nodo-dinámico G_d en dos instantes distintos. A la izquierda, el instante t_0, en donde $G_d^{t_0} = (V_{t_0}, E_{t_0})$, con $V_{t_0} = \{A, B, C\}$ y $E_{t_0} = \{\{A, B\}, \{B, C\}\}$. A la derecha, el instante t_1, en donde $G_d^{t_1} = (V_{t_1}, E_{t_1})$, con $V_{t_1} = \{A, B, C, D\}$ y $E_{t_1} = \{\{A, B\}, \{B, D\}\}$

En particular, es fácil entender por qué nuestro caso de estudio se puede modelar como un grafo eje y nodo-dinámico: a medida que pasa el tiempo, nuevos jugadores (nodos) se irán sumando a la comunidad en línea y, cada vez habrá más partidas entre jugadores (ejes).

La siguiente pregunta es cómo hacer un análisis (topológico, por ejemplo, utilizando las medidas de centralidad descriptas en 3.2) sobre grafos dinámicos. La mayoría de los estudios realizados sobre grafos se enfocan en grafos estáticos, por lo que la cantidad de opciones es limitada. Sin embargo, hay al menos dos alternativas: utilizar medidas que consideran el tiempo, y hacer un análisis estático utilizando la técnica de ventana deslizante.

3.1.1. Medidas que consideran el tiempo

Las medidas que consideran el tiempo en el que se observa el grafo permiten, a partir del grafo completo, estudiar ciertas propiedades estructurales o del flujo de información en función del tiempo.

Como se dijo anteriormente, no existen demasiados trabajos sobre medidas de este tipo para grafos dinámicos [30]. Algunos de los que existen se centran principalmente en la evolución de los ejes pero no de los nodos. Otros utilizan conceptos que no parecen ser relevantes para nuestro caso de estudio, como buscar caminos "mínimos y más rápidos", es decir, los que permitan llegar de un nodo a otro no solo utilizando la distancia más corta, sino también en la menor cantidad de tiempo. Por otro lado, puede ser realmente costoso calcular algunas de estas medidas cuando el grafo es tan grande, como es nuestro caso.

Desde un punto de vista más pragmático, la cantidad de medidas disponibles es significativamente menor que las existentes para grafos estáticos, y no existen paquetes de software fácilmente disponibles para calcularlas.

Por estas razones, decidimos descartar esta opción para esta tesis.

3.1.2. Análisis estático con ventana deslizante o *snapshots*

La otra alternativa es dividir la línea temporal en la que evoluciona el grafo dinámico en pequeñas ventanas. Cada una de estas ventanas tendrá algún tiempo de duración preestablecido (por ejemplo, un día). Para cada ventana, se obtendrá el subgrafo de todos los nodos y ejes que estén presentes en ese período de tiempo. Cada uno de estos subgrafos será un grafo estático, por lo que podrán utilizarse todas las medidas y algoritmos existentes para grafos estáticos.

Como una forma de validar las conclusiones obtenidas, se repiten los análisis utilizando distintos tamaños de ventanas, tanto más grandes como más pequeñas, con el fin de estar seguros que los resultados no solo aparecen en un tamaño de ventana determinado, sino que es un fenómeno estructural. Un ejemplo de este tipo de análisis se puede ver en el trabajo de Kossinets y Watts [31]. Hemos tomado este trabajo

como referencia para parte de nuestro análisis.

3.2. Medidas de centralidad

Existen diversas propiedades de los grafos que pueden ser usados como una manera de caracterizarlos. Entre ellas, podemos mencionar al coeficiente de *clustering*, *largo promedio o característico de los caminos*, *modularidad*, y a la *centralidad*. Nos enfocaremos en la última.

La **centralidad** es una medida que se calcula para cada vértice, y representa qué tan importante (para alguna definición de importancia) resulta ese vértice dentro del grafo. Hay numerosas definiciones de *importancia* y, por ende, numerosas propuestas de medidas de centralidad.

A continuación explicamos, brevemente, las medidas de centralidad que utilizamos como parte de nuestro análisis. Las mismas fueron elegidas con el objetivo de caracterizar el proceso de transmisión de información en una red, lo que se alinea con nuestro objetivo de analizar el efecto del intercambio de la información cultural.

Es importante resaltar que todas las medidas que mencionen a la *distancia* entre nodos, consideran la distancia **pesada** entre ellos.

3.2.1. Degree Centrality

La centralidad en base al grado de los nodos es la medida de centralidad más simple. Bajo esta definición de centralidad, la importancia de un nodo se define como el grado (la cantidad de conexiones con otros nodos): a mayor conectividad, mayor será la centralidad del nodo en el grafo. Para esta definición de grado, no tendremos en cuenta el peso de los ejes; es decir, es el grado no pesado de cada nodo.

Aunque es muy simple de computar, no tiene en cuenta otras características potencialmente importantes a la hora del intercambio de información entre agentes, como posiciones "privilegiadas" en el grafo (ser puente entre dos componentes por ejemplo).

3.2.2. Closeness

Closeness es la inversa del promedio de las distancias mínimas de un nodo con todos los otros:

$$C(p) = \frac{n-1}{\sum_{q \in V_p} d(p,q)} \tag{3.1}$$

donde $d(p,q)$ es la mínima distancia entre p y q, normalizado por el número total de nodos alcanzables $(n-1)$, y V_p es el conjunto de nodos del grafo alcanzables desde p. Las distancias son números enteros y no se puede estar a distancia menor a uno. Al ser un promedio invertido, su imagen es el intervalo $[0,1]$. Cuanto más alta la centralidad de un nodo, menor será la distancia a los otros nodos. Intuitivamente, este valor nos marca *qué tan cerca* está (en promedio) un nodo de todo el resto.

Esta medida se ve afectada tanto por la cardinalidad del grafo como por su conectividad: cuanto más grande la red, menor será la centralidad de los nodos que estén "lejos" de todos los otros. En el caso de los nodos que se encuentren cerca del resto, entonces su centralidad será alta. En los siguientes ejemplos, el valor dentro de cada nodo es el valor de *closeness* para ese nodo.

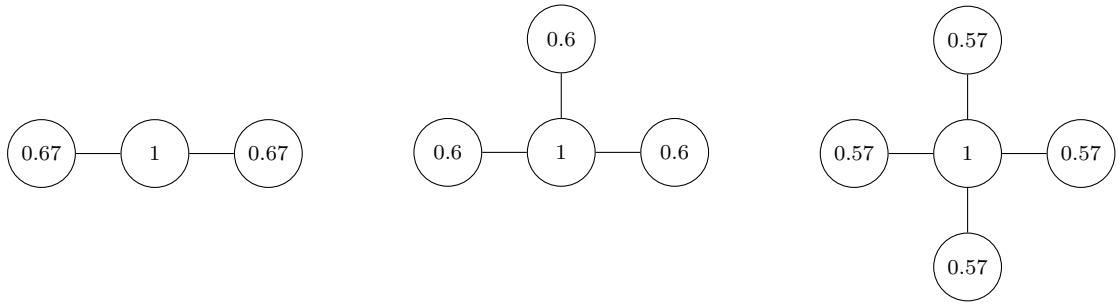


Figura 3.2: Ejemplos de distintos grafos y el valor de *closeness* para cada nodo. Intuitivamente, se ve que los nodos con mayor valor son los que se ubican en la parte central del grafo.

Como se ha dicho, en estos grafos se puede observar que al agregar nodos conectados al nodo central, la centralidad de todos los periféricos disminuye en igual proporción, mientras que la del nodo central se mantiene en 1 por estar conectado a todos los otros. En estos casos no estamos aumentando la conectividad (es decir, no estamos agregando ejes) sino que agregamos nodos.

Por otro lado, al conectar todos los nodos entre sí sin agregar nodos nuevos, las centralidades aumentan: pasarán a ser 1, porque se puede llegar a cualquier otro nodo en un único paso (ha aumentado la conectividad).

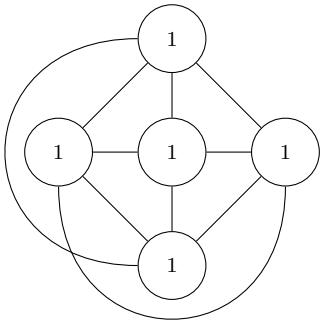


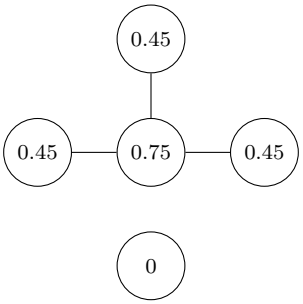
Figura 3.3: Ejemplo de un grafo completo en el que el *closeness* es el mismo en todos los nodos debido a la alta conectividad que presenta este tipo de grafos

Por último, resta mencionar qué sucede en el caso de grafos disconexos. Existe una propuesta, llamada *Wasserman and Faust Closeness*, que escala proporcionalmente el valor de centralidad en base al tamaño de la componente conexa en la cual se encuentra el nodo:

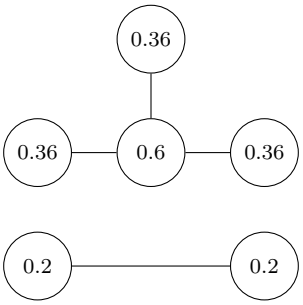
$$C_{WF}(p) = \frac{n-1}{N-1} \cdot \frac{n-1}{\sum_{q \in V_p} d(p,q)} \tag{3.2}$$

Como se dijo anteriormente n representa la cantidad de nodos alcanzables desde p. Llamaremos N a la cantidad total de nodos en el grafo. Si hay una única componente conexa, notemos que $n = N$. **Nosotros utilizaremos esta modificación en nuestro trabajo.**

Veamos dos ejemplos con más de una componente para ilustrar esto. En el caso de que haya una segunda componente con un único nodo, como no tiene nodos adyacentes, su valor de *Closeness* será igual a cero:



Por el contrario, si la segunda componente sí tiene más de un nodo, los valores de *Closeness* no serán nulos. Sin embargo, notar que son menores que los de la componente conexa más grande gracias la modificación de *Wasserman and Faust*:



De esta forma, esta modificación propone una forma de pesar la conectividad cuando hay distintas componentes conexas en un grafo, una situación que emerge naturalmente en nuestro trabajo debido a que la interacción entre jugadores muchas veces genera componentes disconexas.

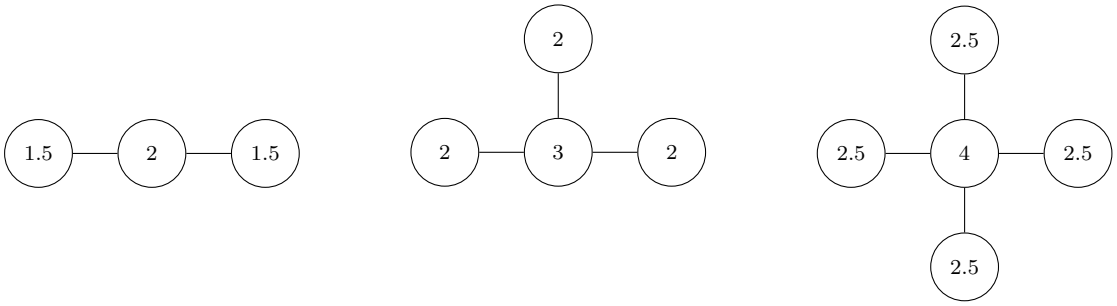
3.2.3. Harmonic Centrality

Harmonic Centrality puede ser vista como una modificación de *Closeness Centrality* en la cual se toma la suma de los inversos de las distancias en lugar del inverso de la suma:

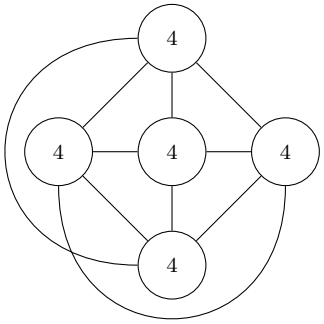
$$C_H(p) = \sum_{q \in V_p : p \neq q} \frac{1}{d(p,q)} \tag{3.3}$$

donde V_p representa el subconjunto de nodos alcanzables desde p. Notar que esto no es un problema para grafos con más de una componente conexa, ya que cada uno será tratado, a efectos de calcular la medida, como un grafo individual.

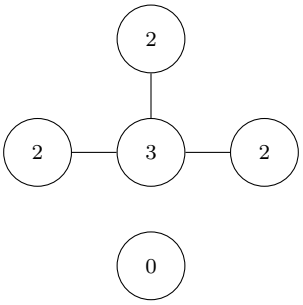
Otra diferencia respecto a *Closeness* es que, aunque en este caso también el valor de la centralidad depende de la cantidad de nodos, aquí incrementa para todos los nodos al aumentar la cardinalidad:



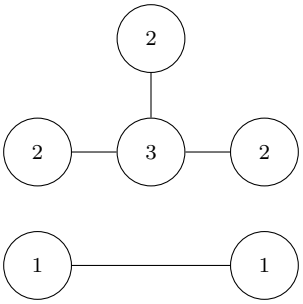
Por otro lado, notar que la imagen se encuentra en el intervalo $[0, N-1]$. En el caso de los nodos que se encuentran a distancia 1 de todo el resto, su valor de *Harmonic Centrality* será igual a la cantidad de nodos del grafo menos 1. Para un grafo completo se verá así:



Como esta medida está definida para grafos no conexos, en el caso de un nodo que se encuentre desconectado de todos los otros, su valor de centralidad será igual a cero:



Notar además que el agregar nodos únicamente afecta la centralidad de aquellos otros que se encuentren en la misma componente conexa a la que fue agregado el nuevo nodo.



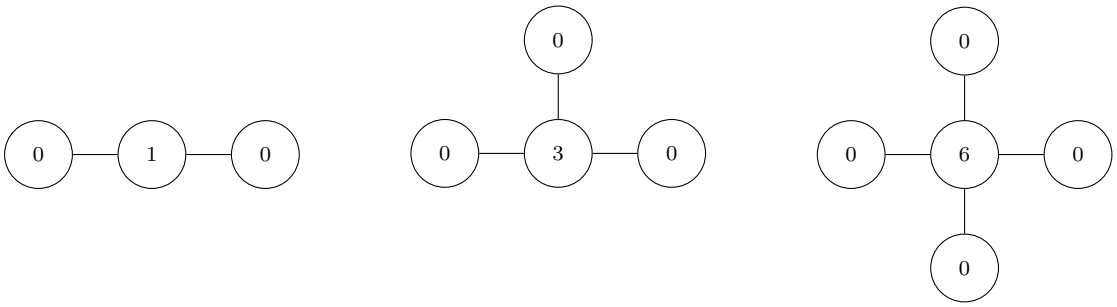
3.2.4. Betweenness

Es una medida de centralidad también basada en las distancias entre nodos, pero que beneficia a los nodos "puente". En este caso se calculan las distancias entre todos los pares de nodos del grafo y se calcula la proporción de ellos que pasan por un nodo en particular. El valor de *betweenness* de un nodo p se define como la suma de las proporciones de caminos mínimos entre todos los pares q y r que pasan por p.

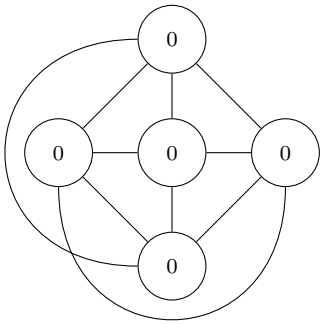
$$B(p) = \sum_{q \neq p \neq r \wedge q < r} \frac{\sigma_{qr}(p)}{\sigma_{qr}} \tag{3.4}$$

donde σ_{qr} es la cantidad de caminos mínimos entre q y r, y $\sigma_{qr}(p)$ es la cantidad de caminos mínimos entre q y r que pasan por p. Su imagen se encuentra en el intervalo $[0, E]$, siendo $E = (N-1)(N-2)/2$ en grafos no dirigidos. El máximo se obtiene en el caso de que todos los caminos mínimos del grafo (entre cada par de nodos) pasen por el nodo p. El valor máximo E puede ser utilizado para obtener una versión normalizada de *betweenness centrality*.

