

Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California



Doctorado en Ciencias en Ciencias de la Computación

Problemas de formación de equipo

Tesis

para cubrir parcialmente los requisitos necesarios para obtener el grado de
Doctor en Ciencias

Presenta:

Julio Antonio Juárez Jiménez

Ensenada, Baja California, México

2022

Tesis defendida por

Julio Antonio Juárez Jiménez

y aprobada por el siguiente Comité

Dr. Carlos Alberto Brizuela Rodríguez
Director de tesis

Dr. Cipriano Santos

Dr. José Alberto Fernández Zepeda

Dr. Ubaldo Ruiz López



Dr. Pedro Gilberto López Mariscal
Coordinador del Posgrado en Ciencias de la Computación

Dr. Pedro Negrete Regagnon
Director de Estudios de Posgrado

Julio Antonio Juárez Jiménez © 2022

Queda prohibida la reproducción parcial o total de esta obra sin el permiso formal y explícito del autor y director de la tesis

Resumen de la tesis que presenta Julio Antonio Juárez Jiménez como requisito parcial para la obtención del grado de Doctor en Ciencias en Ciencias de la Computación.

Problemas de formación de equipo

Resumen aprobado por:

Dr. Carlos Alberto Brizuela Rodríguez

Director de tesis

En vista del creciente interés en los equipos de trabajo efectivos y en cómo formarlos, surge un nuevo y desafiante problema de optimización combinatoria denominado Problema de Formación de Equipo (PFE). EL PFE atiende la siguiente problemática: dado un conjunto de candidatos y una función de efectividad de equipo, ¿cuáles de los candidatos formarían un equipo de máxima efectividad para una tarea en particular? Tras poco más de dos décadas de investigación sobre el PFE, aún se sabe muy poco del problema en sí. Mientras tanto la formulación de modelos particulares sigue en aumento. Esto se debe en parte a que los equipos existen en diversos contextos como la ingeniería, los deportes, los servicios médicos y de emergencia, entre muchos otros. Por otra parte, los modelos también varían en cuanto a su función de efectividad de equipo y a sus datos de entrada. Los primeros PFEs modelan la efectividad de equipo simplemente con respecto a las habilidades individuales de los candidatos, mientras que los modelos subsecuentes se enfocan más en la interacción de los integrantes. En esta tesis, se propone una clasificación taxonómica de los PFEs desde la perspectiva de la estructura de los modelos. Adicionalmente, se estudian los elementos para un modelo de formación de equipo más general. Particularmente, se propone un modelo no simétrico para medir y clasificar la fuerza del vínculo entre dos personas, tomando como entrada una red social profesional. Finalmente, se usa este modelo no simétrico para diseñar un criterio de densidad de colaboración entre miembros del equipo, y se implementa en un modelo de optimización multi-objetivo de PFE. En primera instancia, la taxonomía propuesta ayuda a organizar los modelos de la literatura, promueve la identificación de elementos esenciales en los modelos, y también establece una relación entre los PFEs y otros problemas combinatorios conocidos. En segunda instancia, el modelo no simétrico propuesto para medir y clasificar la fuerza del vínculo entre personas, revela una dimensión adicional en los datos que permite analizar estructuras de interacción más complejas. Finalmente, el criterio de densidad de colaboración propuesto muestra un mejor desempeño en comparación con varios criterios de distancia de comunicación, al implementarse en combinación con un criterio de habilidad técnica del equipo, mediante un enfoque multi-objetivo. Colectivamente, estos resultados, junto con evidencia de otras disciplinas referente a la efectividad de equipos, ayudan a entender mejor el problema y las consideraciones para diseñar modelos más generales, y el desarrollo de una teoría alrededor del problema.

Palabras clave: problema de formación de equipo, taxonomía, análisis de redes sociales, modelo de interacción, criterios de efectividad de equipo

Abstract of the thesis presented by Julio Antonio Juárez Jiménez as a partial requirement to obtain the Doctor of Science degree in Computer Science.

Team formation problems

Abstract approved by:

Dr. Carlos Alberto Brizuela Rodríguez
Thesis Director

In view of the growing interest in effective work teams and how to form them, a new and challenging combinatorial optimization problem called the Team Formation Problem (TFP) has emerged. The TFP addresses the following problem: given a set of candidates and a team effectiveness function, which of the candidates would form a team of maximum effectiveness for a particular task? After a little more than two decades of research on the TFP, we still know very little about the problem itself. Meanwhile the formulation of particular models continues to increase. This is partly because teams exist in diverse contexts such as engineering, sports, medical and emergency services, among many others. Moreover, models also vary in terms of their team effectiveness function and its inputs. Early TFPs model team effectiveness simply with respect to the individual skills of the candidates, while subsequent models focus more on the interaction of the members. In this thesis, a taxonomic classification of TFPs is proposed from the perspective of the models' structure. Additionally, the elements for a more general team formation model are studied. In particular, a non-symmetric model is proposed to measure and classify the tie strength between two people, taking as input a professional social network. Finally, this non-symmetric model is used to design a criterion based on the density of collaboration between team members, and implemented in a multi-objective optimization TFP model. In the first instance, the proposed taxonomy helps to organize the models in the literature, promotes the identification of essential elements in the models, and also establishes a relationship between TFPs and other well-known combinatorial problems. Second, the proposed non-symmetric model for measuring and classifying the tie strength reveals an additional dimension in the data that allows us to analyze more complex interaction structures. Finally, the proposed criterion shows a better performance compared to multiple criteria based on communication distance, when implemented in combination with a criterion of technical skill level in a team, using a multi-objective approach. Collectively, these results, along with the evidence from other disciplines regarding team effectiveness, help us to better understand the problem and the considerations for designing more general models, and the development of a theory around the problem.

Keywords: team formation problem, taxonomy, social network analysis, interaction model, team effectiveness criteria

Dedicatoria

A Mamá.

A Papá[†].

Agradecimientos

Primeramente debo agradecer a mi director, el Dr. Carlos Brizuela, quien me dio la libertad y la confianza para proponer este tema de tesis. Le agradezco especialmente todo el tiempo dedicado, todas esas reuniones, todas esas discusiones que sin duda me ayudaron a madurar el pensamiento crítico. Eso es algo que me acompañará por el resto de mi vida.

A los miembros del comité de tesis, el Dr. Pano Santos, el Dr. Alberto Fernández y el Dr. Ubaldo Ruiz. Su siempre sincera y oportuna retroalimentación hicieron de éste un mejor trabajo de tesis. Debo agradecer además por las veces que amablemente me asesoraron en otros temas, fue de suma importancia para complementar mi formación.

A mis compañeros, en especial a aquellos que pasaron por este proceso antes que yo y de quienes tuve la oportunidad de aprender tanto. Sus consejos me evitaron muchos problemas en el camino.

A mi familia, por siempre estar al pendiente de mí. Especialmente a mi madre, quien constantemente me reiteró su apoyo para que siguiera adelante y a quien nunca podré compensar por el tiempo que estuve ausente.

A Cristina, por compartir mis mejores momentos, pero principalmente por su cálida compañía en los momentos de mayor dificultad.

Al Centro de Investigación Científica y de Educación Superior de Ensenada por brindarme las herramientas necesarias y suficientes para completar mi doctorado.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por brindarme el apoyo económico para realizar mis estudios de doctorado por medio de la beca No. 280320.

Tabla de contenido

	Página
Resumen en español	ii
Resumen en inglés	iii
Dedicatoria	iv
Agradecimientos	v
Lista de figuras	ix
Lista de tablas	xi
 Capítulo 1. Introducción	
1.1. Antecedentes	2
1.2. Planteamiento del problema	3
1.3. Objetivos	4
1.3.1. Objetivo general	4
1.3.2. Objetivos específicos	4
1.4. Contribuciones del trabajo de tesis	5
1.5. Nota aclaratoria	6
1.6. Estructura de la tesis	7
 Capítulo 2. Revisión de la literatura	
2.1. Equipos y su efectividad	9
2.2. Problema de formación de equipo	11
2.3. Componentes de los modelos y sus variables	14
2.3.1. Habilidades técnicas	15
2.3.1.1. Matriz de incidencia candidato–posición	15
2.3.1.2. Variables lingüísticas y números difusos	16
2.3.1.3. Proceso jerárquico analítico	18
2.3.2. Rasgos de personalidad	19
2.3.2.1. El indicador Myers-Briggs	19
2.3.2.2. Otros indicadores	21
2.3.3. Habilidades de trabajo en equipo	21
2.3.3.1. Comunicación mediante red social	22
2.4. Discusión	23
 Capítulo 3. Propuesta de taxonomía de los problemas de formación de equipo	
3.1. Modelos basados en asignación	25
3.1.1. Formación de un equipo	26
3.1.2. Formación de muchos equipos	29
3.1.3. Formación de equipos afines	33
3.2. Modelos basados en comunidad	35
3.2.1. Habilidades binarias	36
3.2.2. Habilidades ponderadas	38

3.2.3. Habilidades probabilísticas	39
3.3. Discusión	40

Capítulo 4. Propuesta de una medida de interacción no simétrica

4.1. Preliminares	43
4.1.1. Modelos de interacción en la literatura de formación de equipos	43
4.1.2. El flujo de información en las redes sociales	45
4.1.2.1. La fuerza de los vínculos débiles	45
4.1.2.2. El portero tecnológico	46
4.2. Propuesta de un modelo de interacción no simétrico	47
4.3. Metodología	48
4.3.1. Datasets	49
4.3.1.1. DBLP	49
4.3.1.2. GitHub	50
4.3.1.3. IMDb	50
4.3.2. Construyendo el grafo	50
4.3.2.1. Clasificando vínculos por su fuerza	51
4.4. Análisis y resultados	53
4.4.1. Vínculos fuertes y débiles	53
4.4.2. Puentes locales	55
4.4.3. Rumbo a la identificación de porteros tecnológicos	57
4.4.4. Análisis de estructuras 3-clique	59
4.5. Discusión	62

Capítulo 5. Estudio comparativo de criterios de optimización desde una perspectiva multi-objetivo: densidad versus distancia

5.1. Preliminares	65
5.1.1. Criterios de optimización para el Problema de Formación de Equipos Basados en Comunidades	65
5.1.1.1. Criterios basados en distancia	66
5.1.1.2. Criterios basados en densidad	66
5.1.1.3. Debate entre criterios basados en distancia y criterios basados en densidad	67
5.1.1.4. Otros criterios de optimización	69
5.1.2. Problemas de formación de equipo multiobjetivo	69
5.2. Propuesta de criterio de densidad basado en interacción no-simétrica y su implementación en un modelo multiobjetivo	70
5.2.1. Formulación multiobjetivo	71
5.3. Metodología	72
5.3.1. Conjunto de datos y el grafo social	72
5.3.2. Casos de prueba	73
5.3.3. Algoritmo evolutivo multi-objetivo	74
5.3.3.1. Representación de la solución	75
5.3.3.2. Inicialización de la población	76
5.3.3.3. Operadores genéticos	76
5.3.4. Diseño de experimento	78

Página

5.3.4.1. Detalles adicionales	80
5.4. Análisis y resultados	83
5.4.1. Experimento A: considerando frente consolidado	83
5.4.2. Experimento B: sin considerar el frente consolidado	84
5.4.3. Tiempo de ejecución	85
5.5. Discusión	87
 Capítulo 6. Conclusiones	
6.1. Conclusiones generales	89
6.2. Limitaciones	90
6.3. Trabajo a futuro	92
Literatura citada	93

Lista de figuras

Figura	Página
1. Funciones de pertenencia de variables lingüísticas.	17
2. Ejemplo de una estructura jerárquica genérica del Proceso Jerárquico Analítico. Existe un peso relativo de importancia asociado a cada conexión. . .	18
3. Taxonomía propuesta de los problemas de formación de equipo.	25
4. Ejemplo de modelos generales basados en asignación. Existe un peso de idoneidad asociado a cada conexión. Izquierda: un grafo bipartito que representa una asociación candidato a puesto de equipo. Derecha: un grafo bipartito que representa una asociación candidato a equipo (el puesto no es importante).	26
5. Ejemplo de un modelo basado en comunidades. Los candidatos están relacionados entre sí mediante una red social, y cada candidato está asociado a un conjunto de habilidades.	36
6. Un ejemplo de los modelos de interacción. Arriba izquierda: número individual de proyectos desempeñados por los individuos A y B . Arriba centro: número total de proyectos distintos desempeñados por A y B . Arriba derecha: número de proyectos donde A y B coinciden. Abajo: izquierda: $w_s(A, B)$ = conteo de colaboraciones de A y B , y $w_J(A, B)$ = índice de Jaccard entre A y B . Abajo derecha: $w_{\rightarrow}(A, B)$ = fuerza de vínculo de A sobre B , y $w_{\rightarrow}(B, A)$ = fuerza de vínculo de B sobre A	48
7. Tres estructuras diádicas. Arriba: los colaboradores cercanos, caracterizados por vínculos fuertes recíprocos o un <i>vínculo completamente fuerte</i> . Centro: el maestro-discípulo, caracterizados por un vínculo fuerte y un vínculo débil en direcciones opuestas, también llamado como <i>vínculo parcialmente fuerte</i> . Abajo: los colaboradores esporádicos, caracterizados por vínculos débiles recíprocos o un <i>vínculo completamente débil</i> . Un vínculo fuerte (débil) está representado mediante una línea continua (discontinua).	49
8. Dos representaciones (dirigida, no-dirigida) de un subgrafo DBLP de la red en torno al Profesor Héctor García-Molina.	52
9. Proporción de vínculos <i>completamente fuertes</i> (fuerte-fuerte), <i>parcialmente fuertes</i> (fuerte-débil), <i>completamente débiles</i> (débil-débil) en los grafos DBLP, GitHub e IMDb.	55
10. Proporción de los puentes locales <i>completamente fuertes</i> , <i>parcialmente fuertes</i> y <i>completamente fuertes</i> en los grafos DBLP, GitHub e IMDb.	56
11. Las siete familias y las 16 configuraciones posibles de 3-cliques considerando una distinción de los vínculos como fuertes (línea continua color naranja) y débiles (línea punteada color cian). En el lado izquierdo se indica el número de vínculos fuerte en la estructura; en el lado derecho, cada posible estructura según la combinación de las configuraciones de sus vínculos.	60