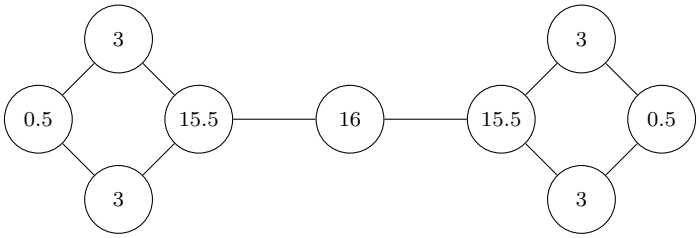
Si usamos los mismos grafos utilizados para ejemplificar las centralidades anteriores, podemos ver que, como todos los caminos pasan por el nodo central, éste siempre tendrá el valor máximo de centralidad. Por otro lado, como los nodos periféricos no forman parte de ningún camino que no los contenga, su valor de *betweenness* es siempre cero:



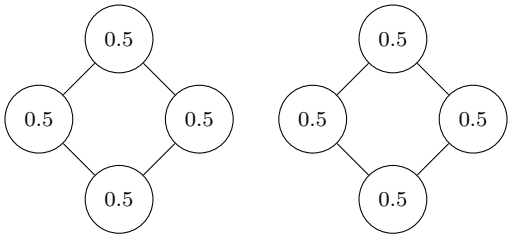
Sin embargo, más interesante es el caso en donde el grafo es completo: en este escenario, como es posible llegar a todo par de nodos sin pasar por un tercer nodo "puente" (es decir, un nodo que, de quitarlo, incrementaría la cantidad de componentes conexas del grafo), el valor de *betweenness* para todos los nodos es nulo.



Por el contrario, cuando el grafo sí contiene un nodo puente, su valor de centralidad será el más alto:



Por último, cuando el grafo es disconexo, no es un inconveniente, dado que se calculará el valor de centralidad de cada componente de manera individual. Por ejemplo:



3.2.5. Communicability Betweenness

Communicability Betweenness es una medida de centralidad que se basa en el concepto de *communicability*, definido como

$$G_{pq} = \sum_{k=0}^{\infty} \frac{1}{k!} W_{pq}^{k} \tag{3.5}$$

donde W_{pq}^{k} es el número de caminos (no necesariamente mínimos) que conectan a p y q con largo k. Esta medida se define en el trabajo de Estrada et al. [32] con la motivación de no tener en cuenta únicamente al camino mínimo entre dos nodos para saber qué tan bien conectados están, sino todos los caminos posibles, pero pesando los términos de la suma para darle menos influencia a los caminos más largos (razón por la cual se divide por el factorial del largo). Como se puede ver, esta medida es muy similar a *Betweenness*, como su nombre lo indica, pero teniendo en cuenta la proporción de todos los caminos que pasan por un nodo en lugar de únicamente los caminos mínimos.

Esta forma de pesar la suma permite una interpretación física de la comunicabilidad, que representa la probabilidad de que una partícula, partiendo de un nodo p, llegue al nodo q luego de moverse por la red por fluctuaciones de temperatura. Si se supone que esta fluctuación de temperatura es de origen aleatorio, podemos pensar a la partícula como, por ejemplo, un portador de información en una sociedad.

Notemos que la expresión (3.5) puede expresarse de la siguiente forma, donde $\mathbf{A}$ es la matriz de adyacencia del grafo:

$$G_{pq} = \sum_{k=0}^{\infty} \frac{\left(\mathbf{A}^{k}\right)_{pq}}{k!} \tag{3.6}$$

que, por propiedades de las series de Taylor, converge a

$$G_{pq} = \left(e^{\mathbf{A}}\right)_{pq} \tag{3.7}$$

Definimos *Communicability Betweenness* como

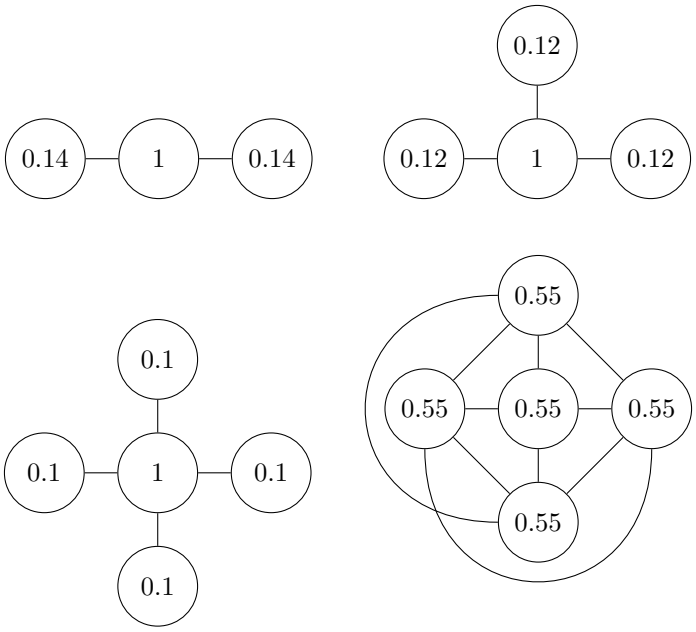
$$C_{Co}(r) = \frac{1}{C} \sum_{p} \sum_{q} \frac{G_{prq}}{G_{pq}}, p \neq q, p \neq r, q \neq r \tag{3.8}$$

con $C = (n-1)^2 - (n-1)$ como factor de normalización, y G_{prq} como la comunicabilidad pero sólo teniendo en cuenta los caminos que pasan por r. Esto último puede expresarse matemáticamente de la siguiente forma:

$$G_{prq} = \left(e^{\mathbf{A}}\right)_{pq} - \left(e^{\mathbf{A}(r)}\right)_{pq} \tag{3.9}$$

donde $\mathbf{A}(r)$ representa la matriz de adyacencia de un grafo G' obtenido a partir de la matriz de adyacencia A de un grafo G, pero poniendo ceros en la fila y columna r; es decir, eliminando todos los ejes conectados al nodo r.

En resumen, *Communicability Betweenness* captura qué proporción de todos los caminos (no necesariamente mínimos) del grafo pasan por r.



Para los grafos utilizados anteriormente como ejemplos, podemos ver que, como es usual, en los grafos no completos, el nodo central siempre tiene la mayor centralidad. En el completo, todos los nodos tienen la misma centralidad, ya que todos los vecinos están a un paso de distancia.

Es interesante ver el siguiente ejemplo, ya que deja en evidencia cómo afecta el largo de los caminos al valor de la centralidad:

En este grafo, hay un único camino entre cada par de nodos. Sin embargo, cuanto más alejados los nodos entre sí, más drásticamente disminuye el valor de la centralidad. Esto se explica por el factorial del largo del camino dividiendo la suma.

3.2.6. Load Centrality

Load Centrality en general reduce a la fracción de todos los caminos mínimos que pasan por p, pero su definición formal utiliza un proceso de flujo:

$$C_L(p) = \sum_{q \neq p \neq r : q \in V \wedge r \in V_q} \rho_{q,r}(p) \tag{3.10}$$

donde $\rho_{q,r}(p)$ representa la cantidad de flujo (enviada del nodo q al r) que pasa por p, V es el conjunto de nodos del grafo, y V_q es el subconjunto de nodos alcanzables desde q.

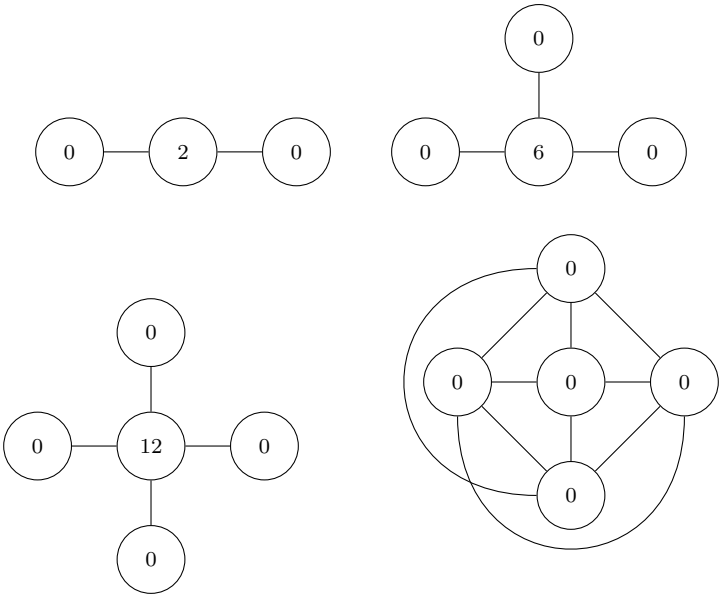
Este flujo parte de q y se transmite entre nodos de la siguiente forma: para cada unidad de flujo x que llegue a un nodo v con destino r, v dividirá x en partes iguales entre todos sus vecinos que tengan distancia mínima a r. Si hay un solo camino con distancia mínima, todo el flujo irá por ahí.

Por ejemplo, si $x = 1$ y p tiene tres vecinos:

- El primer vecino v_1 está a la distancia mínima de r: 3;
- El segundo vecino v_2 también está a distancia 3, lo que significa que hay otro camino mínimo que lo incluye;
- El tercer vecino v_3 está a distancia 5 de r, por lo que no será tenido en cuenta

Como v_1 y v_2 están a distancia mínima de r, cada uno recibirá $x = 0{,}5$. v_3 recibirá 0 ya que no es parte de un camino mínimo entre q y r. Este proceso se repite para cada nodo, y el flujo o suma total que pase por cada nodo v se considerará la *carga* (o sea, la *load centrality* de v).

Load centrality es una modificación de *Betweenness*, ya que en lugar de considerar la proporción de caminos mínimos sobre caminos totales entre cada par de nodos, toma en cuenta la fracción de **todos** los caminos mínimos que pasan por un nodo. Otra diferencia tiene que ver con que para *betweenness*, el camino $a \to b$ y el $b \to a$ son el mismo en un grafo no dirigido y solo se cuenta una vez, mientras que para *load centrality*, ambos caminos se cuentan por separado. Esto resultará en que muchas veces, *load* será el doble de *betweenness*.



Sin embargo, existen casos en los cuales esto no se cumple. En Goh et al. [33] se menciona, particularmente, el caso de los grafos en los cuales el grado de sus nodos tiene una distribución de ley de potencias. Esto hace que sea difícil mostrar una intuición en grafos pequeños, pero la idea básica consiste en que, si existen múltiples caminos mínimos entre los nodos q y r y *para cada vértice intermedio* hay múltiples ejes que son parte de los distintos caminos mínimos, entonces el flujo enviado inicialmente desde la fuente se verá cada vez más "diluido" a medida se avance en el camino.

Esto es consecuencia de que, en cada nodo, el flujo será repartido en partes iguales entre todos los ejes (asumiendo que todos comparten el mismo peso mínimo). En el sentido opuesto, los nodos a los que lleguen múltiples de estos caminos recibirán mayor flujo y tendrán una mayor centralidad.

Veamos el siguiente ejemplo. Arriba se encuentran los valores calculados utilizando *betweenness*, y abajo, usando *load centrality*.

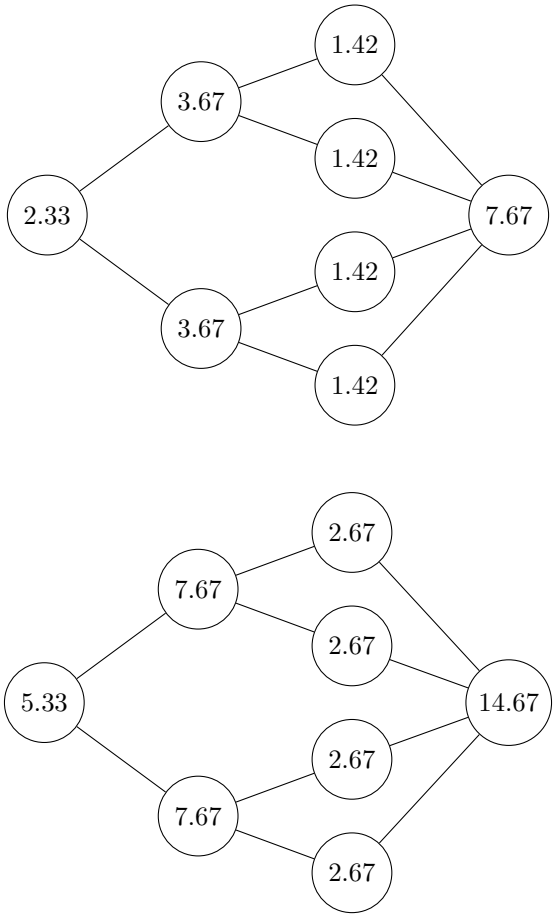


Figura 3.6: Grafos pequeños donde el grado de sus nodos tiene una distribución de ley de potencias. El primero muestra los valores de centralidad *Betweenness* para cada nodo, y el de segundo, utilizando load centrality.

Como se puede observar, los nodos de la tercer y cuarta columnas en la figura para *load centrality* tienen valores menores que el doble de su valor correspondiente de *betweenness*, producto de la "dilución" del flujo. En las dos primeras el efecto es el contrario, y su valor es mayor que el doble.

3.2.7. Current-Flow Betweenness

Esta medida modifica la definición un tanto básica de *Betweenness* para incluir conceptos de redes de flujo eléctrico, que definiremos a continuación.

Una red eléctrica $N = (G; c)$ se define como un grafo G con ejes pesados en base a $c : E \to \mathbb{R}_{>0}$ que indican la conductividad eléctrica de cada eje, es decir, cuánto flujo son capaces de transportar. Nos va a interesar analizar la red en términos de cómo se mueve el flujo eléctrico por la misma.

Los nodos de la red pueden ser puntos de entrada (s) o de salida (t) de la corriente eléctrica (ambos llamados *outlets*), o ser neutros. Esto estará definido por la función de suministros $b : V \to \mathbb{R}$. A los propósitos de esta definición, se considera que una única unidad de corriente entrará solamente por un nodo $s \in V$ y saldrá únicamente por un nodo $t \in V \setminus \{s\}$, con lo cual b queda definida de la siguiente forma:

$$b_{st}(v) = \begin{cases} 1 & v = s, \\ -1 & v = t, \\ 0 & \text{sino} \end{cases} \tag{3.11}$$

Notar que debe haber tanta corriente entrando como saliendo de la red, por lo que $\sum_{u \in V} b(v) = 0$. Como sólo estamos teniendo en cuenta los casos en los que una única unidad de corriente entra a la red por un único suministro, notemos que entonces debe haber también un único punto de salida.

Por último, definiremos una corriente eléctrica en términos matemáticos como una función $x : \overrightarrow{E} \to \mathbb{R}$ que satisface las dos leyes de Kirchhoff:

- **Ley de corrientes:** la suma de corrientes que entran a un nodo v será igual a la suma de corrientes que salen de v
- **Ley de tensiones:** para cada ciclo en G, la suma de las tensiones (potenciales eléctricos) será 0

Estas leyes se definen sobre redes dirigidas, así que entendemos a $\overrightarrow{E}$ como la versión dirigida del conjunto de ejes de G, después de haberles asignado alguna orientación que no tiene importancia: un valor de $x(\overrightarrow{e}) > 0$ implica que la corriente fluye en la dirección de $\overrightarrow{e}$, mientras que un valor de $x(\overrightarrow{e}) < 0$ significa que la corriente fluye en dirección opuesta a $\overrightarrow{e}$ (donde $\overrightarrow{e} \in \overrightarrow{E}$).

Volvamos a *Current-Flow Betweenness*. En *Betweenness*, calculábamos la fracción de caminos mínimos que pasaban por un nodo v. Ahora, podremos definir la versión en redes eléctricas como la fracción de una unidad de corriente que pasa por v:

$$c_{CB} = \sum_{s,t \in V} t_{st}(v) \tag{3.12}$$

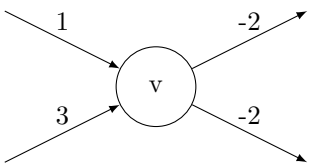
para todo $v \in V$, y donde t_{st} es el *throughput*, definido de la siguiente manera:

$$t_{st}(v) = \frac{t'_{st}(v)}{2} \tag{3.13}$$

$$t'_{st}(v) = -|b_{st}(v)| + \sum_{e : v \in e} |x(\overrightarrow{e})| \tag{3.14}$$

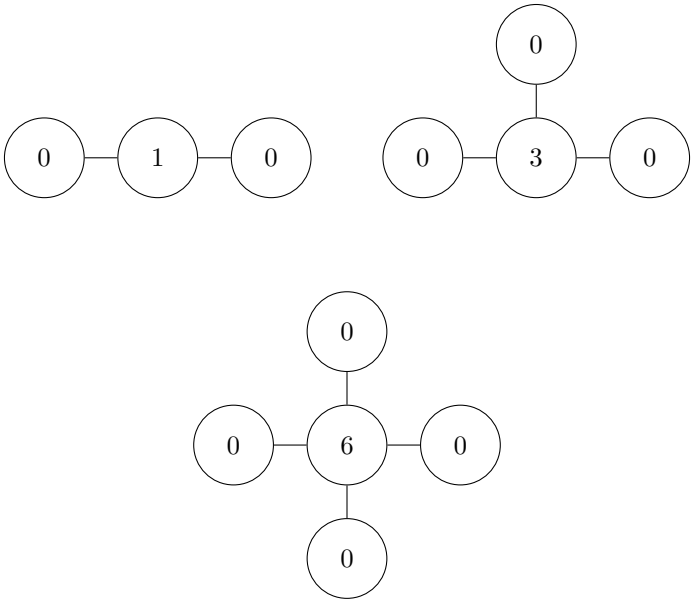
Intuitivamente, t representa *cuánto* flujo eléctrico pasa por v, sin tener en cuenta las unidades de corriente externas que salen o entran de la red a través de v (en caso de que $v = s$ o $v = t$), por lo cual se resta el valor de $b_{st}(v)$ en t'. Análogamente a *betweenness*, c_{CB} puede ser dividida por $n_B = (n-1)(n-2)$ para obtener una versión normalizada: esto se debe al caso en el que el grafo sea completo, y considerando que se enviará una única unidad de corriente para cada par (s, t), con lo cual nunca será posible que la suma de *throughputs* para un nodo sea mayor a n_B.

t' se divide a la mitad porque la sumatoria representa todo el flujo eléctrico que entra y que sale de v: por la primera ley de Kirchhoff, todo el flujo que entra a v debe también salir. Como se toma el módulo de $x(\overrightarrow{e})$ para todos los ejes que contienen a v, los flujos que entran y salen de v terminan duplicándose en lugar de cancelándose. Veamos el siguiente ejemplo:



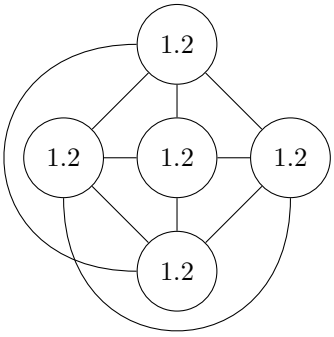
En este caso, $\sum_{e:v\in e} |x(\overrightarrow{e})| = 8$, ya que la conectividad de los dos ejes que llegan a v es en total 4, y la conectividad de los ejes que salen de v es -4. Como hemos dicho antes, dado que se cumple la ley de corrientes, t' siempre será el doble del *throughput* de v, y por eso será necesario dividir a la mitad.

Para los ejemplos simples usados para *Betweenness*, veremos que los valores de centralidad son los mismos para *Current-flow Betweenness*:



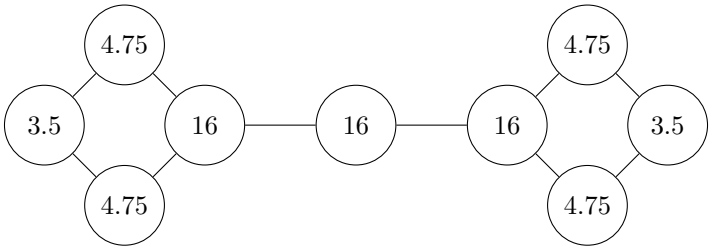
Por razones similares a *Betweenness*, únicamente el nodo central tendrá un valor no nulo porque la corriente eléctrica deberá pasar, sin excepciones, por él. También es importante notar que todos los nodos que son hojas, como corolario de la *ley de corrientes*, tendrán un valor de c_{CB} nulo, ya que no es posible que un valor de corriente distinto a cero entre y salga del nodo, si el grado de dicho nodo es 1.

En el caso de un grafo completo, ya no obtendremos siempre valores nulos como para *betweenness*:



Esto se debe a que ya no estamos viendo la proporción de caminos mínimos que deben pasar por un nodo v, sino que estamos viendo el *throughput* de información o corriente eléctrica que va a pasar por v. La información se moverá por toda la red, y si todos los nodos están conectados, llegará a todos de igual manera.

Finalmente, veamos el ejemplo con el nodo puente:



Esta vez, el nodo con mayor centralidad no es únicamente el puente, sino los dos nodos directamente adyacentes al puente. Esto se debe a que, para poder pasar por el nodo puente, la corriente eléctrica *debe* pasar por los nodos adyacentes. En términos de *throughput* entonces, los tres nodos serán iguales.

En resúmen, habiendo tomado información como flujo eléctrico, esta medida de centralidad permite ver cuáles son los nodos que mayor *throughput* de información tienen en la red.

3.2.8. Information Centrality

También conocida como *Current-Flow Closeness Centrality*, es (al igual que *Current-Flow Betweenness*) una modificación de su análoga *Closeness Centrality* basada en caminos mínimos, pero considerando la red como una red eléctrica, y tomando como medida de distancia la resistencia efectiva en lugar de la distancia geodésica:

$$c_{CC}(s) = \frac{1}{\sum_{t \neq s} p_{st}(s) - p_{st}(t)} \tag{3.15}$$

donde $p_{st}(s) - p_{st}(t)$ es la resistencia efectiva entre s y t.

La resistencia efectiva es una medida de distancia entre dos nodos. Esto significa que podemos interpretar a esta centralidad como "*Closeness*, pero utilizando el equivalente a distancia mínima para redes eléctricas". Veamos más en detalle qué es la resistencia efectiva y de dónde sale.

En un grafo cualquiera, solemos definir a la distancia entre dos nodos como igual a la suma del "peso" de los ejes que los conectan. En redes eléctricas, el peso asociado a un eje se entiende como la *conductividad* del eje ($c(e)$), es decir, cuánta corriente eléctrica es capaz de pasar por ese eje. Por definición, la conductividad equivale a $\frac{1}{r(e)}$, donde $r(e)$ es la *resistencia* del eje, es decir, la medida en la cual el eje se opone al flujo de una corriente eléctrica. Con un simple reemplazo, la resistencia se puede definir en términos de conductividad como $\frac{1}{c(e)}$.

Otra definición necesaria es la de *energía potencial*: es la energía que posee un objeto simplemente como consecuencia de su posición en el espacio (respecto a otro objeto). La *diferencia potencial* es una magnitud de cuánto trabajo es necesario para mover un objeto desde un punto a otro (en este caso, entre nodos). Una corriente eléctrica se relaciona con la diferencia de potencial a través de la *Ley de Ohm*:

$$V = I \cdot R \tag{3.16}$$

donde V corresponde a la diferencia de potencial aplicada, I es la corriente eléctrica que circula y R es la resistencia del circuito. Aplicando esta Ley al contexto de redes eléctricas, podemos reescribirla como:

$$V(e) = I(e) \cdot R(e) \quad \forall e \in E \tag{3.17}$$

para referirnos a que la diferencia de potencial de un eje estará definido por la corriente transmitida a través de mismo y la resistencia que se opone.

Cuanto más fuerte sea la corriente, o cuanto mayor sea la resistencia impuesta por el canal, mayor será el trabajo a realizar (el potencial a aplicar) para poder mover a la unidad de carga a través de e. Para ilustrar el concepto, podemos momentáneamente hacer una analogía con una bomba de agua

moviendo agua a través de tuberías: la corriente será el caudal de agua y la resistencia el diámetro de la tubería (a menor diámetro, mayor resistencia). El trabajo que deberá hacer la bomba será mucho mayor en la medida que se requiera un caudal mayor y/o a que la tubería se vuelva más estrecha.

Como dijimos anteriormente en esta sección, también sabemos que:

$$I(e) = x(\overrightarrow{e}) \tag{3.18}$$

$$R = \frac{1}{c(e)} \tag{3.19}$$

Llamando $\hat{p}$ a V, tenemos entonces:

$$\hat{p}(\overrightarrow{e}) = \frac{x(\overrightarrow{e})}{c(e)} \tag{3.20}$$

Diremos que una función $p : V \to \mathbb{R}$ asigna *potenciales absolutos* si:

$$\hat{p}(\overrightarrow{e}) = \hat{p}(v, w) = p(v) - p(w) \quad \forall (v, w) \in \overrightarrow{E} \tag{3.21}$$

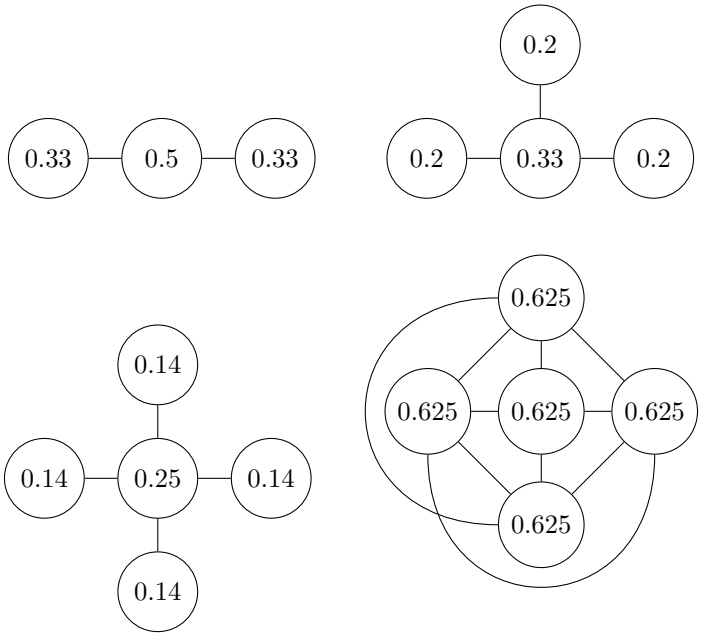
Intuitivamente, el potencial absoluto de un objeto es, al igual que la diferencia potencial, una medida de cuánto trabajo es necesario para mover una partícula de un punto a otro, pero fijando el punto de referencia en infinito. De esta forma, podemos comparar la energía potencial de dos objetos desde el mismo punto de referencia (ya que infinito siempre estará a la misma distancia de cualquier objeto), y usar estos valores para saber qué tanto "dista" un punto de otro.

Dada una red eléctrica $n = (G; c)$ con un vector de suministros b_{st}, sabemos que p existe, y es posible encontrarla utilizando la matriz Laplaciana (la demostración completa de este postulado puede encontrarse en el trabajo de Brandes y Fleischer [34]). Si llamamos p_{st} a los potenciales absolutos para ese b en particular, esta diferencia de potenciales absolutos se conoce como la **resistencia efectiva**. En otras palabras, la resistencia efectiva entre dos puntos nos dice qué tan alejados están dos puntos en términos de energía potencial.

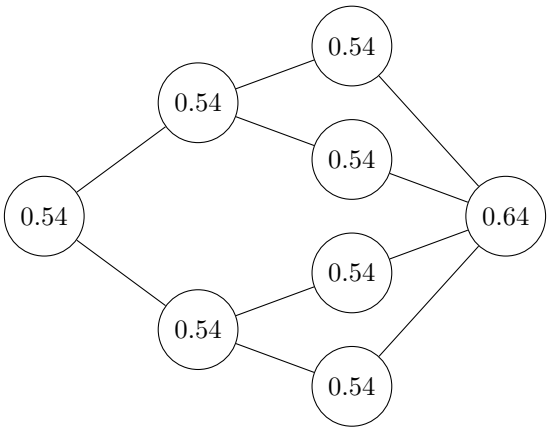
En resumen, la única modificación hecha con respecto a *Closeness Centrality* fue cambiar la definición de distancia mediante caminos mínimos a su análogo en redes eléctricas, ya que cuanto más pequeña la resistencia efectiva, más corta será la distancia entre dos nodos.

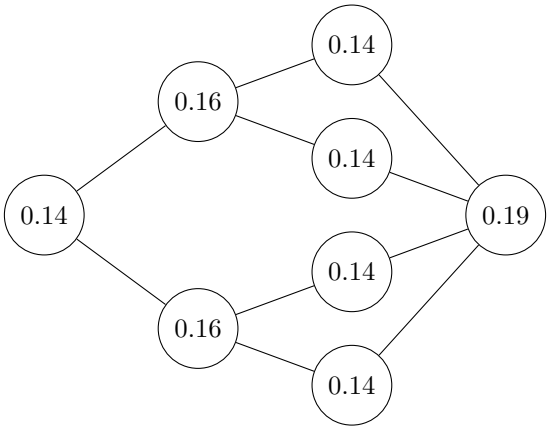
Information Centrality surgió como una respuesta a las críticas hechas a *Closeness Centrality*: las medidas que únicamente tienen en cuenta la distancia mínima entre nodos no tienen en cuenta el flujo de la información a través de caminos que no son mínimos. Al igual que con *Current Flow Betweenness*, la idea de esta medida de centralidad es considerar que la información en realidad se mueve por la red como corriente en una red eléctrica.

Si calculamos esta nueva medida para los grafos utilizados como ejemplos para *Closeness Centrality*, podemos ver que en general, las diferencias entre nodos se mantienen iguales, pero los valores absolutos son menores, dejando en evidencia que la información se reparte entre todos los posibles ejes y no únicamente los involucrados en los caminos mínimos.



Las diferencias pueden apreciarse más cuando tenemos un grafo con múltiples caminos entre nodos, no todos de la misma distancia. En ese caso, *Information Centrality* tendrá en cuenta el *spread* de la información a través de todos los ejes, y no únicamente de los que son parte de los caminos mínimos. Para ilustrar esto utilizaremos el mismo grafo usado como ejemplo en *Load Centrality*. La primera figura corresponde a los valores de centralidad para *Closeness*, y los de la segunda, para *Current-flow Closeness*:





Como se puede ver, en el primer caso todos los nodos excepto el de mayor grado tienen el mismo valor de centralidad, mientras que en el segundo, los valores parecen estar distribuidos en relación a qué tan conectados están los nodos: por ejemplo, el nodo con mayor valor es el de mayor grado, le siguen los dos nodos de grado 3 a la izquierda del grafo, y el resto (de grado 2) tiene el menor valor de entre todos los nodos. Esto tiene que ver con que, como el flujo se reparte por todos los ejes en un flujo eléctrico, los nodos con mayor adyacencia, recibirán más flujo de otros nodos.

3.2.9. Eigenvector Centrality

La centralidad por autovectores es una medida de influencia de un nodo en una red. A cada nodo se le asigna un puntaje relativo a qué tan influyentes son sus vecinos en la red. Un nodo conectado a muchos otros con bajo puntaje no va a tener un valor de centralidad alto, mientras que uno conectado a otros nodos con valores altos de centralidad por autovectores tendrán a la vez valores propios altos de centralidad.