12. Ejemplo de dos configuraciones de estructuras triádicas. Izquierda: el gran maestro–maestro–discípulo, caracterizado por un nodo (el gran maestro) con dos vínculos parcialmente fuertes hacia dos nodos, estos últimos (el maestro–discípulo) con un vínculo parcialmente fuerte entre ellos. El gran maestro debe presentar dos vínculos fuertes salientes y dos vínculos débiles entrantes. Derecha: el ciclo maestro–discípulo, caracterizado por una relación cíclica de maestros–discípulos. Note que esta última configuración es hipotética, es improbable que exista en la realidad. 61
13. Mapa de calor de las configuraciones de los 3-cliques por conjunto de datos. 62
14. Izquierda: grafo social conectado, no denso. Derecha: grafo social desconectado, relativamente denso. 68
15. Ejemplo de un grafo de una red social. 75
16. Ejemplo de metodologías para evaluar conjuntos de soluciones no dominadas. 79
17. Esquema parcial del experimento. Se realizan quince ejecuciones independientes del NSGA-II, resultando en quince frentes no dominados. Estos frentes se combinan de manera distinta para formar los conjuntos de soluciones *CND* y *UND*. 81
18. Esquema parcial del experimento. Se repite el bloque de la Figura 17 quince veces, resultando en quince *CND* y quince *UND*. 82
19. Tiempos de ejecución promedio del NSGA-II por cada caso de prueba, por cada criterio de optimización empleado. 86
20. Ejemplo de conectividad y densidad en un grafo. A mayor densidad, potencialmente mayor conectividad. 88

Lista de tablas

Tabla	Página
1. Tabla de variantes del Problema de Formación de Equipo.	13
2. Tabla de compatibilidad en cuanto a pares de tipos de personalidad del MBTI.	20
3. Tabla normalizada de puntuaciones de compatibilidad entre las tipologías del MBTI de acuerdo a Chen y Lin (2004); Zhang y Zhang (2013).	20
4. Comparativa de los diferentes trabajos de PFEs clasificados por medida de interacción.	45
5. Resumen en números de los grafos por conjunto de datos. Las clasificaciones de fuerza de vínculo se calcularon con k -medias, por cada conjunto de datos, por cada modelo de interacción, de manera independiente. Se usa el conjunto de aristas no dirigidas E para los modelos simétricos (simple y Jaccard) y $\tilde{E}$ para la propuesta.	53
6. Tabla de correlaciones de centralidad de intermediación y recuento de aristas por tipo de fuerza de vínculo.	58
7. Benchmark de casos de prueba generados de manera aleatoria. . . .	74
8. Tabla de ejemplo con los ránquines promedio de los hipervolúmenes generados al optimizar para una tupla de criterios y evaluar para otra tupla de criterios, de los 30 casos de prueba.	82
9. Tabla de criterios a utilizar según su clasificación y su modelo de interacción.	83
10. Tabla de parámetros de ejecución usados para el NSGA-II.	83
11. Ránquines promedio de los hipervolúmenes generados al optimizar para la tupla $(Z, \cdot)$ de la horizontal y evaluar para $(Z, \cdot)$ de la vertical, de los 30 casos de prueba, para el Experimento A considerando el frente consolidado. En verde están aquellos ránquines que encabezan la primera mitad de su renglón y en rojo aquellos ránquines en la peor mitad.	84
12. Ránquines promedio de los hipervolúmenes generados al optimizar para la tupla $(Z, \cdot)$ de la horizontal y evaluar para $(Z, \cdot)$ de la vertical, de los 30 casos de prueba, para el Experimento B sin considerar el frente consolidado. En verde están aquellos ránquines que encabezan la primera mitad de su renglón y en rojo aquellos ránquines en la peor mitad.	85

Capítulo 1. Introducción

La estructura de trabajo en las organizaciones está cambiando, pasando de un paradigma basado en individuos a uno basado en equipos de trabajo (Katzenbach y Smith, 2015; Kozlowski y Bell, 2003; Cohen y Bailey, 1997; Kozlowski y Ilgen, 2006). Los equipos, a diferencia de los individuos, ofrecen a las organizaciones una rara mezcla de habilidades, conocimientos y experiencia que a menudo se requiere para tener éxito en actividades complejas y dinámicas, como la producción de bienes o la prestación de servicios (Swezey y Salas, 1992; Kozlowski y Bell, 2003).

En cuanto a los equipos, es casi seguro que distintas conformaciones de equipo producirán resultados distintos sobre una misma tarea. Por una parte, reclutar a un grupo de personas y darles un objetivo no garantiza que éste funcione como equipo (Solansky, 2011). Por otra parte, algunos equipos podrían ser más efectivos que otros simplemente por la composición de las características de sus integrantes (Campion *et al.*, 1993; Foushee y Helmreich, 1988). Desafortunadamente, no se sabe completamente qué composiciones son las más favorables para la efectividad de equipo. En algunos casos los resultados son contradictorios: varios trabajos reportan que la diversidad demográfica tiene un efecto positivo en la efectividad de equipo (Gladstein, 1984), mientras que otros reportan un posible efecto contrario (Jackson *et al.*, 1991).

Es bien sabido que formar un equipo inefectivo o indispuerto tiene consecuencias en cuanto a resultados no deseados, retrasos y costos innecesarios para la organización. Suponga como ejemplo un escenario de diseño industrial y desarrollo de productos. Se estima que entre el 70 % y el 90 % del costo total del ciclo de vida de un producto se define en las etapas tempranas de diseño (Korpi y Ala-Risku, 2008). Una forma de reducir este costo es formando un equipo de ingeniería concurrente, en el que diferentes especialistas (de fabricación, diseño, ventas, calidad) trabajan de forma integral para reducir los plazos de entrega, anticipando problemas e inconvenientes inherentes al ciclo de vida del producto (Prasad, 1996; Sohlenius, 1992). Sin embargo, debido a los intrincados procesos de esta metodología, sumado a los elevados costos de diseño del producto (Korpi y Ala-Risku, 2008), es esencial formar el equipo “adecuado” de personas.

1.1. Antecedentes

En esencia, para formar equipos efectivos se necesita conocer qué características favorecen la efectividad y cómo seleccionar a sus integrantes. Estas y otras preguntas de investigación referentes a los equipos se han abordado desde la perspectiva de múltiples disciplinas como la antropología y la biología (Gold, 2004), la psicología social y organizacional (Kozlowski y Ilgen, 2006; Swezey y Salas, 1992; Katzenbach y Smith, 2015; Guzzo y Dickson, 1996; Kozlowski y Bell, 2003; Campion *et al.*, 1993), y recientemente, las ciencias de la computación (ver una descripción detallada en (Juárez *et al.*, 2021)).

En las ciencias de la computación, la investigación de equipos se enfoca en el problema combinatorio de seleccionar un conjunto de candidatos con diferentes características de manera que se maximice una cierta función de efectividad de equipo (Zakarian y Kusiak, 1999; Lappas *et al.*, 2009). A este problema y a sus variantes se les denomina: Problema de Formación de Equipo (PFE). A poco más de 20 años de su formulación, el PFE cuenta con una amplia gama de planteamientos en distintos contextos como equipos de ingeniería (Zakarian y Kusiak, 1999; Wang *et al.*, 2003; Fitzpatrick y Askin, 2005; Hlaoittinun *et al.*, 2008), equipos deportivos (Boon y Sierksma, 2003; Tavana *et al.*, 2013; Pérez-Toledano *et al.*, 2013; Ahmed *et al.*, 2013) y equipos de robots (Liemhetcharat y Veloso, 2012, 2014; Okimoto *et al.*, 2016), por mencionar algunos.

Además del contexto del equipo, los modelos de PFE también se diversifican y evolucionan en cuanto a su representación de efectividad de equipo. Los primeros modelos se construyen alrededor de la composición de habilidades técnicas para formar un equipo efectivo (Zakarian y Kusiak, 1999; Boon y Sierksma, 2003; Dereli *et al.*, 2007). Sin embargo, la efectividad de equipo no sólo está influenciada por las habilidades técnicas de sus miembros, sino también por sus interacciones sociales (West, 2012).

Los modelos más recientes incorporan sistemas de perfilado de personalidad de sus miembros. El objetivo es optimizar alguna composición de personalidades y comportamientos entre los miembros del equipo que promuevan su sinergia (Chen y Lin, 2004; Fitzpatrick y Askin, 2005). No obstante, a diferencia de un escenario general, los sistemas de perfilado fallan cuando se trata de predecir el comportamiento de un

individuo en lo particular (Shoda *et al.*, 1994).

La siguiente generación de modelos está caracterizada por tomar en cuenta la interacción de los candidatos mediante su red social. El objetivo es optimizar la composición de los miembros que defina una red efectiva de trabajo al interior del equipo (Lappas *et al.*, 2009; Kargar y An, 2011; Gajewar y Das Sarma, 2012; Selvarajah *et al.*, 2019; Chen *et al.*, 2020). Este tipo de formulación es la más común de los últimos años, y hasta ahora la más conveniente para representar parcialmente la interacción de los candidatos. Adicionalmente, existe evidencia reciente del rol fundamental que juega la interacción de los integrantes de un equipo en su efectividad.

1.2. Planteamiento del problema

En conjunto, los modelos de PFE ofrecen un esquema para formar equipos de personas de manera analítica. Adicionalmente, dan una idea de cómo representar los requisitos de una tarea en función de habilidades, capacidades y características humanas, para luego usarse como entrada en modelos de optimización. Más aún, la diversidad en los modelos de PFE son una muestra de su versatilidad y de su potencial para representar escenarios prácticos. Sin embargo, el área aún está muy lejos de determinar qué características son más adecuadas para un modelo efectivo de formación de equipo—comenzando por cómo se mide la efectividad de equipo y qué criterios cuantitativos son más adecuados para representarla. Por un lado, la diversidad en los modelos de PFE dificulta su comparación, el diseño de mejores algoritmos que los resuelvan y la generalización de conocimiento (Juárez *et al.*, 2021). Por otro lado, el entendimiento de los mecanismos y promotores de la efectividad de equipos son heredados de otras disciplinas como la psicología organizacional y social, donde esto aún se considera trabajo en progreso (Salas *et al.*, 2018).

Está claro que el PFE no es (ni debe ser) un monolito, pero también es evidente que su diversidad en contextos y modelos está interrumpiendo el desarrollo de una teoría alrededor de éste. Otra forma de verlo es que la teoría no está avanzando a la par que el estudio de sus casos particulares. Una comprensión sólida del PFE requiere alejarse del contexto inscrito en cada formulación para describir la estructura y las relaciones fundamentales de los modelos, al margen de supuestos y limitaciones particulares.

En primer instancia, el área puede beneficiarse de una clasificación taxonómica de los PFEs desde una perspectiva computacional. En vez de clasificar en cuanto a sus aplicaciones (como deporte o búsqueda y rescate), se puede clasificar en cuanto a los modelos *per se* y su relación con otros problemas combinatorios. Esto ayudaría a estudiar los PFEs desde un marco estructurado, relacionar modelos aparentemente distintos, y exponer huecos de conocimiento.

En segunda instancia, se necesita estudiar los aspectos más generales de cualquier PFE. Si realmente se pretende que el PFE sirva como un modelo en la práctica, se debe preguntar qué elementos, en cuanto a variables, funciones objetivo, o enfoque se consideran esenciales para formar equipos efectivos. Esto además es importante porque se sabe muy poco del problema en sí. Entender mejor el problema general ayudaría a establecer las consideraciones básicas (o ideales) para cualquier PFE, invariablemente de su contexto.