Estas ideas son utilizadas, por ejemplo, por el algoritmo de PageRank. Los nodos con mayor puntaje pueden pensarse como los que brindan o tienen acceso a la mejor información en la red.

Sea A la matriz de adyacencia de un grafo G, donde $a_{p,q} = 1$ si existe un eje entre p y q y 0 sino. Matemáticamente, *Eigenvector Centrality* se define como:

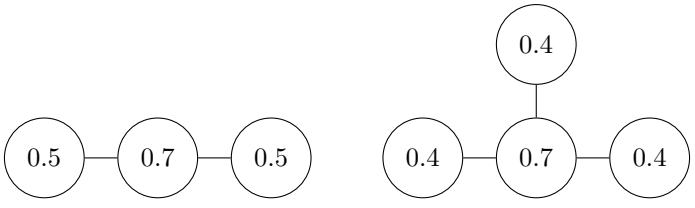
$$C_E(p) = x_p = \frac{1}{\lambda} \sum_{q \in vecinos(p)} x_q = \frac{1}{\lambda} \sum_{q \in G} a_{p,q} x_q \tag{3.22}$$

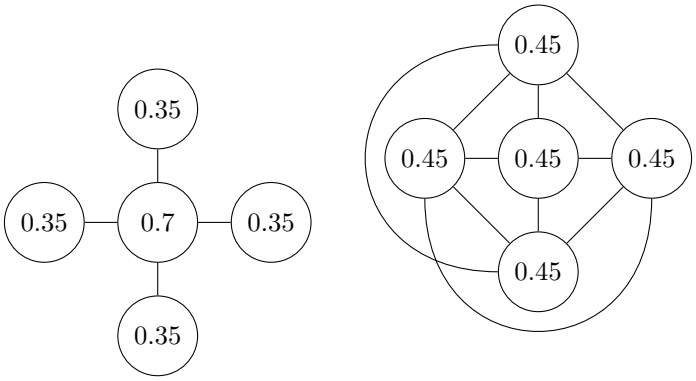
Que equivale, en notación vectorial, a la ecuación de autovectores:

$$\mathbf{Ax} = \lambda \mathbf{x} \tag{3.23}$$

Como la matriz de adyacencia tiene sólo números positivos, existe un único autovalor máximo λ', y los valores de centralidad buscados estarán dados por el autovector asociado a λ'.

Veamos los mismos ejemplos que hemos visto para las centralidades anteriores:





Como en los otros casos, el nodo central es siempre el de mayor centralidad, excepto cuando todos los nodos son adyacentes.

Capítulo 4

Datos

En este capítulo explicaremos cómo se obtuvieron los datos analizados en esta tesis y cómo fueron procesados: su orígen, la forma de almacenarlos, cómo se limpió el dataset, e imprevistos que surgieron.

4.1. Selección del tipo de base de datos

Podemos dividir a las bases de datos en dos grandes grupos: relacionales, y no relacionales. Las relacionales se parecen a una hoja de cálculo, donde cada entrada es una fila con distintas columnas, organizadas en distintas tablas. Por otro lado, las no relacionales no fuerzan una organización tan estricta sobre los datos. Podemos pensarlas, en lugar de como tablas en una hoja de cálculo, como directorios en un *file system*, donde cada entrada es un archivo distinto.

Hay distintos tipos de bases de datos no relacionales, cada una utilizando distintos conceptos para mantener organizada la información. Uno de ellos (y el que nos interesará en este trabajo) son las llamadas *bases de datos basadas en grafos*. En este caso, la base de datos en su totalidad será interpretada como un grafo. Cada entrada será un nodo con distintas propiedades, y los nodos estarán interconectados mediante ejes, que a su vez pueden también tener propiedades.

Es más natural estructurar la información como un grafo cuando los datos describen redes. En nuestro caso es fácil ver cómo esto aplica: podemos interpretar a los jugadores como nodos, unidos por un eje cada vez que hayan jugado una partida. Es también más natural hacer consultas en este escenario, dado que, por ejemplo, podemos pedir todos los vecinos de un nodo para saber inmediatamente todas las personas con las que ha jugado partidas.

Una desventaja de esta mayor expresividad (comparándolas con una base de datos relacional estándar), es que estas bases de datos suelen ocupar más espacio, y no ser tan eficientes a la hora de hacer consultas complejas o sobre un gran dataset. Dicho esto, finalmente decidimos utilizar una base de datos basada en grafos en este trabajo.

4.2. Scraping inicial del sitio web OGS

En primera instancia, para conseguir los datos de las partidas, utilizamos un script (escrito en `Python`) para visitar todas las páginas de partidas del sitio y parsearlas para extraer información de cada una que sería guardada en un archivo de texto.

Con esto, conseguimos cerca de un millón de archivos, aunque muchos de ellos contenían errores o se referían a partidas inexistentes. Hubo un proceso arduo de limpieza, reconstrucción y re-obtención de datos que siguió durante bastante tiempo luego de comenzada esta tesis. Los puntos principales serán

mencionados a lo largo de la presente sección.

Una vez filtrados los archivos que no contenían información de partidas (conteniendo en su lugar, por ejemplo, errores `HTTP 404` de sitio web no encontrado, como consecuencia de la forma en que se hizo el scraping), finalmente estábamos en condiciones de almacenar estos datos de una forma que permitiera un análisis más profundo.

4.3. Creación de la base de datos

Como fue adelantado en la sección 4.1, dado el enfoque de este trabajo y la naturaleza de los datos, decidimos que la mejor forma de almacenar y consultar los datos era en forma de grafo. Por esta razón decidimos usar una base de datos no relacional basada en grafos: **Neo4j**.

`Neo4j`, al igual que otras implementaciones de bases de datos basadas en grafos, permite guardar nodos de distintos tipos, unidos mediante distintas relaciones (ejes). Como todos los scripts creados a lo largo del desarrollo de esta tesis, utilizamos `Python` para escribir un programa que parseara los archivos filtrados de datos crudos y los insertara en la base de datos. Este proceso fue lento, dado que los datos eran muchos y la inserción en este tipo de bases de datos no siempre es tan rápida.

4.3.1. Esquema de la base de datos

Las bases de datos basadas en grafos permiten definir distintos tipos de nodos, conectados mediante relaciones, que también puede ser nombradas con distintas etiquetas. En nuestro caso, decidimos tener dos tipos de nodos: **jugadores**, y **partidas**. Algunas de las propiedades para cada tipo de nodo son:

- **Jugador**
  - ogsId: el identificador del jugador en OGS
- **Partida**
  - ogsId: el identificador de la partida en OGS
  - startDate: la fecha y hora de comienzo de la partida
  - endDate: la fecha y hora de finalización de la partida
  - blackPlayer / whitePlayer: el nombre de usuario de cada jugador
  - blackGlicko / whiteGlicko: el valor de Glicko de cada jugador al momento de comenzar la partida
  - outcome: el resultado de la partida - hubo un ganador, fue empate, alguno de los jugadores abandonó, o hubo timeout
  - winner: el nombre del jugador ganador, si lo hubo

Cada nodo de tipo jugador estará conectado únicamente a nodos de tipo partida, y cada nodo de tipo partida estará conectado a exactamente dos nodos distintos de tipo jugador.

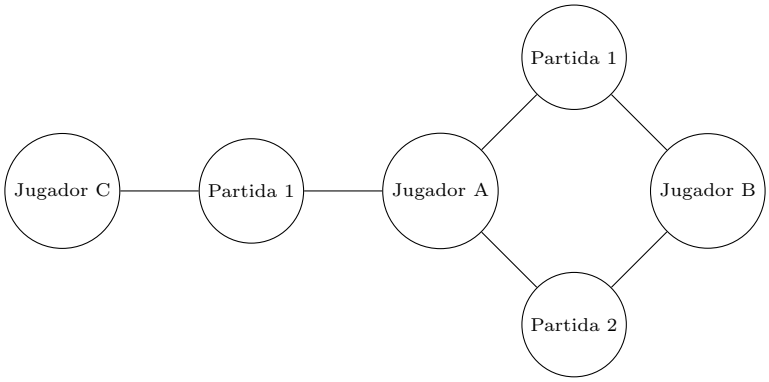


Figura 4.1: Un ejemplo ilustrando cómo está modelada la red de partidas y jugadores en `Neo4j`. Notar que se trata de un grafo bipartito, donde uno de los conjuntos de nodos será únicamente el subconjunto de las partidas, y el otro el de los jugadores.

Tomamos la decisión de modelar las partidas como nodos (en lugar de tener un multigrafo únicamente con nodos de tipo jugador) porque así sería más fácil y rápido poder almacenar y hacer consultas sobre la información de las partidas. `Neo4j` no permite índices sobre propiedades de ejes, y hacer consultas sobre propiedades en los ejes también es bastante limitado.

4.4. Limpieza y descripción del dataset

Inicialmente, la base de datos original contaba con más de 7 millones de jugadores y casi 15 millones de partidas entre Noviembre del 2005 y Julio de 2018.

4.4.1. Bots, el primer imprevisto

Al comenzar a hacer unas primeras exploraciones sobre los datos, notamos que algunos jugadores tenían una cantidad de partidas jugadas extremadamente alta en períodos cortos de tiempos (cientos en una semana). Inicialmente creíamos que tal vez eran jugadores muy activos, pero al buscar los perfiles de algunos de ellos en el sitio web de OGS, confirmamos que en realidad eran bots: "jugadores" artificiales (controlados por una computadora) contra los que otros individuos pueden jugar para practicar.

En una primera instancia, intentamos ver si estos jugadores eran identificables por alguna característica que ya estuviera presente en la base de datos, como por ejemplo algún prefijo en el nombre de usuario. Sin embargo, la respuesta fue tristemente no. La solución fue escribir un nuevo script en `Python` que hiciera una llamada al *endpoint* de la API de OGS usado para cargar la información del perfil de cada usuario. En el *payload* de la respuesta, una de las propiedades marcaba si el jugador era o no un bot. Con esto, conseguimos una lista de todos los jugadores que eran bots, y agregamos una propiedad booleana `isBot` a cada jugador en la base de datos, para poder filtrarlos.

4.4.2. Fusión con otro sitio web, el segundo imprevisto

Luego de un análisis más en detalle de los datos (que será explicado en más profundidad en la sección 5.1), nos encontramos con que en octubre del 2013, OGS había decidido fusionarse con otro importante sitio de partidas de Go online: *Nova*. Esto, similar a fenómenos poblacionales como migraciones, generó un cambio rotundo en la red de individuos.

Para no lidiar con estos efectos, decidimos únicamente trabajar con datos hasta Octubre de 2013,

es decir, previos a la fusión de ambos sitios. Inesperadamente, aprendimos que los bots habían surgido como parte de la fusión, o posteriormente a ella, con lo cual ningún jugador creado antes de Octubre de 2013 era un bot.

Una vez eliminadas todas las partidas posteriores al período que habíamos calificado de interés (y por consecuente habiendo removido todos los bots), finalmente obtuvimos el dataset definitivo: 445.773 jugadores y 818.000 partidas, entre Noviembre del 2005 y Octubre de 2013. A pesar de ser mucho menor en cardinalidad que el original, sigue siendo de un tamaño considerable y representativo.

4.4.3. Estimaciones erróneas, el tercer imprevisto

Al comenzar a estudiar la evolución de los valores de habilidad para los jugadores, notamos otra inconsistencia: por largos períodos de tiempo (a veces para *todas* las partidas de un jugador), la habilidad era la misma.

Revisando en detalle el sitio web de OGS y sus APIs, pudimos confirmar que el valor de ranking para cada jugador que figura en la página de la partida es en realidad el ranking actual, y no el del momento de la partida. Esto significaba que todos los valores de habilidad que teníamos en ese instante en la base de datos eran en realidad incorrectos.

Afortunadamente, la API proveía la información de cuál era la habilidad de cada jugador justo antes de cada partida, por lo que otro script fue escrito en el cual descargamos esta información para cada partida de la base de datos, y la actualizamos. Con este último cambio, tuvimos confianza en que nuestro dataset ahora estaba finalmente "limpio".

4.5. Procesamiento de los datos

Las bases de datos basadas en grafos brindan una forma intuitiva de ver los datos y de trabajar con ellos de manera simple. Sin embargo, es complicado o imposible escribir ciertas consultas (por ejemplo, calcular centralidades), y pueden ser muy lentas para otras (como encontrar caminos mínimos). Por esta razón, decidimos exportar los datos de la base de datos a un formato que pudiera ser interpretado por alguna biblioteca en `Python`, y poder hacer análisis más profundos.

`Neo4j` permite exportar ciertas consultas en formato `graphml`. Este formato es uno de los estándares para describir grafos, y es interpretado por distintos programas y bibliotecas. Una de esas bibliotecas es `NetworkX` de `Python`. `NetworkX` ofrece una API simple pero con muchísima funcionalidad a la hora de trabajar sobre grafos: brinda métodos para manipularlos fácilmente, y para calcular algoritmos complejos sobre los mismos, como centralidades, *clustering*, caminos mínimos, etc.

Como adelantamos en la sección 3.1, utilizamos el mecanismo de *snapshots* para hacer el análisis sobre este dataset. Tomamos la decisión de crear un script que haga consultas a la base de datos utilizando el SDK de `Neo4j` para descargar el dataset en formato `graphml`. Inicialmente, decidimos que cada consulta se corresponda con una *snapshot* o ventana de dos semanas de largo, con una semana de solapamiento entre ventana. Como resultado, de cada una se obtiene un archivo `graphml` con todas las partidas jugadas en la ventana correspondiente y los nodos de los jugadores involucrados. En efecto, cada ventana será el subgrafo del dataset completo en el cual todos los nodos de tipo partida hayan tenido fecha de comienzo dentro de la ventana consultada.

A la hora de hacer varios de los análisis, hicimos modificaciones a los subgrafos generados a partir de cada ventana. En particular, para calcular las centralidades para los jugadores, transformamos los grafos obtenidos de `Neo4j` para que solo contengan nodos de tipo jugador.

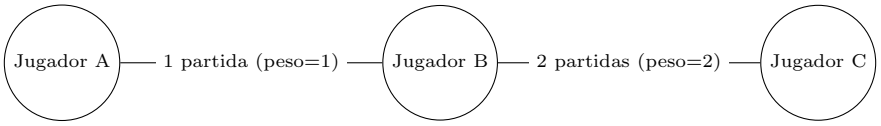


Figura 4.2: El mismo ejemplo de la figura 4.1 pero con la transformación para obtener un grafo únicamente con nodos de jugadores y ejes pesados.

Los nodos de tipo partida fueron reemplazados por un único eje entre los dos jugadores que unían, asignando un peso al eje igual a la cantidad de partidas jugadas entre ambos. Este peso fue utilizado como el peso de los ejes en las medidas de centralidad que lo incorporan en sus definiciones.

A lo largo del desarrollo de la tesis, utilizamos otros tamaños de ventana (dos semanas sin solapamiento, o de una semana de largo por ejemplo) para validar la elección de la ventana temporal utilizada en el análisis. Estos resultados se presentarán en la siguiente sección.

Por último, es necesario aclarar que las fechas tomadas para decidir a qué ventana corresponde cada partida son las fechas de finalización de las mismas, y no las de comienzo. Esto es porque la actualización de la habilidad se realiza al finalizar la partida, y no al comenzar.

Capítulo 5

Resultados

En este capítulo presentaremos los resultados obtenidos a partir del análisis de los datos. Comenzaremos con una breve descripción del dataset. Presentaremos algunas características de la red de jugadores que permiten entender cómo evolucionan ciertos aspectos a lo largo del tiempo, como la distribución de partidas jugadas o la evolución del coeficiente de clustering. También analizaremos las curvas de aprendizaje para poblaciones con diferentes niveles de "actividad" (es decir, con alguna cantidad de partidas jugadas como mínimo) de modo de caracterizar las curvas de aprendizaje por experiencia típica. Finalmente nos enfocaremos en los análisis que buscan determinar el efecto de la topología sobre el aprendizaje: ¿existe una conexión entre la centralidad de los jugadores y su habilidad? Para separar el efecto de otros factores, aplicamos criterios de filtrado, generando distintas curvas de aprendizaje poblacionales según los distintos casos estudiados.