1.3. Objetivos

Para atender la problemática antes planteada, se proponen los siguientes objetivos general y específicos.

1.3.1. Objetivo general

Proponer una taxonomía para los problemas de formación de equipo desde una perspectiva computacional, y estudiar los elementos de un modelo general para el problema de formación de equipo.

1.3.2. Objetivos específicos

A continuación se enumeran los objetivos específicos de este trabajo de investigación.

1. Estudiar los modelos más relevantes de formación de equipo e identificar las componentes principales de cada uno.

2. Establecer criterios de clasificación de los problemas de formación de equipo basados en la estructura de sus modelos.
3. Señalar la relación de los modelos del problema de formación de equipo con otros problemas combinatorios de la literatura.
4. Identificar los elementos de un modelo general del problema de formación de equipo.
5. Proponer adaptaciones para un modelo general del problema de formación de equipo.

1.4. Contribuciones del trabajo de tesis

Una contribución principal de este trabajo de tesis está directamente ligada con una parte del objetivo general: la taxonomía. La taxonomía propuesta en esta tesis ayuda a organizar los trabajos de formación de equipos en términos del modelo que se pretende resolver, a diferencia del tipo de equipo que se quiere formar. Esto facilita la identificación de los elementos esenciales en los modelos, y su relación con otros problemas combinatorios conocidos. En general, la taxonomía propuesta contribuye con la generalización de conocimiento, la identificación de huecos de investigación, y el desarrollo de una teoría alrededor del problema. Hasta antes de este trabajo de tesis, hasta donde el autor del presente trabajo conoce, no existía una taxonomía para el problema de formación de equipo.

En esta tesis también se destaca el papel de los patrones de interacción como parte de los modelos de efectividad de equipo. Esto se complementa con la propuesta de un novedoso método no simétrico para medir y clasificar la fuerza del vínculo entre personas, dada una red social profesional. Este método revela una dimensión adicional en los datos que permite analizar estructuras de interacción más complejas, que no se había explorado anteriormente en la literatura de formación de equipos.

Finalmente, esta tesis proporciona evidencia a favor del uso de criterios de optimización basados en densidad de colaboración sobre los criterios basados en distancia de comunicación. Esto es importante puesto que existe un debate abierto en la literatura sobre cuál de las dos es más conveniente usar como función objetivo para los

problemas de formación de equipo en el contexto de las redes sociales. Más aún, estas dos familias de criterios son las más usadas para la optimización de equipos en dicho contexto.

1.5. Nota aclaratoria

Esta sección cumple un propósito de prefacio pos-introductorio; es una nota aclaratoria al contenido de este trabajo de tesis.

Primeramente, es importante señalar la distinción entre la formación de equipos como proceso y la formación de equipos como problema en las ciencias de la computación. El proceso de formación de equipos involucra un cierto dinamismo, siendo que los equipos y su ambiente se encuentran en constante evolución. En cambio, en las ciencias de la computación, el PFE está centrado principalmente en la selección de los candidatos. En otras palabras, el PFE es sólo una componente de un proceso más grande y más complejo. Tuckman (1965), por ejemplo, nombra en inglés a las etapas del ciclo de vida de los equipos como *Forming*, *Norming*, *Storming* y *Performing*. El PFE se lleva a cabo en la etapa de *Forming*, que es cuando el equipo se conoce y sus integrantes aceptan trabajar juntos.

Aún así, el PFE es un problema desafiante como para abordarse de lleno en una sola tesis. Esta tesis se enfoca en una visión teórica, fundamentada en un proceso lógico, al margen de la evidencia con la que se cuenta. Con esto se debe recalcar que el enfoque de este trabajo de tesis está basado en el estudio del PFE en un contexto más general o abstracto, sin considerar una aplicación particular del equipo y sin considerar el dinamismo al que está generalmente sujeto. Existen muchos trabajos en la literatura que estudian este problema en diversos contextos y aplicaciones, y es precisamente lo que se desea evitar en este trabajo. Esto en parte porque existe una amplia variedad modelos de formación de equipo en distintos escenarios, lo que se traduce en una falta de claridad de aquello que es esencial en los modelos, dejando un hueco en el estudio de la problemática general. Adoptar una visión más abstracta de la formación de equipos puede generar una cierta resistencia, puesto que va en contra de las experiencias personales que se tienen formando parte de algún equipo o viendo trabajar a otros equipos en contextos particulares. Sin embargo, estudiar a

los modelos de formación de equipo separados de su contexto es un primer paso, que puede ayudar a entender mejor el problema general que hay por detrás.

También es importante mencionar que los resultados que se presentan más adelante, de ninguna manera representan una verdad absoluta para este problema, ni significa que a partir de este trabajo se sabrá cómo formar equipos más efectivos en general. Por un lado, no se cuenta con datos cuantitativos suficientes, ni con la calidad suficiente, como para validar las propuestas que aquí se presentan. Y los datos que pudieran existir con esas características seguramente no están disponibles libremente para uso de la comunidad científica. Por otro lado, con los recursos con los que se cuenta, no es posible ni práctico, reunir un grupo de personas lo suficientemente grande y diverso para probar los modelos. Incluso aunque se contara con tal recurso, los controles requeridos para evaluar la efectividad de los equipos en la práctica podría introducir sesgos ocultos. Sin mencionar que una validación profunda de los modelos requeriría evaluar la efectividad sobre una batería muy extensa de tareas diversas que van más allá del alcance de este trabajo. No obstante, las grandes compañías de tecnología e información como Google probablemente cuentan con la información (o pudieran generarla) necesaria de sus empleados y de sus equipos que permita validar parte de los resultados o implicaciones de este trabajo. También, existen investigadores de la talla de Eduardo Salas cuyo conocimiento y experiencia sobre los procesos del trabajo en equipo, su entrenamiento y los promotores de su efectividad, son clave para diseñar experimentos prácticos pertinentes para la validación de ciertos elementos de este trabajo de tesis.

En resumen, esta tesis hace una contribución teórica a la investigación de equipos en lo que concierne a las ciencias de la computación. El enfoque recae en las implicaciones de calcular un modelo en comparación con otro, y en la transformación o procesamiento de datos de manera que su uso sea más efectivo, en el marco del problema de formación de equipo.

1.6. Estructura de la tesis

La estructura del resto de la tesis es la siguiente. El Capítulo 2 contiene una revisión de la literatura relacionada con los equipos y su efectividad, y las componentes

de los modelos de formación de equipo. En el Capítulo 3 se presenta una propuesta de taxonomía de los problemas de formación de equipo desde una perspectiva computacional. En el Capítulo 4 se presenta una propuesta de medida de interacción no-simétrica entre pares de personas en una red social. En el Capítulo 5 se presenta un estudio comparativo de criterios de optimización basados en distancia de comunicación versus basados en densidad de colaboración, desde una perspectiva multiobjetivo. Finalmente, en el Capítulo 6 se presentan las conclusiones generales de este trabajo de tesis, así como las limitantes y perspectivas para trabajo a futuro.

Capítulo 2. Revisión de la literatura

En este capítulo se presenta una revisión de la literatura sobre la investigación de equipos en un orden lógico. Se comienza hablando sobre los equipos en general y sobre la efectividad de equipo. Luego se discute brevemente la literatura del problema de formación de equipo. Finalmente, se describen las componentes de los modelos de los problemas de formación de equipo y de dónde provienen sus variables de entrada.

2.1. Equipos y su efectividad

Los equipos de personas son cada vez más socorridos en las organizaciones de todo el mundo como unidad básica de trabajo (Salas *et al.*, 2008). Éstos no sólo ofrecen respuestas ágiles y flexibles a tareas dinámicas y complejas (Swezey y Salas, 1992; Kozlowski y Bell, 2003), sino que también permiten alcanzar objetivos que están fuera del alcance de cualquier individuo actuando solo (Hackman, 2011).

De acuerdo con Swezey y Salas (1992), un equipo en general se define como un conjunto distinguible de dos o más personas que interactúan de forma dinámica, independiente y adaptativa hacia una meta, misión u objetivo en común, a quienes se les han asignado funciones o roles específicos para desempeñar, y que tienen una vida útil limitada. Esta definición es esencial para áreas como la investigación de equipos, así como para delimitar al problema de formación de equipo de otros problemas similares.

Uno de los objetivos más codiciados en la investigación de equipos es contar con un conjunto de herramientas que permitan medir la efectividad de un equipo de manera fiable (Salas *et al.*, 2018). Sin embargo, determinar la eficacia de un equipo es una tarea difícil debido a los numerosos factores que influyen en ésta, como las habilidades técnicas, el trabajo en equipo o los aspectos sociales. Además, estos factores influyen en mayor o menor medida la tarea o el tipo de equipo (Salas *et al.*, 2008). Es por ello que muchos expertos consideran que la definición de efectividad de equipo es amplia y carece de la precisión de un constructo teórico (Guzzo y Dickson, 1996; Kozlowski y Bell, 2003; Kozlowski y Ilgen, 2006).

La investigación sobre la efectividad de los equipos abarca más de medio siglo, la

mayor parte de ella en el ámbito de la psicología social y la psicología organizacional (Witte y Davis, 1996; Kozlowski y Ilgen, 2006). Hoy en día, la cantidad y el tipo de datos que se pueden recopilar y analizar abre la oportunidad de observar estructuras sociales dinámicas—como los equipos—fuera de los límites de una teoría preconcebida o un escenario controlado (Lazer *et al.*, 2009; Onnela *et al.*, 2014; Kozlowski *et al.*, 2016). Entre los hallazgos de estos estudios se destaca uno en particular: la seguridad psicológica en un equipo es uno de los predictores más importantes de su éxito. La seguridad psicológica en un equipo se define como la creencia compartida de que el equipo representa un espacio seguro para la toma de riesgos interpersonales (Edmondson, 1999).

Woolley *et al.* (2010) examinaron el rendimiento de 152 equipos, formados por entre dos y cinco miembros, a lo largo de un conjunto de tareas (Woolley *et al.*, 2010). En su análisis observaron que el rendimiento de un equipo en diversas tareas es predecible mediante un factor denominado por ellos como inteligencia colectiva del equipo. Análogo a la inteligencia individual, la inteligencia colectiva la definen como la capacidad general de un grupo para realizar una amplia variedad de tareas. Adicionalmente, su análisis mostró que la inteligencia colectiva se correlaciona significativamente con la sensibilidad social media y el comportamiento de toma de turnos de conversación en el equipo. Estas dos características están directamente relacionadas con la seguridad psicológica en un equipo.

Pentland (2012), en un artículo de la revista *Harvard Business Review*, resumió su investigación sobre la interacción de los equipos en el Laboratorio de Dinámica Humana del MIT (Pentland, 2012). Pentland y sus colegas analizaron el comportamiento comunicativo de los equipos en departamentos de varios sectores, utilizando carnets electrónicos para seguir las interacciones de las personas en el lugar de trabajo. Los datos mostraron que los patrones de comunicación desempeñan un papel fundamental en la eficacia de un equipo. En general, los equipos etiquetados como exitosos presentaban tiempos de conversación más o menos iguales entre sus miembros durante las conversaciones de equipo, interacciones de comunicación recíproca entre cada par de miembros y comunicación con otros equipos.

Rozovsky (2015), en una entrevista con el *New York Times*, habló de su papel principal en el Proyecto Aristóteles—un esfuerzo de Google por encontrar las caracterís-

ticas que explican la eficacia de los equipos (Duhigg, 2016). Rozovsky y su equipo recogieron y analizaron datos de más de 250 atributos de 180 equipos a lo largo de tres años. Sorprendentemente, los resultados revelaron que la seguridad psicológica destaca consistentemente en los equipos de alto rendimiento (Rozovsky, 2015). Particularmente, la seguridad psicológica en el equipo promueve un entorno seguro para que los miembros del equipo asuman riesgos y compartan sus ideas y preocupaciones abiertamente. Aunque los comentarios de Razovsky en la entrevista son en su mayoría anecdóticos, gran parte de sus conclusiones coinciden con hallazgos independientes de la investigación del comportamiento grupal sobre la seguridad psicológica en los equipos (Edmondson, 1999; Bradley *et al.*, 2012; Kostopoulos y Bozionelos, 2011).

Colectivamente, estos estudios señalan la importancia que tiene la seguridad psicológica y el rol que juega como promotor de la efectividad en los equipos. Adicionalmente, indican que los modelos de formación de equipo requieren un mayor énfasis en representar esta característica ya sea de manera directa o indirecta. Estos trabajos además son la sobrecama de la propuesta en el Capítulo 4.

2.2. Problema de formación de equipo

El Problema de Formación de Equipo (PFE) atiende la siguiente pregunta: dado un conjunto de candidatos y una función de efectividad de equipo, ¿cuáles de los candidatos formarían un equipo de máxima efectividad para una tarea en particular? Este es un problema de optimización combinatoria de reciente formulación que ha capturado el interés de comunidades de investigación de distintas áreas de ingeniería, matemáticas y ciencias de la computación. Por el momento, se evita formalizar la definición del problema puesto que esto una de las dificultades de esta tesis como se discute a continuación.