5.1. Propiedades generales de la red

Al comenzar a trabajar en esta tesis, todo lo que sabíamos de la red de jugadores eran los atributos que teníamos para cada uno, la información para cada partida, y las fechas entre las cuales habían sucedido las mismas. Para poder tener una visión más concreta del dataset, comenzamos por analizar medidas y propiedades generales de la red, para tener una visión general que nos permitiera entender las características principales de la red.

Utilizamos el método de ventana deslizante explicado con anterioridad para calcular algunas medidas generales. Un resultado interesante, aunque esperado, de esta fase, fue notar (ver figura 5.1) que la distribución del grado o cantidad de partidas jugadas en la población seguía una distribución exponencial, con la mayoría de los jugadores teniendo pocas partidas en su historial por semana, y muy pocos teniendo más de 20.

Intentando generalizar, analizamos cuál sería la distribución de partidas para todos los jugadores, considerando el grafo de todo el período estudiado en lugar de por ventana. Aquí también pudimos ver (figura 5.2) que la población decaía cuanto más alto era el número de partidas, o en otras palabras, presenta una distribución exponencial.

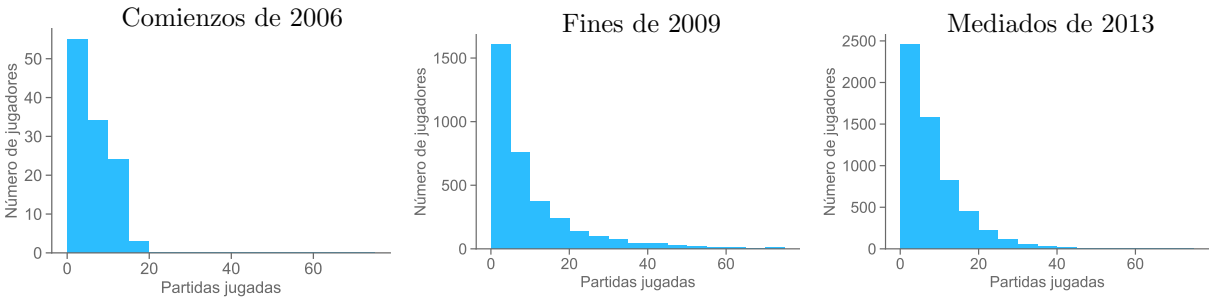


Figura 5.1: Histogramas de distribución de partidas jugadas por jugador en ventanas específicas. Cada histograma se corresponde con una ventana distinta: la primera es de una semana a comienzos de 2006, la segunda a fines de 2009, y la tercera a mediados de 2013. Los jugadores incluidos en cada histograma (y sus respectivas partidas) son únicamente los contenidos en esa ventana. Notar que la escala de los ejes de las ordenadas para cada histograma son distintas, ya que a medida que pasa el tiempo, aumenta el número de jugadores que participa y, por ende, la cantidad de jugadores en cada ventana.

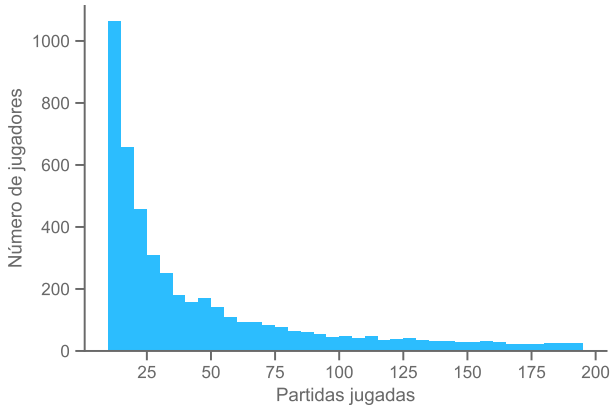


Figura 5.2: Cantidad de partidas por jugador considerando el grafo entero de jugadores, sin dividir por ventanas.

Además, hicimos un análisis temporal de la red. Para esto, nos basamos en un paper de Kossinets y Watts [31], que utiliza ventanas de diferente tamaño con el objetivo de ver su impacto en el análisis temporal. Tomamos ventanas de una semana de largo con dos días de solapamiento, de dos semanas de largo con una semana de solapamiento, y de un mes de largo con una semana de solapamiento, y graficamos las curvas para las mismas medidas tomadas de estas tres maneras. La figura 5.3 resume este análisis.

La primera sub-figura muestra la evolución del grado medio de los jugadores en cada ventana, es decir, la cantidad de partidas jugadas. Notemos cómo este valor varía dependiendo el tamaño de ventana elegido. Cuanto más grande es la ventana, mayor es la cantidad de partidas en promedio que un jugador habrá jugado en la misma y por lo tanto naturalmente el grado crece.

La segunda sub-figura muestra la fracción del grafo que representa la componente conexa más grande en cada ventana: en los casos más extremos, será 1 si hay una única componente en la ventana, y $\frac{1}{n}$ si hubiera n jugadores y nadie hubiera jugado con nadie. En este caso, como se puede observar, este valor casi no se ve afectado por el tamaño de ventana elegido. La única excepción a esto es en los primeros meses de existencia del sitio (donde no había tantos jugadores o partidas entre los nuevos), y en el breve decaimiento en conectividad que se ve a fines de 2010 (hablaremos de esto más adelante). A lo largo de las otras ventanas, siempre parece haber una componente conexa que contiene a la gran mayoría de los jugadores.

Figura 5.3: La primera figura muestra la evolución del grado medio de los jugadores en cada ventana. La segunda muestra la fracción del grafo que representa la componente conexa más grande en cada ventana. La tercera muestra el largo promedio de los caminos mínimos en la componente conexa más grande. La última figura muestra el coeficiente de clustering del subgrafo en cada ventana.

La tercera sub-figura muestra el largo promedio de los caminos mínimos en la componente conexa más grande. Ésto da una idea de qué tan interconectados se encuentran los jugadores: en general, todos se encuentran, en promedio, a aproximadamente cuatro grados de separación del resto.

Por último, la cuarta sub-figura muestra el coeficiente de *clustering* del subgrafo en cada ventana. El coeficiente de *clustering* es otra forma de cuantificar la densidad de un grafo: se basa en contar la cantidad de triángulos en el mismo. Un primer resultado general que podemos observar es que, cuanto más amplia es la ventana de tiempo elegida, menos ruido tendrá la curva correspondiente, mientras que cuanto más chica, más sujetas a ruido estarán.

Un claro ejemplo de esto último son los picos que podemos observar a fines del 2010. Los mismos son particularmente pronunciados en las curvas amarillas (para ventanas de una semana de duración), y en particular en el segundo gráfico, donde parecería haber una repentina desconexión entre los jugadores de la red. Al ver esto, intentamos buscar alguna explicación sobre por qué la conectividad del grafo podría haber caído tan repentinamente. Una primera explicación posible podría haber sido un gran incremento en la cantidad de usuarios registrados, pero sin que hubieran jugado muchas partidas. Con un poco de investigación, encontramos evidencia en un historial del estado del servidor de OGS durante esos años, de que el servidor sufrió problemas técnicos y estuvo caído por un par de semanas a fines de 2010, lo cual encaja a la perfección con la caída.

Inicialmente habíamos generado este mismo gráfico para todas las partidas de la base de datos, pero como hemos mencionado antes, decidimos acotar el análisis hecho en esta tesis a las partidas ocurridas hasta Octubre de 2013. Tomamos esta decisión a partir de una observación hecha sobre el gráfico que habíamos generado. En ese caso, a fines de 2013 se podía ver una fuerte caída (aún más pronunciada que la de 2010).

Buscando en los foros de la comunidad de OGS, encontramos que a fines de 2013 hubo una fusión entre OGS y otro portal de Go, Nova, como mencionamos en la sección 4.4. En términos topológicos, la unión de ambos sitios afectó al grafo porque un conjunto completamente nuevo y disjunto de vértices fue introducido en la red preexistente: llevó un tiempo hasta que ambas redes se conecten fuertemente.

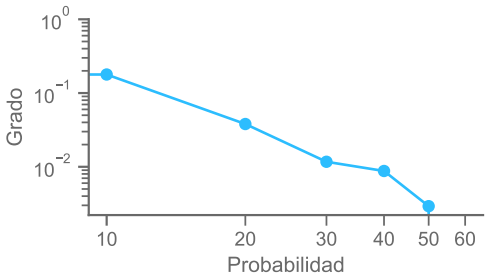
Como no estábamos interesados en analizar este fenómeno, que potencialmente podría ser objeto de análisis en sí mismo, decidimos, de aquí en adelante, trabajar con el dataset reducido desde los orígenes a finales de 2005, hasta antes de la unión en Octubre de 2013. Todos los ejes y nodos creados en una fecha posterior fueron descartados y no han sido tenido en cuenta para los experimentos (ni resultados) de las siguientes secciones.

Volviendo al análisis sobre el tamaño de las ventanas, las curvas resultantes de tomar ventanas más grandes no se ven tan afectadas por estos fenómentos que, a fin de cuentas, no son más que ruido. Es fácil entonces ver que una ventana tan pequeña no es recomendable para hacer este tipo de análisis.

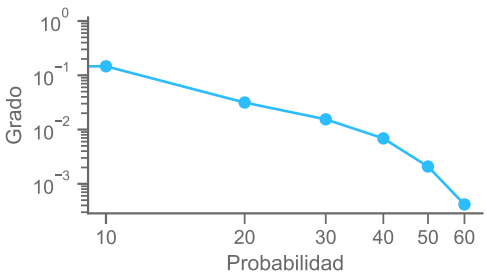
Por otro lado, una ventana de un mes de largo puede generar curvas que son demasiado suaves y no permiten observar algunos efectos interesantes. Por ejemplo, la curva verde en este caso no presenta ninguna caída durante Diciembre de 2010.

Finalmente, vimos que haber elegido ventanas de tamaño "medio" como finalmente hicimos, no parece tener un efecto negativo respecto a la proporción de la red que es parte de la componente conexa mayor, ni a la interconexión observada en la misma (gráficos 3 y 4 en la figura 5.3): independientemente del tamaño de ventana, estos valores se mantienen muy similares.

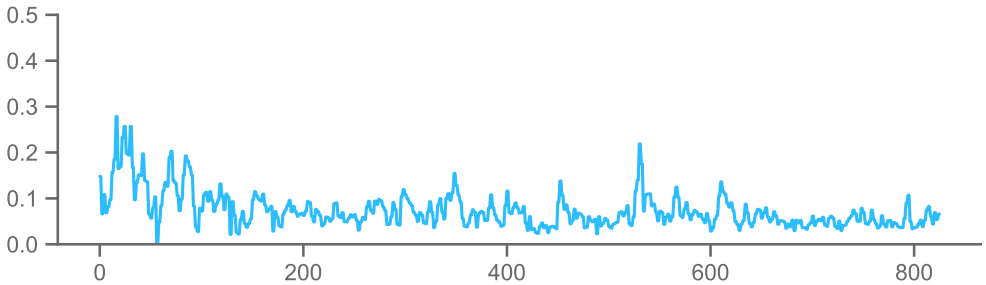
Posteriormente, decidimos hacer un segundo análisis presentado también en Kossinets y Watts [31]. Esta vez, la estrategia fue tomar una ventana como punto de referencia, para compararla con todas las otras y poder observar qué tan distintas o similares son entre ellas.



(a) Una semana a comienzos de 2006



(b) Una semana a mediados de 2013



(c) Estadística-D de Kolmogorov comparando distribución de 5.4a con las otras ventanas. Esta medida resume la distancia entre la distribución tomada como referencia, y otra distribución. En otras palabras, es un indicador de qué tan distintas son dos distribuciones: cuanto más cercano a 0, más parecidas serán, y cuanto más cercano a 1, más distintas.

Figura 5.4: Distribución media del grado por ventana. La ventana tomada como referencia es la de comienzos de 2006.

La figura 5.4a muestra la distribución del grado, en escala logarítmica, de los jugadores en una ventana a mediados de 2006. En 5.4b se encuentra la distribución del grado para una ventana a comienzos de 2013. Estos son casi los dos extremos (temporalmente hablando) en el subconjunto de datos que decidimos analizar. Podemos ver que ambas curvas presentan una forma similar, sugiriendo que los grados estarían distribuidos de manera relativamente similar, con la diferencia de que aunque hay una mayor población, habrá aún menos jugadores con altos grados.

Por último, la figura 5.4 muestra la *prueba de Kolmogorov-Smirnoff* (también conocida como *D-statistic*) comparando la distribución del grado en la ventana de referencia contra todas las otras ventanas. Cuanto menor sea este valor, mayor similitud habrá entre las distribuciones. Podemos ver que, a pesar de que hay bastantes oscilaciones, el valor suele mantenerse por debajo de 0,1, alcanzando valores cercanos a 0,2 solo en dos ocasiones. Esto significa que todas las ventanas tienen una distribución de nodos similar.

5.2. Curvas de aprendizaje poblacional

Ya con un conjunto reducido de partidas y jugadores a partir de la conclusión de 5.1, decidimos calcular la curva de aprendizaje poblacional para esa red (figura 5.5). Como se explicó, inicialmente contábamos con los valores de habilidad *Glicko*, y notamos que, en escala *log-log*, la curva de habilidad no era una función lineal de la experiencia.

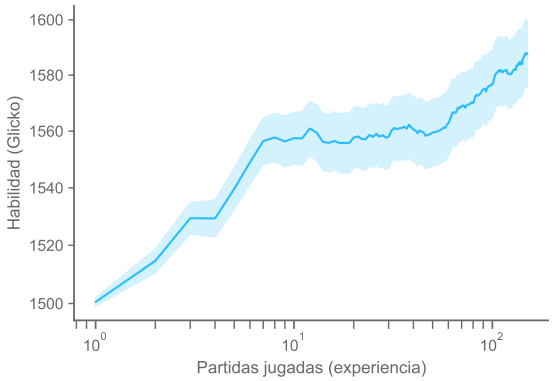


Figura 5.5: Curva de aprendizaje poblacional con los valores originales de Glicko brindados por OGS.

La primera suposición fue que, como diferentes poblaciones (o subgrupos) tienen diferentes curvas de aprendizaje [8], era posible que tener todos estos grupos mezclados estuviera generando curvas no representativas. Para evitar que la misma decaiga cuando crece el número de partidas, decidimos separar a la población en cohortes por cantidad de partidas jugadas, y graficar las curvas de aprendizaje para cada una de ellas (figura 5.6).

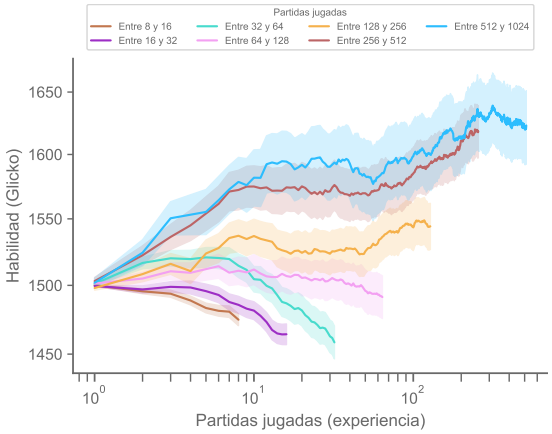


Figura 5.6: Curva de aprendizaje poblacional con los valores originales de Glicko brindados por OGS, separando a la población por cohortes en base a la cantidad de partidas jugadas.