La literatura de PFE se agrupa principalmente en dos campos de interés: la Investigación de Operaciones (IO) y la Minería de Datos (MD), cada uno de los cuales concibe y modela el problema según su fuente de información. En IO, el problema se alimenta de las solicitudes y las demandas de las organizaciones, principalmente de la industria (Zakarian y Kusiak, 1999). Se considera el problema de identificar un equipo con la “mejor” coincidencia entre los candidatos y las posiciones en el equipo. En MD, en

cambio, el problema se rige por los datos de las redes sociales, principalmente las redes profesionales (Lappas *et al.*, 2009). Se considera el problema de identificar a los candidatos con las interacciones más estrechas, cuyo conjunto de habilidades sea igual al conjunto de habilidades requeridas por la tarea. Independientemente del enfoque, ambos requieren explorar sobre el número total de equipos posibles, lo que se relaciona con problemas combinatorios complejos que han sido probados o sugeridos como problemas NP-difíciles (Juárez *et al.*, 2021).

Otros problemas similares a la formación de equipos pueden encontrarse en áreas como los Sistemas Multiagente (SMA). Sin embargo, el enfoque de SMA suele carecer de la noción de tarea y roles de equipo, característica de la definición de equipo (Adams *et al.*, 2011; Fatima y Wooldridge, 2019). Además, los agentes suelen decidir si forman o no parte de un grupo, mientras que el PFE supone que la decisión está centralizada y no contempla las decisiones individuales de los candidatos (Adams *et al.*, 2011). Las formulaciones multiagente requieren otro tipo de consideraciones relacionadas con la teoría de juegos y la teoría de subastas. Estas formulaciones quedan fuera del alcance de esta tesis.

El PFE, a diferencia de otros problemas combinatorios clásicos como el problema del vendedor viajero¹, carece de un modelo básico que enmarque concretamente el problema. En cambio, invariablemente de los dos enfoques distinguibles, la literatura referente al PFE se encuentra como una colección de problemas en distintos contextos. Por ejemplo, la primer formulación del Problema de Formación de Equipos, atribuido a Zakarian y Kusiak (1999), ocurre en el contexto de la industria automotriz. En su trabajo, Zakarian y Kusiak (1999) intentan formar tantos equipos de ingeniería como requerimientos hay en el diseño de un vehículo, maximizando la idoneidad total de candidatos a equipos de manera que se cubran los requerimientos más importantes. Por su parte, en el contexto deportivo, Boon y Sierksma (2003) presentan una formulación del PFE para encontrar alineaciones efectivas para equipos de fútbol y voleibol. En el contexto de la robótica, Liemhetcharat y Veloso (2014) introducen un modelo de PFE que considera formar un equipo efectivo a partir de robots heterogéneos para desempeñar en tareas de búsqueda y rescate. Consecuentemente, se puede generalizar muy poco a partir de estos y otros trabajos en lo individual.

¹También conocido como problema del vendedor ambulante, problema del agente viajero o problema del viajante

Más aún, el PFE cuenta con múltiples variantes. Por ejemplo, Chen y Lin (2004) abordan el problema de seleccionar candidatos de varios departamentos funcionales de una organización para formar un equipo con conocimientos multifuncionales, capacidad de trabajo en equipo y una buena relación de trabajo. Baykasoğlu *et al.* (2007) por su parte, proponen asignar candidatos a equipos de proyecto en función de su adecuación a los requisitos de los proyectos, sujeto a limitaciones de calendario, presupuesto y relaciones interpersonales. Wi *et al.* (2009) exploran asignar candidatos a equipos de proyecto en función de su conocimiento personal y del conocimiento ampliado por la red de expertos que conoce. Éstas y otras variantes se resumen en la Tabla 1.

Tabla 1. Tabla de variantes del Problema de Formación de Equipo.

Variante	Referencia
Sobre/sub-cualificación	(Hlaoittinun <i>et al.</i> , 2007; Rostami y Neshati, 2019)
Presupuesto y ganancia	(Baykasoğlu <i>et al.</i> , 2007; Rangapuram <i>et al.</i> , 2013; Ahmed <i>et al.</i> , 2013) (Tavana <i>et al.</i> , 2013; Kargar <i>et al.</i> , 2013; Liu <i>et al.</i> , 2015) (Golshan <i>et al.</i> , 2014)
Tamaño del equipo (fijo)	(Boon y Sierksma, 2003; Van de Water <i>et al.</i> , 2007; Fitzpatrick y Askin, 2005) (Hlaoittinun <i>et al.</i> , 2007; Ahmed <i>et al.</i> , 2013; Feng <i>et al.</i> , 2010) (Zhang y Zhang, 2013)
Tamaño del equipo (acotado)	(Zakarian y Kusiak, 1999; Rangapuram <i>et al.</i> , 2013; Strnad y Guid, 2010)
Carga de trabajo y disponibilidad	(Wang <i>et al.</i> , 2003; Abdelsalam, 2009; Baykasoğlu <i>et al.</i> , 2007) (Anagnostopoulos <i>et al.</i> , 2010, 2012; Huang <i>et al.</i> , 2017) (Durfee <i>et al.</i> , 2014)
Robusto y recuperable	(Okimoto <i>et al.</i> , 2015, 2016; Demirović <i>et al.</i> , 2018)
Niveles dinámicos de habilidades	(Hlaoittinun <i>et al.</i> , 2008)
Niveles dinámicos de interacción	(Awal y Bharadwaj, 2014; Wi <i>et al.</i> , 2009)
Candidatos con una habilidad	(Zakarian y Kusiak, 1999; Fitzpatrick y Askin, 2005; Feng <i>et al.</i> , 2010)
Candidatos con múltiples habilidades	(Lappas <i>et al.</i> , 2009; Chen y Lin, 2004; Wi <i>et al.</i> , 2009) (Kargar y An, 2011; Tseng <i>et al.</i> , 2004; Yaakob y Kawata, 1999) (Baykasoğlu <i>et al.</i> , 2007)
Formación de un equipo	(Chen y Lin, 2004; Wi <i>et al.</i> , 2009; Lappas <i>et al.</i> , 2009) (Kargar y An, 2011; Boon y Sierksma, 2003)
Formación de muchos equipos	(Zakarian y Kusiak, 1999; Fitzpatrick y Askin, 2005; Baykasoğlu <i>et al.</i> , 2007) (Tseng <i>et al.</i> , 2004; Zhou <i>et al.</i> , 2017; Okimoto <i>et al.</i> , 2016) (Bahargam <i>et al.</i> , 2019)
Pertenencia a un equipo	(Chen y Lin, 2004; Van de Water <i>et al.</i> , 2007; Fitzpatrick y Askin, 2005) (Lappas <i>et al.</i> , 2009; Feng <i>et al.</i> , 2010)
Pertenencia a múltiples equipos	(Zakarian y Kusiak, 1999; Baykasoğlu <i>et al.</i> , 2007; Yaakob y Kawata, 1999) (Wang <i>et al.</i> , 2003; Abdelsalam, 2009; Strnad y Guid, 2010)

A pesar del trabajo que se ha hecho, se conoce muy poco del problema en sí. Las

variaciones en el contexto, las componentes del modelo y sus variables desdibujan las semejanzas de los PFEs. Por ejemplo, los modelos (o porciones de modelo) de formación de equipos propuestos en (Boon y Sierksma, 2003; Hlaioittinun *et al.*, 2007; Durfee *et al.*, 2014) son esencialmente equivalentes, y cada uno a su vez, estructuralmente similar al problema de asignación clásico—que además se sabe que se puede resolver en tiempo polinomial (Kuhn, 1955). Para estudiar el problema como tal, es necesario alejarse del contexto inscrito en cada formulación para describir la estructura y las relaciones fundamentales de los modelos, al margen de supuestos y limitaciones particulares.

Si estudiamos los PFEs en función de su estructura, entonces se podría comenzar a entender qué partes de los modelos son esenciales y qué partes son secundarias, como restricciones de presupuesto o ubicación geográfica. Como consecuencia, uno se puede enfocar entonces en estudiar los elementos de un PFE más general.

2.3. Componentes de los modelos y sus variables

Para comprender mejor los modelos de formación de equipo, probablemente es más conveniente comenzar por las componentes que los conforman y sus variables de entrada. Existen al menos tres componentes principales que se distinguen en los modelos de forma consistente. De acuerdo con Chen y Lin (2004), estas componentes pueden clasificarse en habilidades técnicas, habilidades de trabajo en equipo y rasgos personales. Estas tres componentes también pueden o no estar presentes de manera conjunta en un mismo modelo.

La primer componente, las habilidades técnicas, a veces denominadas habilidades duras, suelen ser habilidades cuantificables relacionadas con tareas o disciplinas específicas. Estas habilidades pueden ser físicas o cognitivas, por ejemplo: memoria, razonamiento deductivo, expresión oral, orientación espacial, velocidad de reacción, destreza de los dedos, fuerza, equilibrio, estamina, por mencionar algunos.

La segunda y tercer componente, las habilidades de trabajo en equipo y los rasgos personales, comúnmente denominadas habilidades blandas, son las características interpersonales e intrapersonales que no son tan fácilmente cuantificables, como la comunicación interpersonal o la conducta instintiva.

2.3.1. Habilidades técnicas

Muchos modelos, especialmente aquellos en investigación de operaciones, recaen fuertemente en su componente de habilidad técnica para representar la efectividad de un equipo. Las habilidades técnicas de cada candidato se califican con el fin de determinar su puntuación de idoneidad con respecto a los requisitos técnicos de un equipo o de una posición en el equipo, según la tarea a realizar.

En general, dados n candidatos y ℓ puestos de equipo, una componente de habilidades técnicas de un modelo puede encontrarse de la forma dada por la ecuación (1).

$$\text{Total de habilidad técnica en un equipo} = \sum_{i=1}^n \sum_{k=1}^{\ell} w_{ik} x_{ik}, \quad (1)$$

donde $x_{ik} \in \{0, 1\}$ representa la variable de decisión de un candidato i que ocupa una posición k en el equipo². Mientras que w_{ik} es un valor no-negativo que representa la idoneidad del candidato i en ocupar la posición k en consideración a sus habilidades técnicas y los requerimientos técnicos. La entrada entonces puede verse como una matriz $W = (w_{ik})_{n \times \ell}$.

Dependiendo del contexto de la tarea, de las herramientas de evaluación disponibles y de los evaluadores mismos, las puntuaciones de idoneidad de las habilidades técnicas de los candidatos pueden representarse como números reales o enteros (Zakarian y Kusiak, 1999; Boon y Sierksma, 2003) (incluso binarios (Lappas *et al.*, 2009)), números difusos (Yaakob y Kawata, 1999; Baykasoğlu *et al.*, 2007), o incluso variables probabilísticas (Liemhetcharat y Veloso, 2012).

2.3.1.1. Matriz de incidencia candidato–posición

En algunos casos, los responsables de la toma de decisiones, como el departamento de gestión técnica, se encargan directamente de puntuar las aptitudes generales del candidato en relación con los puestos del equipo que se va a formar. Estas evaluaciones se realizan de forma individual, asignando un valor numérico a cada asignación de candidato-puesto. Suponga por ejemplo, que el candidato i recibe una puntuación de $a_{i1} = 0.8$ en una escala continua de 0 a 1, para el puesto 1, y de $a_{i2} = 0.6$ en la

²Una posición también puede ser un rol o un cargo en el equipo.

misma escala, para el puesto 2, y así sucesivamente para cada candidato y puesto hasta que se obtiene una matriz $A = (a_{ik})_{n \times \ell} \in \mathbb{R}^{n \times \ell}$. Este es un caso conveniente ya que se puede definir directamente la matriz $W = (w_{ik} = a_{ik})_{n \times \ell} = A$ como la entrada de la puntuación de idoneidad para algunos modelos de PFE.

Cuando existen múltiples criterios en juego, puntuar directamente la idoneidad de cada candidato por cada posición en el equipo puede resultar más complicado. Una alternativa es solicitar al departamento técnico dos evaluaciones por separado. En primer lugar, puntuar a cada candidato en relación con los criterios técnicos importantes, de modo que se obtenga una matriz candidato–criterios $A \in \mathbb{R}^{n \times p}$. En segundo lugar, anotar un peso prioritario en cada criterio en relación con cada posición en el equipo, de modo que se obtenga una matriz de puntuación posición–criterios $B \in \mathbb{R}^{p \times \ell}$. Así, una matriz de puntuación de idoneidad candidato–posición W puede estar dada por el producto de las otras dos matrices como $W = AB \in \mathbb{R}^{n \times \ell}$. Este enfoque se usa en (Boon y Sierksma, 2003).