Pero nuevamente, las curvas no eran exponenciales como esperábamos. Inicialmente, esto nos sorprendió, pero finalmente teorizamos que esto se debía a las partidas jugadas en privado o no rankeadas: los jugadores estaban aprendiendo, pero el sitio web no era capaz de capturar ese aprendizaje en la estimación de habilidad, generando una curva no exponencial, aunque en la realidad sí lo fuera.

El siguiente paso, entonces, fue estimar nosotros mismos la habilidad de cada jugador después de cada partida, teniendo en cuenta las partidas privadas o no rankeadas. Para esto, utilizamos *TrueSkill*, y posteriormente *TrueSkill Through Time*.

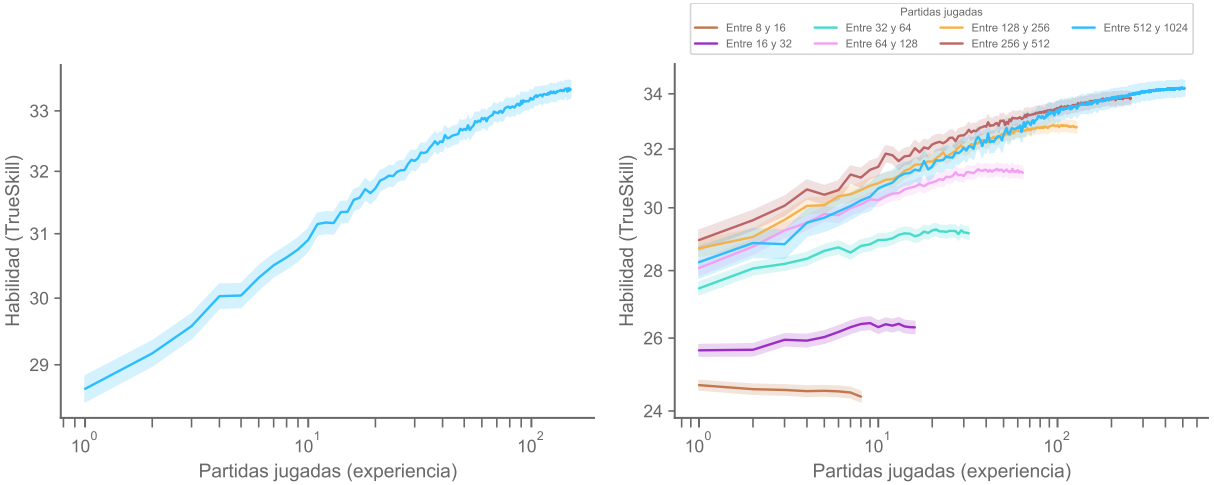


Figura 5.7: Figuras análogas a 5.5 y 5.6 pero utilizando TrueSkill en lugar de Glicko.

Como se puede observar en la figura 5.7, las curvas resultantes sí presentan una relación exponencial, lo cual se condiciona con lo mencionado en 1.3.3. Vale destacar dos observaciones sobre estos gráficos:

- La primera, es que, una vez que los jugadores tienen más de 64 partidas, parecería que las curvas para las cohortes correspondientes no se diferencian mucho entre sí. A partir de este valor de experiencia máxima, se elimina el sesgo de baja habilidad características de las cohortes de jugadores que abandona rápidamente el juego.
- La segunda observación es que cerca del final de cada curva, se pueden ver unas leves caídas. Esto probablemente se deba a que nos estamos acercando a las últimas partidas de los jugadores en

esa cohorte. El hecho de que sea de las últimas partidas significa que el jugador probablemente se encuentre desganado o desmotivado por seguir jugando, y entonces no se enfoca en participar tan activamente.

5.3. Relación entre el aprendizaje y la topología

Tal como se presentó inicialmente, el objetivo de esta tesis es estudiar (si existiera) el efecto de la topología de la red de jugadores sobre el aprendizaje de los mismos a lo largo del tiempo. Idealmente, querríamos poder graficar distintas curvas de aprendizaje para distintas cohortes de la población, calculadas a partir de características topológicas de la red. Para lograr esto, para cada ventana de *sliding window*, calculamos las distintas medidas de centralidad introducidas en la sección 3.2 para cada jugador:

- Information centrality
- Degree centrality
- Load centrality
- Eigenvector centrality
- Closeness centrality
- Betweenness centrality
- Current Flow Betweenness
- Harmonic centrality

Al graficar los histogramas para cada centralidad, incluyendo únicamente a los jugadores con al menos 150 partidas jugadas, podemos observar que la mayoría presentan una distribución exponencial: muy pocos jugadores tienen altos valores de centralidad, mientras que la mayoría tienen valores bajos. Para mostrar mejor la distribución poblacional, decidimos mostrarlos graficando (en la figura 5.8) el eje abscisas en escala logarítmica, para obtener así distribuciones normales.

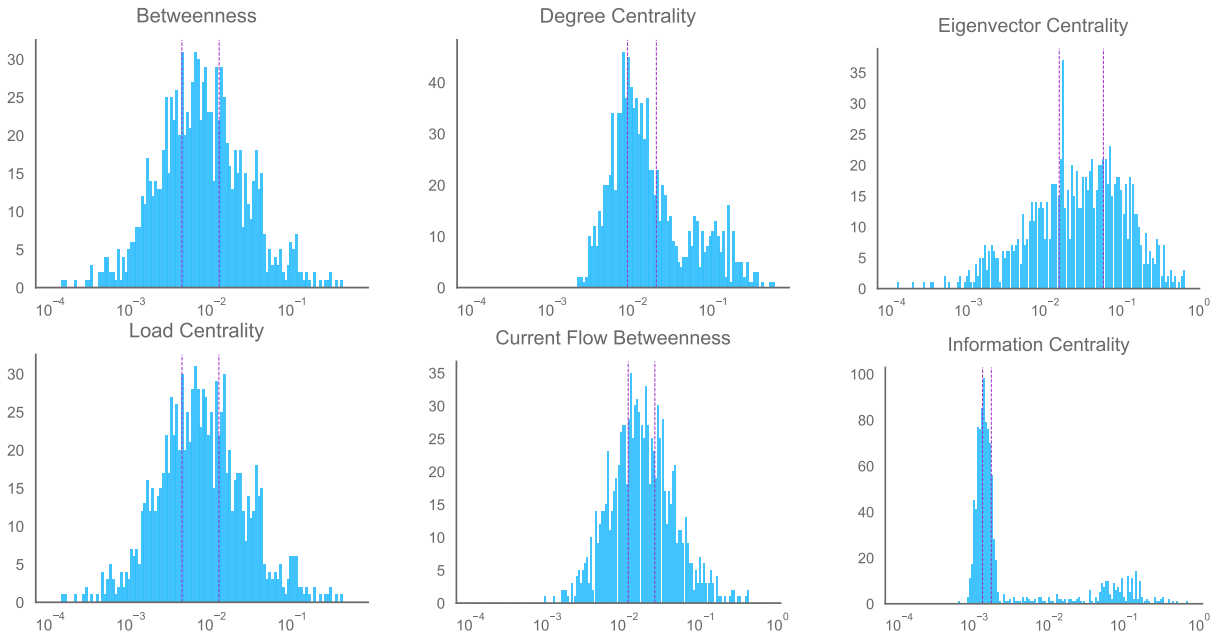
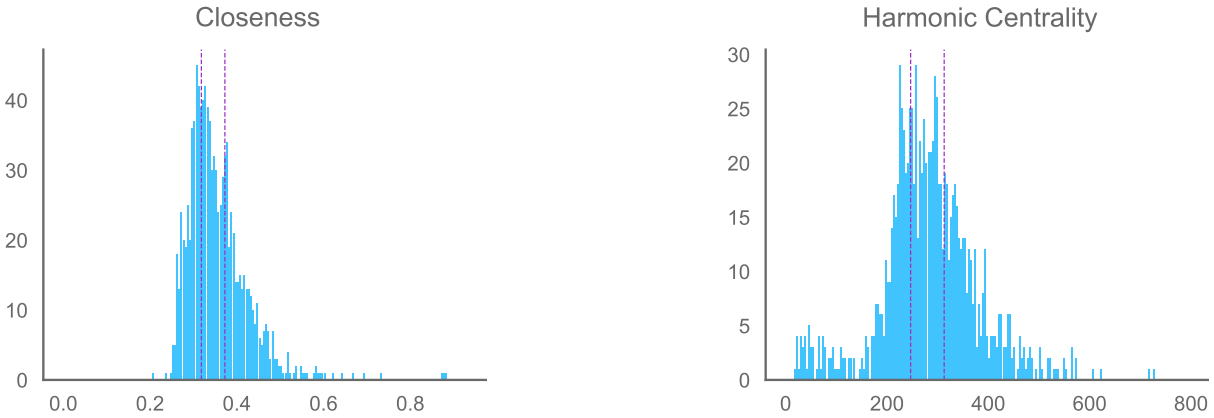


Figura 5.8: Histogramas mostrando la distribución de los valores medios de centralidad, teniendo en cuenta únicamente las primeras 50 partidas. Las líneas punteadas muestran los percentiles utilizados para las divisiones en cohortes de centralidad baja, media y alta. La primera línea se corresponde con el 33avo percentil, y la segunda con el 66avo.

Dos de las centralidades presentan una distribución Gaussiana sin necesidad de graficar los ejes en escala logarítmica: closeness y harmonic.



Cuando calculamos las centralidades, obtenemos una lista para cada centralidad C, para cada jugador J. Cada elemento i en la lista representa la centralidad C del jugador J en la ventana i en la que el jugador estuvo presente. De forma más intuitiva, cada lista representa la evolución a lo largo del tiempo de la centralidad C para el jugador J.

Como nos interesa ver la evolución de la habilidad en función de la cantidad de partidas jugadas (la experiencia), pero la centralidad es una medida que tiene sentido por ventana temporal y no en un instante (ya que no sería de utilidad ver únicamente el subgrafo correspondiente a una partida entre dos jugadores), en principio un jugador puede haber jugado múltiples partidas en una ventana, pero sólo tendrá un valor de centralidad para esa ventana. Para poder mantener la relación de *experiencia vs. habilidad* en relación a las partidas jugadas y no a las ventanas, las centralidades de la ventana i para el jugador J estarán repetidas en cada lista tantas veces como partidas haya jugado el jugador J en la ventana i.

El siguiente paso es clasificar a cada jugador dentro de alguna cohorte definida en base a estos valores de centralidad. Para clasificar a cada jugador, necesitamos alguna manera de resumir la evolución de cada centralidad. Probamos tomando la media, la mediana, el máximo valor, y el último valor. Luego de comparar las curvas con cada una de ellas, concluímos que la media era la mejor forma de resumir ya que separaba mejor las curvas.

Por último resta definir las cohortes en base a condiciones respecto del valor resumen elegido. Decidimos simplemente separar en percentiles para quedarnos con tres grupos: centralidad baja, centralidad media, y centralidad alta. Todos los jugadores cuya media esté por debajo del percentil 33 serán clasificados como centralidad baja, los que estén entre el 33avo y el 66avo, con centralidad media, y por encima del 66avo, con centralidad alta.

Hipótesis Nuestra suposición al decidir separar a la población en cohortes definidas por la centralidad de los individuos era que íbamos a ver curvas marcadamente distintas para cada grupo:

- El grupo de alta centralidad, suponíamos que iba a tener una habilidad por encima de los otros grupos
- A su vez, el grupo de centralidad media iba a estar por encima del grupo de centralidad baja

Esto, conjeturamos, tenía sentido dado que un jugador con mayor centralidad ocupa un lugar de mayor importancia en la red. Desde el punto de vista de transmisión de información cultural, cuanto más central es el lugar que ocupa el jugador, mayor información transmite a los otros jugadores. Si

un jugador es muy central, entonces, debe ser porque tiene mucho que ofrecer al resto y, por ende, su habilidad sería más alta.

Sorprendentemente, la realidad fue completamente opuesta a nuestra hipótesis original. La figura 5.9 muestra la evolución de la habilidad con TrueSkill para distintas cohortes agrupadas de acuerdo a diferentes medidas de centralidad obtenidas. Como se puede observar, las curvas de baja centralidad siempre se encuentran por encima de la curva de centralidad media, la que, a su vez, se encuentra por encima de la curva para la cohorte de centralidad alta. En otras palabras, parece ser que *una centralidad baja se relaciona con una mayor habilidad.*

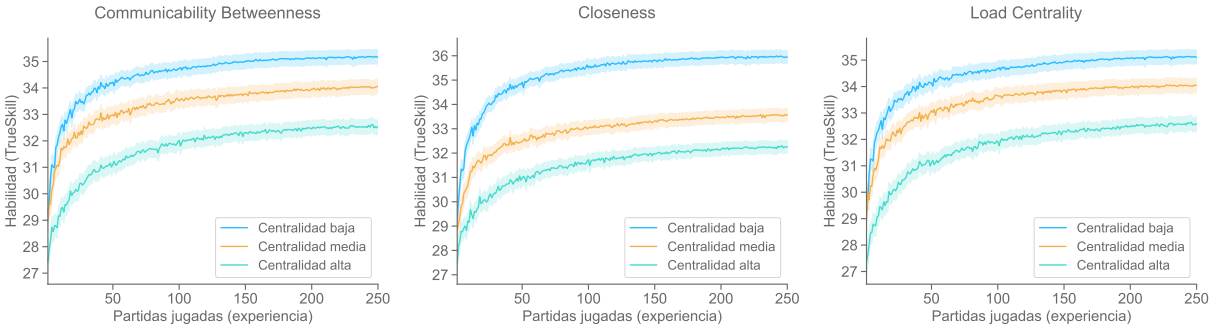


Figura 5.9: Curvas de evolución de la habilidad (TrueSkill) a lo largo del tiempo (medido en partidas jugadas), separado por cohortes en base a valores para distintas medidas de centralidad. En todos los casos, la medida de resumen es la media.

Algo que no nos convencía de utilizar TrueSkill es el hecho de que todos los jugadores comienzan con un valor inicial de habilidad pre-definido y artificial. Realmente, ¿estaremos viendo la evolución de la habilidad en función de la experiencia? ¿O será en realidad que, simplemente, estamos observando la cantidad de partidas que necesita TrueSkill para converger a un valor de habilidad para cada jugador?

Esta situación motivó, fundamentalmente, utilizar TrueSkill Through Time para las estimaciones, para poder considerar que algunos jugadores van a comenzar a utilizar el sitio teniendo ya mucha más experiencia jugando Go que otros. La figura 5.10 muestra los mismos grupos que se incluyeron en la figura 5.9 pero obteniendo la evolución de la habilidad con TrueSkill Through Time.

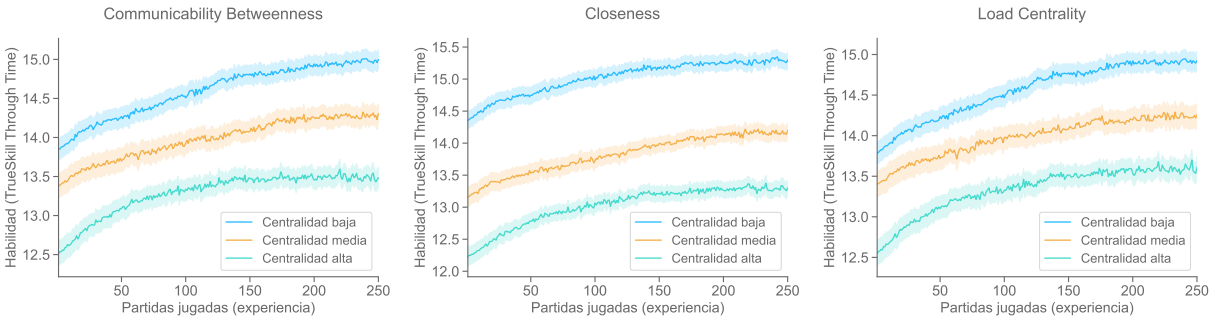


Figura 5.10: Curvas de evolución de la habilidad (TrueSkill Through Time) a lo largo del tiempo (medido en partidas jugadas), separado por cohortes en base a valores para distintas medidas de centralidad. En todos los casos, la medida de resumen es la media.

Ahora es claro que las cohortes tienen puntos de comienzo distinto, gracias a la capacidad que posee TrueSkill Through Time para incorporar también información "hacia atrás". Esto muestra que, entonces, los jugadores en cohortes distintas parecen tener distintos valores iniciales de habilidad.

Por otro lado, las cohortes parecen ser muy buenas separando las distintas curvas de aprendizaje para

las centralidades presentadas. Cada punto de diferencia en la escala de TrueSkill utilizada representa que el jugador con mayor habilidad tendrá un 76 % más de probabilidad de ganar que el jugador con menor valor. En las figuras presentadas, se puede observar alrededor de un punto de diferencia entre curvas, con hasta 2 puntos de diferencia entre las curvas de baja y alta centralidad.

Es necesario mencionar que originalmente, comenzamos sin aplicar ningún filtro a la población considerada. Sin embargo, notamos que a medida que crecía la cantidad de partidas jugadas, la curva se volvía extremadamente ruidosa.

Considerando la distribución exponencial de la cantidad de partidas por jugador (vista en la figura 5.2), y para mantener una población estable, decidimos definir a un *jugador activo* como aquel que tenga al menos 250 partidas jugadas. De no hacer esto, podríamos estar viendo cambios en la curva de aprendizaje como artefacto de la población disminuyendo, y no porque realmente haya una correlación entre la centralidad y el aprendizaje. Además, todos los gráficos tendrán en cuenta partidas hasta la 250, por lo que las poblaciones se mantendrán fijas siempre para cada curva.

El hecho de que las curvas no estaban en el orden que esperábamos, nos hizo comenzar a dudar de que la forma y ubicación de las curvas fuera un artefacto de otras características de la población que no estábamos teniendo en cuenta, por lo que decidimos comenzar a filtrarla para intentar remover algunos de estos factores.

Como dijimos anteriormente, la diferencia de habilidad inicial de los jugadores en las distintas cohortes es importante. Esto puede estar mostrando que distintos subgrupos podrían tener distintas habilidades iniciales a la hora de comenzar a jugar en OGS, razón por la cual tiene sentido agrupar a los jugadores. Sin embargo, aunque parecen tener un período inicial de bastante crecimiento, después se mantienen relativamente constantes, y hasta presentan decrecimiento hacia el final. Teorizamos que esto podría deberse a que, como tomamos la media de centralidad sobre todo el historial de partidas, estaríamos agrupando a los jugadores de forma tal que tuvieran valores similares de habilidad, en lugar de agruparlos para poder ver la evolución a lo largo del tiempo.

Por esta razón, decidimos separar a la población en base a su habilidad o *skill* inicial, y luego aplicar la misma categorización por cohortes usada anteriormente. Para esto, definimos al *skill inicial* como el primer valor de habilidad de cada jugador. Luego, separamos a la población en cinco grupos de acuerdo a su habilidad inicial según *TrueSkill Through Time* (TTT), como puede verse en la Figura 5.11. De estos grupos, nos quedamos con TTT inicial bajo (entre 6 y 10), medio (entre 12 y 15) y alto (entre 17 y 20). Los dos grupos restantes fueron descartados de este análisis para no incluir a jugadores que se encontraran en “zonas grises” y asegurarse que las cohortes estuvieran separadas.

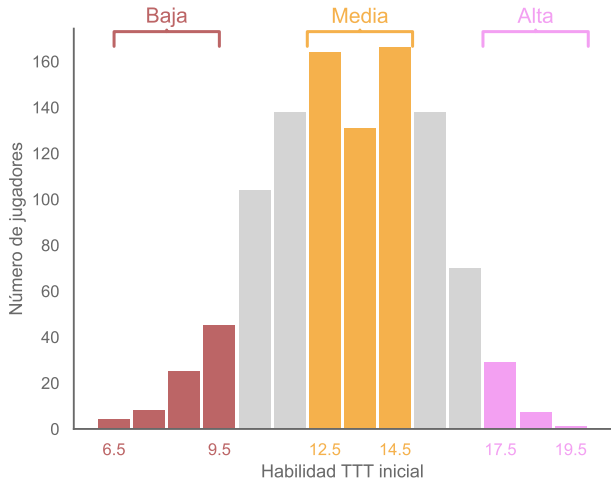


Figura 5.11: Histograma mostrando la ditribución de la habilidad inicial (TTT) para jugadores con al menos 150 partidas jugadas. Las barras coloreadas de rojo, naranja y rosa representan las tres cohortes: habilidad inicial baja, media, y alta. Las barras grises fueron excluidas de las cohortes para no incluir a jugadores que se encontraran en "zonas grises", es decir, entre medio de dos cohortes.

Es importante resaltar que solo fue posible dividir en cohortes por el valor de habilidad inicial una vez tuviéramos los valores estimados de habilidad con TTT. Sin TTT, los valores iniciales de habilidad estimados hubieran sido completamente falsos (ver las secciones de *TrueSkill* y *TrueSkill Through Time* en 2). Con TTT, en cambio, podemos tener más certeza de cuáles eran los verdaderos valores de habilidad, incluso en la primera partida.

Naturalmente, esto disminuyó mucho la población con lo cual necesitamos modificar nuestra definición de *jugador activo* para incluir a más jugadores. De ahora en más, un jugador activo será aquel con al menos 150 partidas.

Esto nos lleva al siguiente problema: ¿será posible que los valores iniciales de habilidad estén influyendo en la topología de la red? Como el sitio web tiene en cuenta la habilidad de cada jugador a la hora de elegir oponentes, ¿no será el caso entonces que la centralidad se verá afectada como consecuencia de la habilidad de los jugadores en lugar de lo opuesto?

A partir de esto, decidimos que únicamente tendremos en cuenta las primeras 50 para resumir la centralidad de cada jugador y categorizarlo en una cohorte. Si tan solo categorizamos a los jugadores por su centralidad en las primeras partidas, podremos ver si la topología efectivamente juega un rol en la evolución de la habilidad, y no al revés. Esto se debe a que queremos asegurarnos de no estar incluyendo en el mismo grupo a jugadores que, a medida que pasa el tiempo aprenden a ritmos muy distintos y eventualmente se separan, pero no hayamos sido capaz de detectarlos por ver todas sus partidas jugadas como un todo.

Estos cambios generaron curvas muy interesantes. Por un lado, aquellas correspondientes al subconjunto de la población con un TTT inicial bajo tienen un fuerte crecimiento, aunque no es posible separarlas correctamente utilizando cohortes por centralidad. Esto puede deberse a que, independientemente de su posición en la red, todos los jugadores aprenden mucho ya que son novatos. Por otro, las curvas correspondientes al subconjunto con un TTT inicial alto tienen mucho ruido (por ser poblaciones pequeñas) y son casi constantes, ya que jugadores experimentados no tienen mucho por aprender.

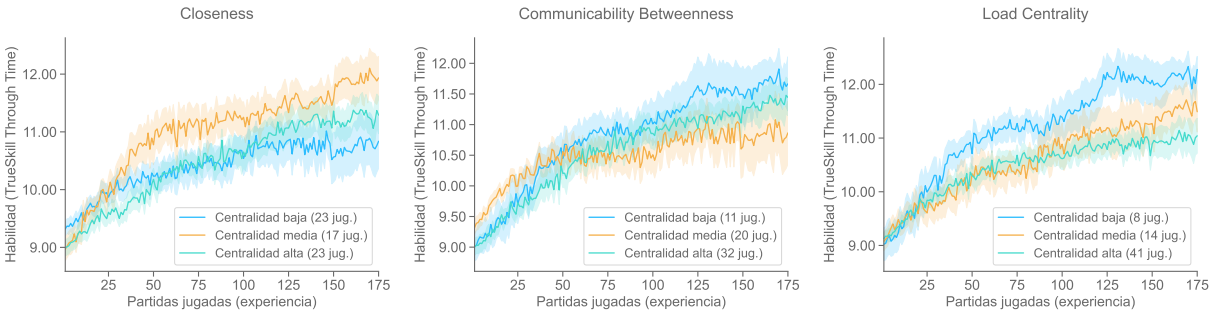


Figura 5.12: Curvas de evolución de la habilidad (TrueSkill Through Time) a lo largo del tiempo (medido en partidas jugadas), separado por cohortes en base a valores para distintas medidas de centralidad. En todos los casos, la medida de resumen es la media, y el valor de TTT inicial es bajo, según las cohortes mostradas en la figura 5.11.

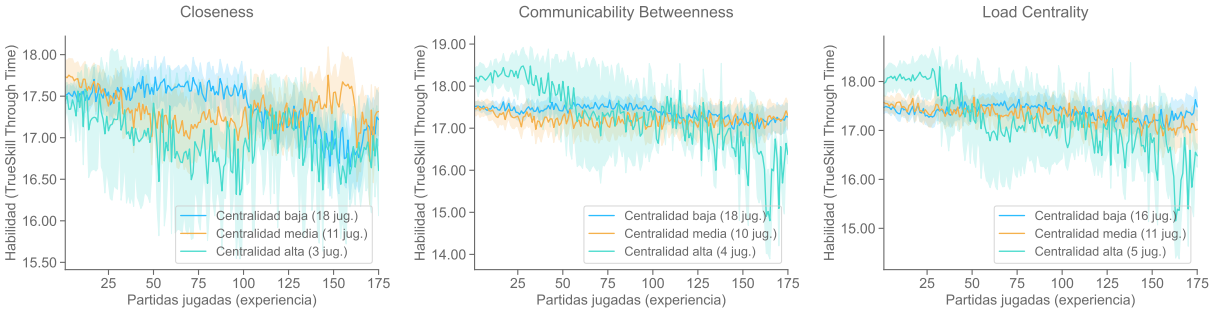


Figura 5.13: Curvas de evolución de la habilidad (TrueSkill Through Time) a lo largo del tiempo (medido en partidas jugadas), separado por cohortes en base a valores para distintas medidas de centralidad. En todos los casos, la medida de resumen es la media, y el valor de TTT inicial es alto, según las cohortes mostradas en la figura 5.11.

Finalmente, las interesantes, son las curvas para la población con TTT inicial medio. En este caso, se mantiene el orden original de las curvas: centralidad baja por encima de media, y media por encima de alta. Además, la distancia entre las curvas de centralidad baja y alta es mucho más grande en este caso.

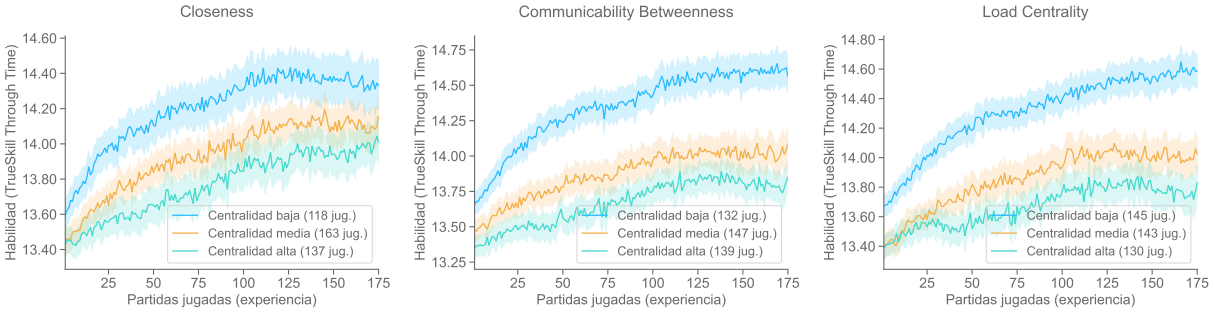


Figura 5.14: Curvas de evolución de la habilidad (TrueSkill Through Time) a lo largo del tiempo (medido en partidas jugadas), separado por cohortes en base a valores para distintas medidas de centralidad. En todos los casos, la medida de resumen es la media, y el valor de TTT inicial es medio, según las cohortes mostradas en la figura 5.11.

Otra nota interesante sobre estos gráficos es el hecho de que se puede ver una correspondencia con la

curva de aprendizaje presentada en la figura 1.4. Observando las figuras 5.12 a 5.14, podemos ver que:

- las curvas de TTT inicial bajo (figura 5.12) se corresponden con la zona **1**, ya que presentan el mayor ritmo de aprendizaje, y el mismo es independiente de su centralidad (lo cual significa que todos los jugadores aprenden mucho independientemente de su posición en la red)
- las de TTT inicial medio (figura 5.14) se corresponden con la zona **2**, ya que todavía presentan un crecimiento en su aprendizaje, pero no es tan pronunciado, y aquí sí depende de su centralidad, lo cual nos da la pauta de que no todos los jugadores están aprendiendo al mismo ritmo (algunos tal vez ya tienen más conocimiento por su posición en la red)
- por último, en las de TTT inicial alto (figura 5.13), como estos jugadores ya tienen un gran conocimiento del juego, no aprenden demasiado (o nada), y se encuentran en la zona **3**

Estos resultados para la población con TTT inicial media nos convencieron de que esta característica no sea un artefacto. Sólo restaba solucionar un problema. Para un jugador, la partida X puede haber sucedido una semana después de la $X - 1$, o 5 años más tarde. Si la diferencia entre partidas es muy grande, entonces es probable que la habilidad del jugador en el sitio no esté actualizada respecto de la realidad: puede que sea mucho más alta o mucho más baja. Esto también puede generar artefactos en las curvas, haciéndolas crecer o decrecer artificialmente.

Para solucionar esto, filtramos nuevamente a la población para solamente tener en cuenta a aquellos jugadores cuya diferencia de tiempo entre su primera y última partidas jugadas sean no más de dos años.

Otro cambio que decidimos introducir fue juntar las curvas para centralidad media y centralidad alta. Hicimos esto porque, como no presentaban grandes diferencias, queríamos asegurarnos de que no era un artefacto por *"separar de más"* el hecho de que la curva de centralidad baja estuviera por encima de ellas. ¿Será posible que, al juntar ambas cohortes, se dieran vuelta las curvas?

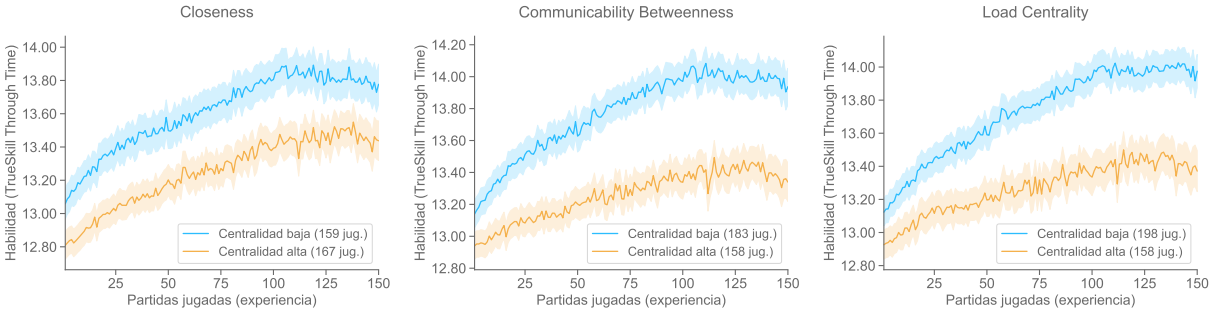


Figura 5.15: Curvas de evolución de la habilidad (TrueSkill Through Time) a lo largo del tiempo (medido en partidas jugadas), separado por cohortes en base a valores para distintas medidas de centralidad. Los jugadores fueron filtrados para no tener en cuenta a aquellos cuya diferencia de tiempo entre su primera y última partida jugadas fuera mayor a dos años. En todos los casos, la medida de resumen es la media, y el valor de TTT inicial es medio, según las cohortes mostradas en la figura 5.11, pero juntando las cohortes media y alta.