Por otro lado, si los valores de ambas matrices $A \in \mathbb{R}^{n \times p}$ y $B \in \mathbb{R}^{p \times \ell}$ están en la misma escala, y los valores de la matriz de puntuación de criterios de posición B representan la calificación mínima requerida. Se puede definir $W = (w_{ik})_{n \times \ell}$ con $w_{ik} = \sum_{o=1}^p \min(0, b_{ok} - a_{io})$, de manera que la matriz W representa la desviación de habilidad técnica candidato–posición. Este enfoque se usa en (Abdelsalam, 2009).

2.3.1.2. Variables lingüísticas y números difusos

Debido a la complejidad inherente de evaluar los atributos humanos, algunas veces se prefiere calificar a los candidatos usando un lenguaje natural (o artificial) en lugar de valores numéricos (Herrera *et al.*, 2000). En otras palabras, se pueden utilizar variables lingüísticas como b = malo, s = suficiente, f = regular, g = bueno, y vg = muy bueno, para describir las evaluaciones del departamento técnico a las habilidades del candidato o a los requisitos mínimos de habilidades de los puestos de trabajo como en (Yaakob y Kawata, 1999; Baykasoğlu *et al.*, 2007).

Estas variables lingüísticas se asignan, a su vez, a números difusos, que se utilizan para describir valores de criterios dudosos (Dubois y Prade, 1978) en el contexto de la teoría de conjuntos difusos (Zadeh, 1965). La forma en la que se utilizan es la

siguiente: en la escala de $\{b,s,f,g,vg\}$, suponga que el candidato i se evalúa con un $\tilde{a}_{i1} = f$, para la posición 1, y un $\tilde{a}_{i2} = g$, para la posición 2, y de manera similar para cada candidato y posición. Suponga también que a cada variable lingüística le corresponden los siguientes números difusos triangulares $b:(0,0.3,0.5)$, $s:(0.4,0.5,0.6)$, $f:(0.5,0.7,0.8)$, $g:(0.7,0.8,0.9)$, $vg:(0.8,1,1)$ cuya función de pertenencia se representa en la Figura 1. Sustituyendo cada variable lingüística por su correspondiente número difuso, se obtiene una matriz de números difusos triangulares $\tilde{A} = (\tilde{a}_{ik})_{n \times \ell}$. Cuando hay un único criterio a considerar, la matriz $\tilde{W} = (\tilde{w}_{ik} = \tilde{a}_{ik})_{n \times \ell} = \tilde{A}$ se define directamente. Finalmente, se transforma cada número difuso $\tilde{w}_{ik} = (y, x, z)$ en un número real $w_{ik} = x$ tomando su valor máximo x , o mediante alguna otra técnica de *defuzzificación* adecuada (Van Leekwijck y Kerre, 1999). Una vez que la matriz difusa $\tilde{W}$ se transforma en una matriz de números reales W , se utiliza como entrada para los modelos de PFE.

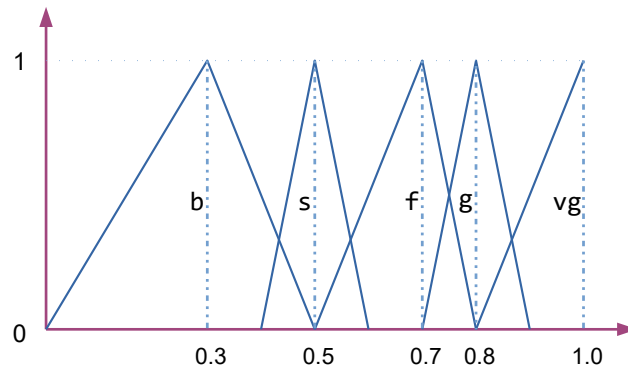


Figura 1. Funciones de pertenencia de variables lingüísticas.

De manera similar a la contraparte no difusa descrita en la subsección anterior, si se da una matriz difusa triangular $\tilde{A} = (\tilde{a}_{ik})_{n \times p}$ que contiene las evaluaciones de las habilidades de los candidatos y una matriz difusa triangular $\tilde{B} = (\tilde{b}_{ik})_{p \times \ell}$ que contiene las evaluaciones de los requisitos mínimos por puesto de equipo, entonces se define el valor difuso de idoneidad candidato–posición como $\tilde{w}_{ik} = \tilde{a}_{i1} \otimes \tilde{b}_{1k} \oplus \tilde{a}_{i2} \otimes \tilde{b}_{2k} \oplus \dots \oplus \tilde{a}_{ip} \otimes \tilde{b}_{pk}$; donde $\otimes$ y $\oplus$ son operadores difusos para la multiplicación y la suma (Dubois y Prade, 1978), respectivamente. La *defuzzificación* se puede realizar posteriormente para convertir $\tilde{W}$ a una matriz $W \in \mathbb{R}^{n \times \ell}$ y usarla como entrada a los modelos de PFE.

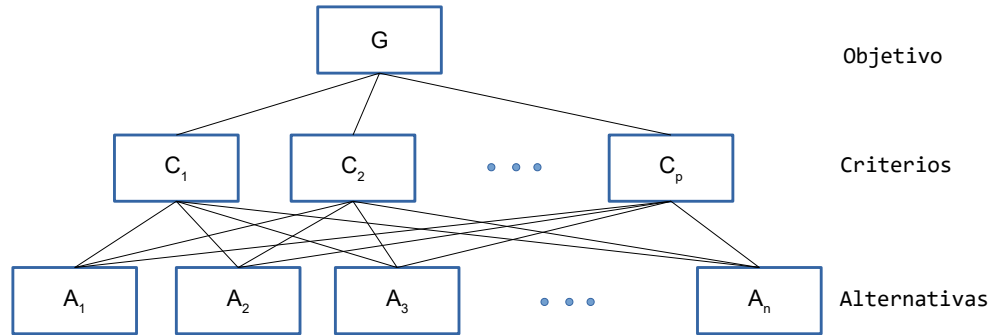


Figura 2. Ejemplo de una estructura jerárquica genérica del Proceso Jerárquico Analítico. Existe un peso relativo de importancia asociado a cada conexión.

2.3.1.3. Proceso jerárquico analítico

En otros casos, los responsables de la toma de decisiones no pueden realizar una evaluación directa de la posición del candidato. En su lugar, se utilizan técnicas de toma de decisiones multicriterio como el Proceso Jerárquico Analítico (PJA) (Saaty, 1990). El PJA sirve para evaluar la prioridad de las alternativas con respecto a un objetivo, en este caso, la idoneidad de los candidatos con respecto a un puesto de equipo. Este proceso relaciona jerárquicamente un objetivo y las alternativas a través de una capa intermedia de criterios relevantes, como se muestra en la Figura 2. Luego, los evaluadores técnicos establecen la importancia relativa de cada elemento mediante comparación por pares en el mismo nivel jerárquico, siguiendo un cuestionario especializado. Para los detalles de este proceso, consultar (Saaty, 1990).

Para la formación de equipos, cada posición del equipo (el objetivo) se descompone en habilidades técnicas (los criterios) que se requieren en los candidatos (las alternativas). Esto demanda una estructura jerárquica (Figura 2) por cada posición. Entonces, para la posición de equipo k , se obtiene el peso relativo b_{ok} de cada habilidad técnica o . Luego, se obtiene el peso relativo a_{io} de cada candidato i , con respecto a cada habilidad técnica o . Finalmente, la puntuación de idoneidad w_{ik} de cada candidato i con respecto a cada posición k se define como $w_{ik} = \sum_{o=1}^p a_{io} \cdot b_{ok}$. Esta técnica para el PFE se usa en (Zakarian y Kusiak, 1999; Tseng et al., 2004; Zhang y Zhang, 2013).

2.3.2. Rasgos de personalidad

Los rasgos personales son atributos que captan el impulso y los motivos del individuo. Algunos investigadores sugieren que la composición de la personalidad del equipo puede ser un predictor sólido de la eficacia del equipo (Kozlowski y Bell, 2003), la tan deseada herramienta de un PFE. Una forma de evaluar los rasgos personales es mediante el uso de pruebas psicométricas de personalidad, como el test del Indicador de Tipo Myers-Briggs (MBTI) (Myers *et al.*, 1998), ampliamente utilizado en las formulaciones de PFE (Chen y Lin, 2004; Abdelsalam, 2009; Zhang y Zhang, 2013; Xu *et al.*, 2012).

En general, dados n candidatos y ℓ puestos de equipo, una componente de rasgos personales de un modelo tiene la forma dada por la ecuación (2).

$$\text{Total de afinidad en un equipo} = \sum_{i=1}^{n-1} \sum_{i'=1}^n \sum_{k=1}^{\ell-1} \sum_{k'=1}^{\ell} r_{ii'} x_{ik} x_{i'k'}, \quad (2)$$

donde $x_{ik} \in \{0, 1\}$ ($x_{i'k'} \in \{0, 1\}$) representa la variable de decisión de un candidato i (i') que ocupa una posición k (k') en el equipo. Mientras que $r_{ii'}$ es un valor que representa la afinidad del candidato i con respecto al candidato i' en consideración de ambos rasgos de personalidad. La entrada entonces puede verse como una matriz cuadrada $R = (r_{ii'})_{n \times n}$. Esta matriz suele ser simétrica debido a que representa la colisión o el complementeo de una tipificación de la personalidad con otra.

2.3.2.1. El indicador Myers-Briggs

El test MBTI, propuesto por Myers y Briggs (Myers *et al.*, 1998), es una aplicación de la teoría de los tipos psicológicos de Jung (Jung, 2016) en la que se asigna a una persona uno de los 16 tipos de personalidad, derivados de cuatro rasgos psicológicos categóricos básicos, cada uno con dos preferencias opuestas. Las cuatro dicotomías son: (1) Extrovertido o Introvertido, en lo que respecta al foco de atención; (2) Sensación o Intuición, en lo que respecta a la adquisición de información; (3) Pensamiento o Sentimiento, en lo que respecta a la toma de decisiones; y (4) Juicio o Percepción, en lo que respecta a la conexión con el mundo exterior. Cualquier combinación de estas cuatro preferencias, normalmente identificadas con letras mayúsculas, define un

tipo de personalidad; por ejemplo, ENTJ denota un tipo de personalidad Extrovertida, iNtuitiva, Pensante y Juzgadora. Además, la teoría establece que existen relaciones positivas (+), neutras (○) y negativas (−) entre los pares de tipos de personalidad que se resumen en la Tabla 2. Para incorporar esta prueba en los modelos de PFE, algunos trabajos como (Chen y Lin, 2004; Zhang y Zhang, 2013) han decidido asignar valores numéricos a las relaciones de personalidad, de manera que los tipos de personalidad de dos individuos pueden recibir una puntuación para evaluar su compatibilidad. Estas puntuaciones se deciden de forma que las combinaciones de todas las relaciones de tipo positivo, como ESTJ y ENFJ, tengan la puntuación más alta, mientras que las combinaciones del caso inverso (todas las relaciones de tipo negativo), como ISTJ e ISTP, tengan la puntuación de compatibilidad más baja. La Tabla 3 muestra las puntuaciones normalizadas para todas las posibles combinaciones de tipos de personalidad, tal y como aparece en (Chen y Lin, 2004; Zhang y Zhang, 2013). Entonces, una vez tipificada la personalidad de cada candidato, se construye la matriz de afinidad $R = (r_{ii'})_{n \times n}$ utilizando las puntuaciones de la Tabla 3, donde $r_{ii'}$ representa el valor correspondiente de compatibilidad de la tipificación del candidato i y del candidato i' . Finalmente, esta matriz se utiliza como entrada en la componente representada por la ecuación (2).

Tabla 2. Tabla de compatibilidad en cuanto a pares de tipos de personalidad del MBTI.

	E	I		S	N		T	F		J	P
E	+	○		○	+		○	+		+	−
I	○	−		+	○		+	○		−	+

Tabla 3. Tabla normalizada de puntuaciones de compatibilidad entre las tipologías del MBTI de acuerdo a Chen y Lin (2004); Zhang y Zhang (2013).

	ESTJ	ESTP	ESFJ	ESFP	ENTJ	ENTP	ENFJ	ENFP	ISTJ	ISTP	ISFJ	ISFP	INTJ	INTP	INFJ	INFP
ESTJ	0.67															
ESTP	0.33	0.67														
ESFJ	0.83	0.50	0.67													
ESFP	0.50	0.83	0.33	0.67												
ENTJ	0.83	0.50	1.00	0.67	0.67											
ENTP	0.50	0.83	0.67	1.00	0.33	0.67										
ENFJ	1.00	0.67	0.83	0.50	0.83	0.50	0.67									
ENFP	0.67	1.00	0.50	0.83	0.50	0.83	0.33	0.67								
ISTJ	0.50	0.17	0.67	0.33	0.67	0.33	0.83	0.50	0.33							
ISTP	0.17	0.50	0.33	0.67	0.33	0.67	0.50	0.83	0.00	0.33						
ISFJ	0.67	0.33	0.50	0.17	0.83	0.50	0.67	0.33	0.50	0.17	0.33					
ISFP	0.33	0.67	0.17	0.50	0.50	0.83	0.33	0.67	0.17	0.50	0.00	0.33				
INTJ	0.67	0.33	0.83	0.50	0.50	0.17	0.67	0.33	0.50	0.17	0.67	0.33	0.33			
INTP	0.33	0.67	0.50	0.83	0.17	0.50	0.33	0.67	0.17	0.50	0.33	0.67	0.00	0.33		
INFJ	0.83	0.50	0.67	0.33	0.67	0.33	0.50	0.17	0.67	0.33	0.50	0.17	0.50	0.17	0.33	
INFP	0.50	0.83	0.33	0.67	0.33	0.67	0.17	0.50	0.33	0.67	0.17	0.50	0.17	0.50	0.00	0.33

2.3.2.2. Otros indicadores

Existen muchos otros modelos de tipificación de personalidad como el Big-Five (John y Srivastava, 1999) que ofrece una batería de instrumentos para medir cinco dimensiones de la personalidad (apertura a la experiencia, escrupulosidad, extroversión, amabilidad, y neurotismo), el 16PF (Cattell *et al.*, 1970), basado en dieciséis elementos estructurales de la personalidad (como la calidez, el razonamiento, la estabilidad emocional, la dominancia, la vivacidad, por mencionar algunos), o CliftonStrengths (Buckingham y Clifton, 2001) que se basa en la permutación de 34 atributos en orden de importancia que tipifican a una persona. El problema de estos indicadores o herramientas de tipificación, incluido el MBTI, es que no está claro cómo se complementan o cómo colisionan estos rasgos entre ellos, especialmente cuando es necesario cuantificarlo para generar la matriz *R*. Adicionalmente, se sabe que estos sistemas de tipificación pueden predecir el comportamiento de una persona en lo general, pero no en lo particular³ (Shoda *et al.*, 1994).

2.3.3. Habilidades de trabajo en equipo

Las habilidades de trabajo en equipo son un conjunto de atributos genéricos de los candidatos que promueven el trabajo efectivo como equipo. De acuerdo con Morgan Jr *et al.* (1986), las habilidades de trabajo en equipo engloban la coordinación, la adaptación a las distintas demandas de la situación, la comunicación efectiva, el comportamiento compensatorio, la supervisión del rendimiento mutuo y el acto de dar y recibir retroalimentación. Chen y Lin (2004), en el contexto de los PFE, resumen los atributos de las habilidades de trabajo en equipo como la experiencia en el trabajo en equipo, las habilidades de comunicación y la flexibilidad en la asignación del trabajo.

Varios de estos atributos se pueden evaluar implícita o explícitamente con herramientas como la evaluación de desempeño de 360°, los mapas de comunicación, y la consulta de procesos de equipo (Recardo *et al.*, 1996). Otros se pueden obtener mediante datos históricos de la participación relacionada con el trabajo en equipo

³El comportamiento general de una persona se refiere al comportamiento promedio esperado invariablemente del escenario, mientras que el comportamiento particular se refiere al comportamiento de una persona dado un escenario y un contexto en particular. Por ejemplo, uno no suele actuar de la misma manera en público que en privado.

dentro de una organización como la comunicación por correo electrónico (Van Staden y Van der Poel, 2016).

Los atributos más importantes desde el punto de vista de los modelos tienen que ver con aquellos que se representan mediante una red⁴, como la comunicación, colaboración o influencia. En general, dados n candidatos y ℓ puestos de equipo, una componente de comunicación de equipo se ve de la forma expresada por la ecuación (3).

$$\text{Costo de comunicación en un equipo} = \sum_{i=1}^{n-1} \sum_{i'=1}^n \sum_{k=1}^{\ell-1} \sum_{k'=1}^{\ell} d_{ii'} x_{ik} x_{i'k'}, \quad (3)$$

donde $x_{ik} \in \{0, 1\}$ ($x_{i'k'} \in \{0, 1\}$) representa la variable de decisión de un candidato i (i') que ocupa una posición k (k') en el equipo. Mientras que $d_{ii'}$ representa la distancia de comunicación del candidato i con respecto al candidato i' . La entrada entonces puede verse como una matriz cuadrada $D = (d_{ii'})_{n \times n}$. A diferencia de la componente de rasgos personales, esta matriz no necesariamente es simétrica.

2.3.3.1. Comunicación mediante red social

Algunos PFE, especialmente aquellos en minería de datos, se limitan a aquellas habilidades de trabajo en equipo relativas a comunicación o colaboración que se puedan obtener de redes sociales profesionales. Estas redes se modelan mediante un grafo social, donde los nodos representan candidatos y las aristas capturan la interacción de comunicación (o colaboración) entre pares de candidatos (Lappas *et al.*, 2009).

Por ejemplo, varios trabajos utilizan el conjunto de datos de un sistema de información de bibliografía en ciencias de la computación, conocido como DBLP, para generar el grafo social que sirve de entrada al problema. En este tipo de grafo social, los nodos representan investigadores, y las aristas representan una conexión entre dos investigadores. Existe una arista entre dos candidatos siempre y cuando ambos aparecen como coautores en al menos un cierto número de publicaciones. Esta representación por si sola encierra alguna información sobre la capacidad de trabajo en equipo entre candidatos, pero puede ampliarse considerando pesos en las aristas para representar costos de comunicación.

⁴Los atributos de habilidades de trabajo en equipo que son meramente individuales, pueden modelarse como se ve en la sección de habilidades técnicas, ya sea que tengan un impacto particular con cada posición o no.

Dado un grafo social, la distancia de comunicación $d_{ii'}$ de dos candidatos i e i' se define mediante modelos sencillos como el inverso del número total de proyectos con participación mutua como $d_{ii'} = \frac{1}{|P_i \cap P_{i'}|}$, o mediante su distancia de Jaccard como $d_{ii'} = 1 - \frac{|P_i \cap P_{i'}|}{|P_i \cup P_{i'}|}$ (Levandowsky y Winter, 1971), donde P_x es el conjunto de proyectos de $x = \{i, i'\}$. Para el ejemplo anterior, los proyectos representan publicaciones, de manera que $P_i \cap P_{i'}$ representa el conjunto de publicaciones con coautoría de i e i' . Otros trabajos como (Wi et al., 2009; Awal y Bharadwaj, 2014), sugieren además considerar alguna función de decaimiento para modelar la disminución de conectividad entre dos candidatos conforme transcurre el tiempo desde su última colaboración. Así, la función de distancia $d_{ii'}$ entre candidatos i, i' está dada como la suma del costo de comunicación de las aristas del camino más corto entre ambos (Lappas et al., 2009; Wang et al., 2015).

Finalmente, las componentes de trabajo de equipo de este tipo de modelos suelen ser mucho más variadas de lo que se presenta en la ecuación (3). Por ejemplo, para el costo de comunicación de equipo suelen considerar medidas basadas en grafos, como el costo del árbol de expansión mínimo o el diámetro del subgrafo inducido por los candidatos de un equipo, entre otros (Lappas et al., 2009; Wang et al., 2015).

2.4. Discusión

Hay dos puntos clave en este capítulo. El primero es que la definición de efectividad de equipo es muy amplia y los factores que influyen en ésta son múltiples. Algunos estudios recientes han señalado a la seguridad psicológica como principal predictor de equipos efectivos. No obstante, la seguridad psicológica no es algo que se pueda medir *a priori* ni de forma individual, sino que es una característica en algunos equipos. Esto implica que nuevamente se requiere un equipo ya formado para evaluar esta característica, cuando la problemática está centrada justamente en una etapa previa: en la selección de candidatos. Alternativamente, la seguridad psicológica está correlacionada con patrones de comunicación uniformes dentro del equipo. Por lo que, dada una red social, se pueden analizar los patrones de interacción entre candidatos (como se menciona en la subsección 2.3.3.1), y posteriormente calcular aquel subconjunto de personas que forme un subgrafo con un patrón de interacción más uniforme, en espera de que la seguridad psicológica emerja en el equipo resultante.

El segundo punto es que existen muchas variantes del PFE tanto en el contexto de su aplicación, como en el contexto de su modelo de efectividad, como en el contexto de sus componentes y restricciones, como en el contexto del origen de sus variables de entrada. Por ejemplo, es muy difícil comparar si una matriz de incidencia candidato-posición produce mejores resultados empleando valores de números enteros o de números difusos si el modelo en el que se emplean y el contexto en el que se usan son distintos. Todas estas diferencias entorpecen el estudio del problema en sí. Particularmente, dificultan la identificación de elementos esenciales en los modelos, al mismo tiempo que obstaculizan la comparación entre ellos.

Ambos puntos dan la pauta para organizar los trabajos de los PFE en cuanto a algo más que sólo su contexto. Por ejemplo, una clasificación taxonómica en cuanto a la estructura de sus modelos. Esto invita a descomponer las formulaciones en sus elementos esenciales, para posteriormente analizarlos y contrastarlos.

Capítulo 3. Propuesta de taxonomía de los problemas de formación de equipo

En un sentido general, las formulaciones del Problema de Formación de Equipo (PFE) pueden clasificarse en dos grandes colecciones basadas en cómo se modela el problema; éstas son los modelos basados en la asignación y los modelos basados en la comunidad.

En este capítulo, se presenta una propuesta taxonómica de los PFE desde una perspectiva de la estructura de sus modelos. La jerarquía de la taxonomía propuesta se muestra en la Figura 3.

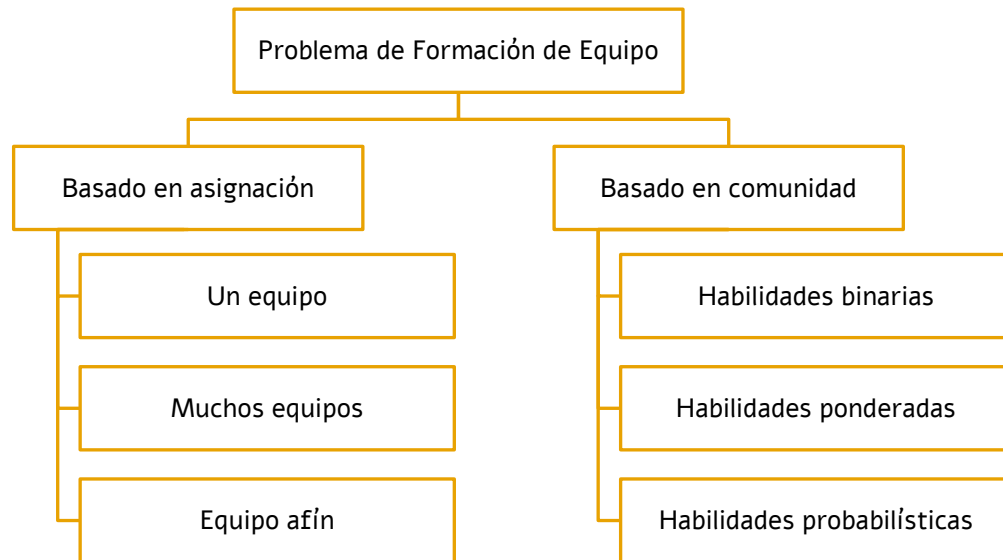


Figura 3. Taxonomía propuesta de los problemas de formación de equipo.

3.1. Modelos basados en asignación

Los modelos basados en asignación implican un emparejamiento bipartito entre un conjunto de candidatos y un conjunto de equipos o puestos de equipo (como se muestra en la Figura 4). Cada emparejamiento potencial tiene un peso—un puntaje—que representa la idoneidad de un candidato con respecto a un equipo o una posición de equipo. Entonces, el objetivo general de los PFEs basados en la asignación es encontrar un emparejamiento con la máxima puntuación de idoneidad. Estas formulaciones se encuentran habitualmente en el campo de la investigación de operaciones.

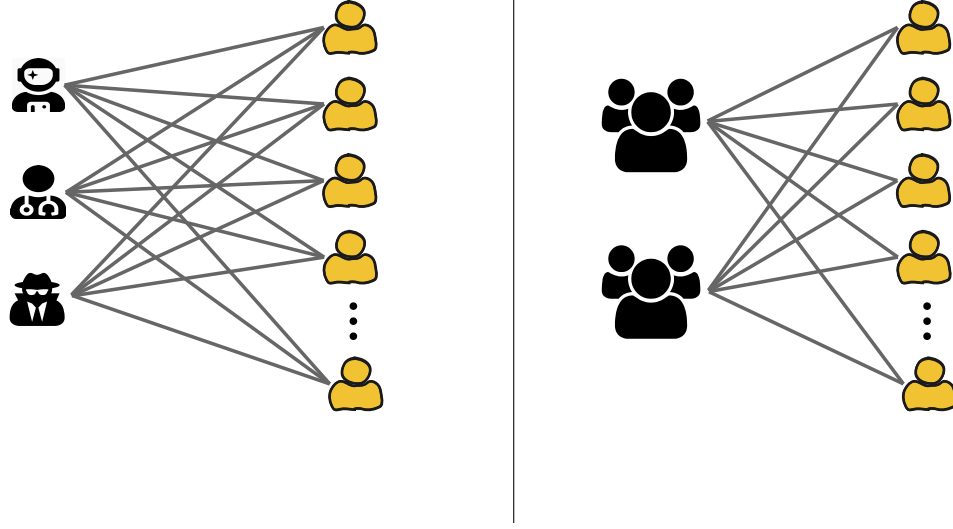


Figura 4. Ejemplo de modelos generales basados en asignación. Existe un peso de idoneidad asociado a cada conexión. Izquierda: un grafo bipartito que representa una asociación candidato a puesto de equipo. Derecha: un grafo bipartito que representa una asociación candidato a equipo (el puesto no es importante).