Como se puede ver en la figura 5.15, no fue el caso. La separación igual se mantuvo, y la curva de la cohorte con centralidad baja permaneció por sobre la otra.

Capítulo 6

Conclusiones y trabajo futuro

En esta tesis nos propusimos estudiar la hipótesis de que la topología de una red de jugadores de Go (basada en partidas tomadas de un portal en línea) tendría un efecto sobre el aprendizaje de los jugadores. Encontramos que efectivamente parece haber una relación entre ambas variables. Los resultados están en línea con estudios previos en áreas similares, que concluyeron que una conectividad no muy alta ni muy baja resulta en un intercambio de información óptimo en comunidades: **son los modelos con conectividad moderada los que muestran una acumulación cultural (CCE) óptima [19, 20].**

El capítulo 1 presentó el marco teórico necesario. Primero se discutieron los distintos trabajos ya existentes que exploran la transmisión de información cultural y la posibilidad mencionada de que una conectividad media sea óptima para el traspaso de esta información dentro de la comunidad. Aquí se hace mención a que grandes cambios en los números de la población (producidas por ejemplo por migraciones) pueden generar grandes cambios en el traspaso de la información. Podemos considerar la fusión de OGS con Nova como un evento disruptivo en este aspecto, parecido a una migración, dado que las comunidades de ambos sitios fueron fusionadas. Es por esta razón que decidimos acotar el estudio hasta justo antes de la fusión de los sitios. Otro concepto explicado en este capítulo es el de barreras estructurales y conductuales. Es posible pensar al sistema de *matchmaking* o de ranking como una barrera estructural, ya que el sitio condiciona contra quiénes puede jugar cada jugador. Asimismo, se puede considerar una barrera conductual el hecho de elegir repetir jugar con las mismas personas (o con ciertos grupos) en base a su habilidad, o a experiencias previas. Por ejemplo, un jugador puede sentir que aprende mucho jugando contra alguien en particular, entonces repite partidas contra ese oponente. A continuación, se incluyó una breve introducción al Go y a sus reglas, así también como al portal en línea OGS, de donde obtuvimos los datos utilizados en esta tesis, y a su sistema de matchmaking. Por último, se presentó brevemente el concepto de las curvas de aprendizaje poblacional y cómo pueden representarse como variantes de la ley potencial, es decir, son curvas exponenciales.

En el capítulo 2 fue presentado un modelo causal que explica cómo es afectada esta información cultural, y como la misma afecta a la adaptación de los individuos. Seguidamente se explicaron distintos modelos de estimación de habilidad utilizados en otros juegos, como ajedrez, y en particular en OGS: Elo, Glicko, TrueSkill, y TrueSkill Through Time. Se mostró cómo cada modelo se basó en el anterior y añadió mejoras. En particular, se presentó cómo TrueSkill Through Time es capaz de utilizar la información de nuevas partidas para no solo actualizar las predicciones a futuro, sino también actualizar valores previos de habilidad.

El capítulo 3 se centra en explicar de manera detallada las medidas de centralidad utilizadas más adelante en nuestra investigación, y las distintas maneras de analizar grafos dinámicos. Se enfatizó el presentar intuiciones sobre las distintas medidas, utilizando pequeños ejemplos, dado que algunos de los conceptos pueden ser difíciles de entender y no existe mucha literatura que las explique de manera sencilla.

El capítulo 4 describe el proceso de investigación llevado a cabo. Se explica cómo se consiguieron los

datos, se describe el esquema de la base de datos, y se presentan los inconvenientes que surgieron y cómo fueron solucionados, hasta obtener un dataset "limpio". También se incluye una breve sección hablando sobre los distintos tipos de bases de datos y por qué se eligió una basada en grafos para este trabajo. Finalmente, se habla sobre qué tecnologías se utilizaron para llevar a cabo los análisis.

El capítulo 5 es el corazón de nuestro trabajo. Primero se muestran resultados exploratorios de los datos, como grado medio o tamaño de la componente conexa más grandes. También se incluyen diversos histogramas, como de cantidad de partidas jugadas por jugador o de valores de centralidad. Aquí también se presenta formalmente la hipótesis central de nuestro trabajo, y se describe el proceso mediante el cual iteramos nuevas ideas a partir de las preguntas que fueran surgiendo de los resultados anteriores. Uno de los más interesantes es tal vez el presentado en las figuras 5.12, 5.13 y 5.14, donde se muestran las curvas de aprendizaje separadas por cohortes en base a la centralidad, y subdividiendo a la población en base a su habilidad inicial.

Como fue notado en la presentación de los resultados, en el caso de la habilidad inicial baja, las curvas no se separan, pero tienen un alto crecimiento. Esto podemos interpretarlo de la siguiente manera: como los jugadores no saben nada, van a aprender de cualquier interacción que tengan. No importa su nivel de compromiso ni su conectividad al resto de la red - en este caso el factor más importante será la experiencia de jugar, mediante la cual los jugadores novatos siempre aprenderán algo.

Por otra parte, los jugadores con una habilidad inicial alta se encuentran en el extremo opuesto: en lugar de un alto crecimiento, vemos una curva de aprendizaje que se mantiene constante. La única similitud es que las cohortes por centralidad tampoco parecen tener curvas muy distintas. Estos jugadores ya sabían jugar desde antes de unirse al portal, por lo que su conectividad será irrelevante a la hora de aprender. Jugarán con otros jugadores de habilidad similar, y no tendrán un gran aprendizaje en cada encuentro.

Por último, tenemos el caso interesante de la habilidad inicial media. Aquí sí hay marcadas diferencias entre las curvas de aprendizaje resultantes de cada cohorte por centralidad. Es muy interesante resaltar que, aunque los jugadores arrancan con una habilidad inicial casi idéntica, rápidamente se diferencian. Como también notamos, la cohorte con centralidad más baja presenta la evolución de habilidad más alta, mientras que las dos cohortes de mayor centralidad muestran una menor habilidad. ¿Cómo explicamos esto? Una posible explicación tiene que ver con el *compromiso* de los jugadores.

Podemos definir compromiso como qué tanto interés tiene el jugador por aprender y mejorar. Los jugadores con bajo compromiso (es decir, que no tienen un particular interés por aprender, que tal vez juegan para pasar el rato), jugarán con cualquier oponente. No les importará demasiado quién es el contrincante ya que solo querrán jugar una partida de vez en cuando. En este caso, tendrán una alta centralidad, ya que no elegirán cautelosamente a sus oponentes: simplemente jugarán con quien sea, y esto significará que todas las partidas serán con contrincantes distintos.

Por otro lado, tenemos a los jugadores con un alto compromiso. Aquellos que realmente quieren aprender y mejorar su habilidad. Para estos jugadores, será más importante elegir con mayor cuidado contra quién jugar. Elegirán mejor a sus oponentes, ya sea porque ya han jugado juntos y quieren aprender más de ellos (lo cual nos remite al concepto de jugadores "demostradores" presentado en la introducción), o porque refinarán más las opciones de *matchmaking* para jugar contra jugadores que les brindarán más conocimiento. Esta mayor cautela a la hora de elegir jugadores, o el conocimiento de saber con quién jugar y con quién no, resultará en una centralidad menor, ya que no jugarán con cualquier persona, y el grupo de contrincantes será más reducido.

Bajo este marco, tiene sentido pensar entonces que los jugadores con más baja centralidad aprenderán más que los otros. Por otro lado, al comienzo de este trabajo, presentamos estudios que parecían haber encontrado una relación entre tener una conectividad moderada en la red analizada y una acumulación cultural óptima [19, 20]. Creemos que nuestros resultados se alinean con esta misma idea.

Nos hemos referido a la centralidad óptima como "centralidad baja". Sin embargo, es importante resaltar que en el contexto estudiado en nuestro caso, es muy difícil tener pocas conexiones entre jugadores, y por ende, es difícil tener jugadores con conectividad extremadamente baja. En los trabajos

presentados en el primer capítulo, se habla de baja conectividad en casos extremos, como el peculiar caso de Tasmania, cuya población queda completamente aislada y sin contacto fuera de la isla. En un portal para jugar online con cientos de otros jugadores alrededor del mundo, es muy fácil relacionarse y no quedar aislado. Esto también significa que la población no sufrirá grandes cambios, sino que mas bien se mantendrá relativamente constante al no existir eventos externos que la modifiquen extremadamente. Ésto justifica además la decisión de cortar nuestro análisis en 2013, ya que lo hicimos como una manera de evitar incluir en el mismo un evento externo como fue la unión de dos bases de datos de jugadoers.

También, recordemos que filtramos a la población analizada para únicamente tener en cuenta a los jugadores con más de 150 partidas jugadas. De haber jugadores realmente muy mal conectados, estos serán filtrados en esta instancia. Si tenemos en cuenta ambas cosas, en realidad es razonable pensar que una conectividad baja en nuestro contexto estaría a la par de una conectividad media en los trabajos mencionados.

Durante nuestro análisis, surgieron algunas preguntas que quedaron sin responder, o ideas que quedaron sin explorar en este trabajo. Algunas de ellas son:

- ¿El matchmaking modifica la centralidad de los jugadores? ¿Cuál es exactamente el efecto del algoritmo de matchmaking sobre la topología del grafo?
- Usar medidas de centralidad para grafos dinámicos, u otros posibles enfoques que no incluyan *snapshots* o ventanas.
- ¿Se puede utilizar el compromiso como otra variable para dividir a la población? ¿Tendrá efectivamente un efecto sobre la topología?
- Una pregunta planteada en Derek et al. [4]: ¿es posible que el nivel óptimo de conectividad dependa de qué tan difícil es la habilidad que se está intentando aprender?
- Al plantear un modelo formal incluyendo la centralidad (por ejemplo, una regresión lineal), ¿será ésta una variable significativa?

Bibliografía

[1] Glickman ME. Parameter Estimation in Large Dynamic Paired Comparison Experiments. J Roy Stat Soc C-App. 1999;48(3):377–394. doi:10.1111/1467-9876.00159.

[2] Elo AE. The Rating of Chess Players, Past and Present. Ishi Press; 2008.

[3] Herbrich R, Minka R, Graepel T. TrueSkill: A Bayesian Skill Rating System. In: Proc of the Conf in Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press; 2006. p. 569–576.

[4] Derex M, Mesoudi A. Cumulative Cultural Evolution within Evolving Population Structures. Trends Cogn Sci. 2020;24(8):654–667. doi:10.1016/j.tics.2020.04.005.

[5] Cavalli-Sforza LL, Feldman MW. Cultural transmission and evolution: a quantitative approach. Princeton University Press; 1981.

[6] Boyd R, Richerson PJ. Culture and the evolutionary process. University of Chicago press; 1985.

[7] Rendell L, Boyd R, Cownden D, Enquist M, Eriksson K, Feldman MW, et al. Why copy others? Insights from the social learning strategies tournament. Science. 2010;328(5975):208–213.

[8] Landfried G, Slezak DF, Mocskos E. Faithfulness-boost effect: Loyal teammate selection correlates with skill acquisition improvement in online games. PLoS ONE. 2019;14(3):e0211014.

[9] Tomasello M, Kruger AC, Ratner HH. Cultural learning. Behavioral and Brain Sciences. 1993;16(3):495–511. doi:10.1017/S0140525X0003123X.

[10] Tennie C, Call J, Tomasello M. Ratcheting up the ratchet: on the evolution of cumulative culture. Philosophical Transactions of the Royal Society B: Biological Sciences. 2009;364(1528):2405–2415. doi:10.1098/rstb.2009.0052.

[11] Shennan S. Demography and cultural innovation: a model and its implications for the emergence of modern human culture. Cambridge archaeological journal. 2001;11(1):5–16.

[12] Henrich J. Demography and cultural evolution: how adaptive cultural processes can produce maladaptive losses—the Tasmanian case. Am Antiquity. 2004;69(2):197–214. doi:10.2307/4128416.

[13] Wisdom TN, Song X, Goldstone RL. Social learning strategies in networked groups. Cognitive Science. 2013;37(8):1383–1425.

[14] Kempe M, Mesoudi A. An experimental demonstration of the effect of group size on cultural accumulation. Evolution and Human Behavior. 2014;35(4):285–290.

[15] Caldwell CA, Millen AE. Human cumulative culture in the laboratory: effects of (micro) population size. Learning & Behavior. 2010;38(3):310–318.

[16] Fay N, De Kleine N, Walker B, Caldwell CA. Increasing population size can inhibit cumulative cultural evolution. Proc Natl Acad Sci USA. 2019;116(14):6726–6731.

[17] Powell A, Shennan S, Thomas MG. Late Pleistocene demography and the appearance of modern human behavior. Science. 2009;324(5932):1298–1301.

[18] Migliano AB, Battiston F, Viguier S, Page AE, Dyble M, Schlaepfer R, et al. Hunter-gatherer multilevel sociality accelerates cumulative cultural evolution. Science Advances. 2020;6(9):eaax5913.

[19] Creanza N, Kolodny O, Feldman MW. Greater than the sum of its parts? Modelling population contact and interaction of cultural repertoires. Journal of The Royal Society Interface. 2017;14(130):20170171. doi:10.1098/rsif.2017.0171.

[20] Derex M, Perreault C, Boyd R. Divide and conquer: intermediate levels of population fragmentation maximize cultural accumulation. Phil Trans R Soc B. 2018;373(1743):20170062.

[21] Derex M, Boyd R. Partial connectivity increases cultural accumulation within groups. Proc Natl Acad Sci USA. 2016;113(11):2982–2987.

[22] Newell A, Rosenbloom PS, Anderson JR. Mechanisms of skill acquisition and the law of practice. In: Cognitive skills and their acquisition. Psychology Press; 1981. p. 1–55.

[23] Elo AE. The USCF rating system. U.S. Chess Federation publication; 1961.

[24] Bishop CM. Pattern Recognition and Machine Learning. Berlin, Heidelberg: Springer; 2006.

[25] Jaynes ET. Probability theory: The logic of science. Cambridge University Press; 2003.

[26] Van Horn KS. Constructing a logic of plausible inference: a guide to cox's theorem. Int J Approx Reason. 2003;34(1):3–24. doi:10.1016/S0888-613X(03)00051-3.

[27] Glickman M. Example of the Glicko-2 system; 2013. `http://www.glicko.net/glicko/glicko2.pdf`.

[28] Dangauthier P, Herbrich R, Minka T, Graepel T. Trueskill through time: Revisiting the history of chess. In: Platt JC, Koller D, Singer Y, Roweis ST, editors. Proc of the Twenty-First Annual Conf on Neural Information Processing Systems. Curran Associates, Inc.; 2007. p. 337–344.

[29] Harary F, Gupta G. Dynamic Graph Models. Mathematical and Computer Modelling. 1997;25(7):79–87. doi:10.1016/S0895-7177(97)00050-2.

[30] Ghanem M, Magnien C, Tarissan F. Centrality Metrics in Dynamic Networks: A Comparison Study. IEEE Transactions on Network Science and Engineering. 2019;6(4):940–951. doi:10.1109/TNSE.2018.2880344.

[31] Kossinets G, Watts DJ. Empirical Analysis of an Evolving Social Network. Science. 2006;311(5757):88–90. doi:10.1126/science.1116869.

[32] Estrada E, Higham DJ, Hatano N. Communicability betweenness in complex networks. Physica A: Statistical Mechanics and its Applications. 2009;388(5):764–774. doi:10.1016/j.physa.2008.11.011.

[33] Goh KI, Kahng B, Kim D. Universal Behavior of Load Distribution in Scale-Free Networks. Physical Review Letters. 2001;87(27):278701. doi:10.1103/PhysRevLett.87.278701.

[34] Brandes U, Fleischer D. Centrality Measures Based on Current Flow. In: Diekert V, Durand B, editors. Proc of the Annual Symp on Theoretical Aspects of Computer Science (STACS). Berlin, Heidelberg: Springer Berlin Heidelberg; 2005. p. 533–544.