3.1.1. Formación de un equipo

Es el PFE más básico, requiere la formación de un solo equipo. Algunos consideran puestos de equipo mientras que otros no.

Con puestos de equipo. El caso del problema consiste en un conjunto de candidatos y un conjunto de puestos de equipo, donde cada candidato tiene una puntuación de idoneidad para cada puesto. El objetivo es designar a los candidatos a los puestos de los equipos de tal manera que la puntuación total de idoneidad del emparejamiento sea máxima. Sin embargo, deben cumplirse dos restricciones (i) cada puesto se debe asignar a al menos un candidato, y (ii) ningún candidato se debe asignar a más de un puesto. Ejemplos de ello pueden ser la búsqueda de la mejor plantilla o alineación posible en deportes, tripulaciones aéreas, unidades militares de combate o equipos quirúrgicos.

Dados n candidatos y ℓ puestos de equipo, la formulación matemática de esta variante de PFE basada en asignación viene dada por la ecuación (4) y las restricciones (5)–(7).

$$\text{maximizar } \sum_{i=1}^n \sum_{k=1}^{\ell} w_{ik} x_{ik} \quad (4)$$

sujeto a

$$\sum_{k=1}^{\ell} x_{ik} \leq 1 \quad i = 1, 2, \dots, n, \quad (5)$$

$$\sum_{i=1}^n x_{ik} = 1 \quad k = 1, 2, \dots, \ell, \quad (6)$$

$$x_{ik} \in \{0, 1\}, \quad i = 1, 2, \dots, n, \quad k = 1, 2, \dots, \ell, \quad n \geq \ell, \quad (7)$$

donde $x_{ik} = 1$ si el candidato i está asignado al puesto k , 0 en caso contrario; w_{ik} puntuación que representa cuán calificado está el candidato i para desempeñar la función k . Las restricciones (5) y (6) garantizan que ningún candidato se asigne a más de un puesto, y que cada puesto se asigne a un candidato, respectivamente. Obsérvese que este modelo de PFE es equivalente a la versión de maximización de utilidad del *problema de asignación clásico* (Kuhn, 1955; Pentico, 2007), que puede resolverse en tiempo polinomial (Papadimitriou y Steiglitz, 1998). La clave en esta equivalencia es la asignación uno a uno sobre una estructura bipartita; en este caso, candidatos y puestos de equipo.

Boon y Sierksma (2003) exploraron este problema buscando la alineación inicial óptima de un equipo de fútbol. Este problema requiere que se asignen once posiciones de campo a once jugadores de la plantilla del equipo. Cada jugador candidato tiene una puntuación de idoneidad según cada posición de campo.

Sin puestos de equipo. La formación de un solo equipo sin puestos de equipo es una formulación más abstracta que su contraparte. Un caso del problema consiste en un conjunto de candidatos, cada uno con una puntuación de idoneidad y un costo, y un presupuesto de equipo. El objetivo es seleccionar a los candidatos que se unirán al equipo de manera que la puntuación total de idoneidad sea máxima y el costo total sea menor o igual que el presupuesto del equipo. Esta formulación es adecuada para equipos cuya definición de los puestos de trabajo no es tan relevante, como en los equipos creativos. En cambio, los requisitos se consideran prototípicos para todos los miembros del equipo.

Dados n candidatos y un presupuesto B para formar el equipo, la formación de equipo único sin posiciones puede definirse como se expresa en la ecuación (8) y las restricciones (9) y (10).

$$\text{maximizar } \sum_{i=1}^n w_i x_i \quad (8)$$

sujeto a

$$\sum_{i=1}^n c_i x_i \leq B, \quad (9)$$

$$x_i \in \{0, 1\}, \quad i = 1, 2, \dots, n, \quad (10)$$

donde $x_i = 1$ si el candidato i se asigna al equipo, $x_i = 0$ en caso contrario, y w_i su puntuación de idoneidad para el equipo. La restricción (9) hace que la suma total de los costes c_i , de cada miembro del equipo i , no supere el presupuesto B . Obsérvese que esta formulación es equivalente al problema conocido como Knapsack Problem (o problema de la mochila), que es NP-difícil (Garey y Johnson, 1979). Para este problema se espera que una mochila de capacidad fija se llene con los artículos más valiosos de entre un conjunto de artículos de varios pesos. En este caso, el equipo representa la mochila, mientras que los candidatos representan los artículos. Así, los costes y las puntuaciones de idoneidad de los candidatos corresponden a los pesos y valores de los artículos, según corresponda. Sin embargo, para el caso especial de formación de equipo que no considera presupuesto sino tamaño de equipo—con B representando el tamaño del equipo y cada c_i siendo unitario en la restricción (9)—, el problema puede reducirse al problema de la mochila donde los artículos tienen el mismo peso, lo cual se considera manejable. Una variante importante de generalización considera que los candidatos tienen un coeficiente multidimensional w_i^k que modela la puntuación de idoneidad respecto a múltiples características, siendo k el índice de la característica. Entonces, en lugar de la función (8), la función objetivo correspondiente sería maximizar $\sum_{i=1}^n \sum_{k=1}^p w_i^k x_i$. Esta generalización sería equivalente al problema conocido como Multiobjective Knapsack Problem o problema de la mochila multiobjetivo.

Ahmed *et al.* (2013) propusieron un PFE en el contexto de los equipos de cricket. Un equipo de cricket tiene once jugadores, en el que sólo dos de ellos tienen un papel único: el capitán y el receptor (wicket-keeper en inglés). Para este problema, los candidatos reciben una puntuación según tres criterios: rendimiento de bateo, rendi-

miento de bolos y rendimiento de campo. Además, cada candidato incurre en un coste de contratación que puede ser diferente al de los demás. El objetivo es encontrar un subconjunto de candidatos que maximice la puntuación total de idoneidad de los tres criterios, siempre que los costes totales estén dentro del presupuesto del equipo. Además de la restricción presupuestaria, existen algunas restricciones particulares: (i) hay un límite de tamaño del equipo de once jugadores, (ii) el capitán y el receptor se deben elegir de sus respectivos grupos, (iii) hay un número máximo de jugadores extranjeros en el equipo.

3.1.2. Formación de muchos equipos

Esta clase es una generalización del problema de formación de un equipo. Particularmente, este conjunto de problemas tiene que lidiar con los detalles de la formación de varios equipos, por ejemplo, si un candidato se podría o no asignar a más de un equipo.

Con puestos de equipo. La formación de muchos equipos con puestos de equipo implica un conjunto de candidatos, cada uno con una capacidad de carga de trabajo, y un conjunto de equipos, cada uno de los cuales consta de un número de puestos de equipo. Al igual que otros PFE basados en asignación, los candidatos poseen una puntuación de idoneidad para cada puesto en el equipo. El objetivo es asignar cada candidato a cada puesto de forma que la puntuación total de idoneidad de todos los equipos sea máxima. Se permite que cualquier candidato ocupe cualquier número de puestos de equipo siempre que (i) ningún candidato ocupe varios puestos en un mismo equipo, (ii) ningún puesto se asigne a más de un candidato, y (iii) ningún candidato supere su capacidad de trabajo.

Dados n candidatos, m equipos, y ℓ puestos en cada equipo, una formulación matemática para la formación de muchos equipos con puestos de equipo puede verse como se expresa en la ecuación (11) y las restricciones (12)–(15).

$$\text{maximizar } \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^{\ell} w_{ijk} x_{ijk} \quad (11)$$

sujeto a

$$\sum_{j=1}^m \sum_{k=1}^{\ell} a_{ijk} x_{ijk} \leq b_i \quad i = 1, 2, \dots, n, \quad (12)$$

$$\sum_{k=1}^{\ell} x_{ijk} \leq 1 \quad i = 1, 2, \dots, n \quad j = 1, 2, \dots, m, \quad (13)$$

$$\sum_{i=1}^n x_{ijk} = 1 \quad j = 1, 2, \dots, m \quad k = 1, 2, \dots, \ell, \quad (14)$$

$$x_{ijk} \in \{0, 1\}, \quad i = 1, 2, \dots, n \quad j = 1, 2, \dots, m, \quad k = 1, 2, \dots, \ell, \quad (15)$$

donde $x_{ijk} = 1$ si el candidato i está asignado al equipo j en el puesto k , 0 en caso contrario, y w_{ijk} es la puntuación de idoneidad de dicha asignación; a_{ijk} es la cantidad de capacidad consumida por el candidato i para el puesto k en el equipo j , y b_i denota su capacidad máxima. La restricción (12) garantiza que el candidato i no supere su capacidad máxima de trabajo b_i . La restricción (13) garantiza que el candidato i no se asigne a más de un puesto en un mismo equipo j , mientras que la restricción (14) especifica que cada puesto de equipo debe estar ocupado. Por último, la restricción (15) exige que la solución tome variables binarias. La implementación de este modelo puede requerir una relajación de la restricción (14) ya que la restricción de capacidad (12) podría no permitir llenar todos los requisitos de los puestos en cada equipo, y el modelo sería inviable para algunas entradas. Una alternativa de relajación podría ser sustituir (14) por $g_{jk} = 1 - \sum_{i=1}^n x_{ijk}$ que actúa como contador de huecos, y luego añadir la expresión $-\sum_{j=1}^m \sum_{k=1}^{\ell} g_{jk} \cdot M$ al objetivo (11), la cual impone una penalización M por cada hueco candidato–puesto g_{jk} . Este modelo es similar al problema conocido como Multi-dimensional General Assignment Problem o problema de asignación general multidimensional, o puede verse también como una generalización del problema conocido como Multilevel Generalized Assignment Problem o problema de asignación general multinivel, ambos problemas están en la clase NP-difícil (Glover *et al.*, 1979; Laguna *et al.*, 1995; Pentico, 2007).

Una característica de esta clase es que, si se añade una restricción a los candidatos a formar parte de un solo equipo, entonces el *problema de formación de muchos equi-*

pos con puestos de equipo puede reinterpretarse como un *problema de formación de un equipo con puestos de equipo*. Esto se hace mediante la fusión de todos los equipos como un solo equipo considerando todos los puestos originales de cada equipo. Luego, se resuelve el problema como uno de formación de un equipo con puestos de equipo. Por último, manteniendo un registro de los puestos que pertenecen originalmente a cada equipo, se obtiene la solución al problema original.

Van de Water *et al.* (2007) examinaron la división de un grupo de estudiantes en varios equipos. Cada equipo se compone de nueve puestos de equipo, y cada estudiante tiene una puntuación de idoneidad para cada puesto. En este problema, los estudiantes no pueden formar parte de varios equipos (la restricción (12) no es necesaria). El objetivo es asignar a los estudiantes a los puestos de los equipos de manera que la puntuación total de idoneidad sea máxima. Este caso considera un conjunto distintivo de restricciones de “balanceo”, que por ejemplo requieren que siempre que dos estudiantes tengan una puntuación alta en el mismo rol, se evite que esos estudiantes estén en el mismo equipo.

Sin puestos de equipo. La formación de muchos equipos sin puestos de trabajo especifica un conjunto de candidatos, un conjunto de equipos y un presupuesto para cada equipo. Los candidatos tienen una puntuación de idoneidad asociada para cada equipo, un coste y una capacidad de carga de trabajo. El objetivo es encontrar un subconjunto de candidatos por cada equipo que maximice la puntuación total de idoneidad entre todos los equipos. Cualquier candidato puede formar parte de varios equipos siempre que no se supere su capacidad de carga de trabajo. Además, el coste total de los candidatos de un equipo no debe superar el presupuesto del equipo.

Dados n candidatos, m equipos, y un presupuesto B_j por cada equipo $j = 1, 2, \dots, m$, el modelo matemático propuesto para esta formulación de PFE basado en asignación está dada por la expresión de la ecuación (16) y las restricciones (17)–(19).

$$\text{maximizar } \sum_{i=1}^n \sum_{j=1}^m w_{ij} x_{ij}, \quad (16)$$

sujeto a

$$\sum_{j=1}^m a_{ij} x_{ij} \leq b_i \quad i = 1, 2, \dots, n, \quad (17)$$

$$\sum_{i=1}^n c_{ij} x_{ij} \leq B_j \quad j = 1, 2, \dots, m, \quad (18)$$

$$x_{ij} \in \{0, 1\}, \quad i = 1, 2, \dots, n, \quad j = 1, 2, \dots, m, \quad (19)$$

donde w_{ij} representa la puntuación de idoneidad del candidato i con respecto al equipo j . La variable de decisión $x_{ij} \in \{0, 1\}$ denota que el candidato i forma parte (o no) del equipo j . Las restricciones definidas por las ecuaciones (17) y (18) impiden que ningún candidato supere su capacidad máxima de trabajo b_i , y que ningún equipo supere su presupuesto B_j , correspondientemente. Este modelo puede interpretarse como una generalización del problema conocido como Multiple Knapsack Problem (MKP) o problema de múltiples mochilas en el que hay más de una mochila (Kellerer *et al.*, 2013). Dado que el MKP es NP-difícil (Kellerer *et al.*, 2013), este problema también es NP-difícil.

Zakarian y Kusiak (1999) estudiaron este problema en el contexto de los equipos de ingeniería para el diseño y desarrollo en la industria automotriz. En este trabajo, el diseño y desarrollo de nuevas características de ingeniería en un vehículo requiere integrar un equipo particular. Cada equipo se define en función del tipo de ingeniería que requiere la característica correspondiente. Cada candidato tiene una puntuación de idoneidad para cada equipo basada en la relevancia de sus especialidades de ingeniería para la tarea. El objetivo es asignar a los candidatos a los equipos de ingeniería para que la puntuación de idoneidad global sea máxima. Los candidatos pueden pertenecer a varios equipos, pero sólo a un número máximo de ellos. Este caso no considera una restricción de presupuesto o de tamaño de los equipos, pero sí incluye una restricción en el número de equipos de ingeniería que pueden formarse.

3.1.3. Formación de equipos afines

Mientras que las clasificaciones basadas en la asignación mencionadas anteriormente sólo consideran las habilidades técnicas como criterio para formar equipos, las presentes formulaciones reconocen el efecto social al considerar la conducta individual, el impulso o la preferencia de los individuos al trabajar juntos. Mucho de esto se modela a partir de las tipificaciones de rasgos personales (que se discuten en la subsección 2.3.2). Así, además de las puntuaciones de idoneidad de los candidatos, también se considera un puntaje de afinidad entre candidatos.

Con puestos de equipo. Un caso del problema de formación de equipos afines con puestos de equipo consiste en un conjunto de candidatos y un conjunto de puestos de equipo. Cada candidato tiene una puntuación de idoneidad para cada puesto de equipo, y cada par de candidatos comparte una puntuación de interacción. El objetivo es encontrar un emparejamiento candidato-puesto de equipo tal que la puntuación total de idoneidad y la puntuación total de afinidad entre pares de miembros del equipo sean máximas.

Dados n candidatos y ℓ puestos de equipo, el modelo matemático está dado por la ecuación (20) y las restricciones (21)–(23).

$$\text{maximizar} \left(\sum_{i=1}^n \sum_{k=1}^{\ell} w_{ik} x_{ik} + \sum_{i=1}^{n-1} \sum_{i'=i+1}^n \sum_{k=1}^{\ell-1} \sum_{k'=k+1}^{\ell} r_{ii'} x_{ik} x_{i'k'} \right), \quad (20)$$

sujeto a

$$\sum_{k=1}^{\ell} x_{ik} \leq 1 \quad i = 1, 2, \dots, n, \quad (21)$$

$$\sum_{i=1}^n x_{ik} = 1 \quad k = 1, 2, \dots, \ell, \quad (22)$$

$$x_{ik} \in \{0, 1\}, \quad i = 1, 2, \dots, n, \quad k = 1, 2, \dots, \ell \quad (23)$$

donde $x_{ik} = 1$ si el candidato i se asigna a la posición k del equipo, 0 en caso contrario; w_{ik} representa la puntuación de idoneidad del candidato–posición, y $r_{ii'}$ es la puntuación de compatibilidad entre los candidatos i e i' (como se describe en la subsección 2.3.2). La restricción (21) garantiza que ningún candidato se asigne a más de un pues-

to en el equipo, mientras que la restricción (22) especifica que cada puesto del equipo se debe asignar a un solo candidato. La restricción (22) podría requerir una relajación para garantizar que el modelo no sea inviable para algunas entradas, de forma similar a lo que se describe para el modelo (11)–(15). Estos modelos comparten un gran parecido con los de otro problema de asignación conocido como Quadratic Assignment Problem (QAP) o problema de asignación cuadrática (Loiola *et al.*, 2007). El QAP busca determinar el empatamiento de un conjunto de instalaciones entre un conjunto de localidades que minimice el costo total, donde existe un coste asociado a la distancia entre cada localidad y el flujo entre las instalaciones. Así, el modelo del QAP se caracteriza por tener un producto de dos variables de decisión y un costo asociado al valor de estas dos variables. La segunda componente de la ecuación (20) es un ejemplo de este parecido. La equivalencia entre el QAP y el *problema de formación de equipos afines con puestos de equipo* ocurre cuando el QAP considera un flujo unitario entre instalaciones.

Zhang y Zhang (2013) consideraron una formación de equipo para el desarrollo de nuevos productos que implica a miembros de diferentes departamentos con buenas relaciones interpersonales. El equipo tiene un número determinado de puestos para miembros de cada departamento. Los candidatos tienen una puntuación de idoneidad sólo en relación con los puestos disponibles para su departamento. Además, cada par de candidatos tiene una puntuación de afinidad que representa la inclinación mutua a cooperar. El objetivo es asignar a los candidatos de los departamentos a sus correspondientes puestos de equipo departamentales de forma que la puntuación total de idoneidad y la puntuación total de afinidad sean máximas.

Sin puestos de equipo. La formación de equipos afines sin puestos de equipo conlleva un conjunto de candidatos y un presupuesto de equipo. Los candidatos tienen un coste, una puntuación de idoneidad para cada equipo y una puntuación de afinidad con cada uno de los otros candidatos. El objetivo es encontrar un subconjunto de candidatos que maximice tanto la puntuación total de idoneidad como la puntuación total de afinidad.

Dados n candidatos y un presupuesto B para formar el equipo, la formulación matemática de este problema viene dada por la expresión de la ecuación (24) y las res-

tricciones (25) y (26).

$$\text{maximizar } \left(\sum_{i=1}^n w_i x_i + \sum_{i=1}^{n-1} \sum_{i'=i+1}^n r_{ii'} x_i x_{i'} \right), \quad (24)$$

sujeto a

$$\sum_{i=1}^n c_i x_i \leq B, \quad (25)$$

$$x_{ij} \in \{0, 1\}, \quad i = 1, 2, \dots, n \quad (26)$$

donde $x_{ij} = 1$ si el candidato i se asigna al equipo j , 0 en caso contrario, y w_{ij} es la puntuación de idoneidad de dicha asignación, mientras que $r_{ii'}$ es la puntuación de afinidad entre los candidatos i e i' . Aunque esta formulación no considera los puestos del equipo, la restricción (25) especifica un presupuesto que limita el número de miembros del equipo. Estos modelos tienen un gran parecido con el problema conocido como Quadratic Knapsack Problem o problema de la mochila cuadrática (Kellerer *et al.*, 2013).

Yaakob y Kawata (Yaakob y Kawata, 1999) propusieron un problema de formación de equipos para redistribuir un conjunto de trabajadores de un entorno industrial en equipos con una cierta afinidad (este problema es tanto de formación de equipos afines como de formación de muchos equipos). La redistribución debe tener en cuenta las habilidades técnicas y las relaciones de afinidad de los trabajadores. Además, los equipos no buscan cubrir puestos, sino encontrar miembros calificados para el trabajo en su conjunto. Cada trabajador candidato tiene una puntuación de idoneidad para cada equipo, y una puntuación de afinidad con cada uno de los otros candidatos. Los trabajadores no pueden formar parte de varios equipos. El objetivo es encontrar una redistribución de los candidatos a los equipos de forma que tanto la puntuación total de idoneidad como la puntuación de afinidad con el equipo sean máximas.

3.2. Modelos basados en comunidad

Los modelos basados en comunidad se encuentran comunmente en el contexto de las redes sociales profesionales. En este enfoque, se captura el conjunto de candidatos expertos y las relaciones diádicas entre ellos (por ejemplo, amistad, colaboración,

dominio académico o profesional). Cada candidato además, está relacionado con un conjunto específico de habilidades. Así, los modelos basados en comunidad consisten en encontrar un subconjunto de candidatos con interacciones estrechas entre ellos y cuyo conjunto combinado de habilidades cumpla con todos los requisitos de una tarea en particular (como se muestra en la Figura 5). Gran parte de los trabajos más recientes sobre la formación de equipos se inclinan por este enfoque, encabezado por comunidades de investigación como la minería de datos y la ciencia de datos.

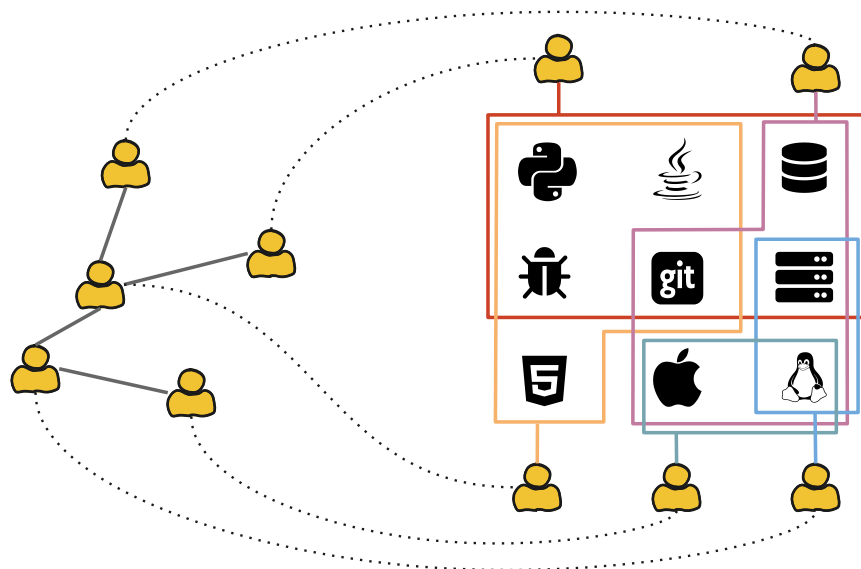


Figura 5. Ejemplo de un modelo basado en comunidades. Los candidatos están relacionados entre sí mediante una red social, y cada candidato está asociado a un conjunto de habilidades.

3.2.1. Habilidades binarias

Las formulaciones basadas en la comunidad con habilidades binarias consisten en aquellos modelos en los que el candidato simplemente posee alguna habilidad o no.

Un caso del problema consiste en un grafo no dirigido ponderado y una tarea definida como un conjunto de habilidades. En el grafo existe un nodo por cada candidato, y una arista por cada par de candidatos que tienen comunicación (o colaboración). Adicionalmente, el peso de una arista simboliza el coste de comunicación entre ambos candidatos, y cada candidato posee un conjunto de habilidades propias. El objetivo es encontrar un subconjunto de candidatos que minimice una función de *costo de comunicación* (o que maximice una función de *fuerza de colaboración*), con la condición

de que el conjunto global de habilidades de los candidatos cubra las habilidades requeridas por la tarea (como se muestra en la Figura 5). En esencia, para cubrir una habilidad de una tarea, basta con que un miembro del equipo posea esa habilidad.

Esta formulación se define por medio de la ecuación (27): Dado un grafo social $G = (\mathcal{X}, E)$ y una tarea $T \subseteq \mathcal{A}$,

$$\underset{\mathcal{X}' \subseteq \mathcal{X}}{\text{minimizar}} f(\mathcal{X}'), \quad (27)$$

sujeto a la restricción (28)

$$T \cap \left(\bigcup_{i \in \mathcal{X}'} X_i \right) = T, \quad (28)$$

donde $f(\mathcal{X}')$ es una función de costo de comunicación, $X_i \subseteq \mathcal{A}$ es el conjunto de habilidades del candidato i , y $\mathcal{A}$ es el universo de habilidades. La función de costo de comunicación puede ser el costo del árbol de expansión mínimo o el diámetro del subgrafo inducido $G[\mathcal{X}']$, de un equipo $\mathcal{X}'$, sobre el grafo de entrada original G . Este problema es NP-difícil con cualquiera de estas dos funciones de costo de comunicación (Lappas et al., 2009). La reducción viene del problema de cobertura de elección múltiple (Arkin y Hassin, 2000) y del problema del árbol steiner de grupo (Reich y Widmayer, 1989), en relación con el diámetro y las funciones de costo MST, respectivamente (Lappas et al., 2009). Nótese que incluso la minimización del tamaño del equipo ($|\mathcal{X}'|$) es un problema NP-difícil ya que la restricción (28) implica resolver un problema conocido como Set Cover o cubrimiento de conjunto, que pertenece a la clase de problemas NP-completa (Garey y Johnson, 1990).

Este problema tiene muchas variantes. Algunos trabajos consideran otras funciones de costo de comunicación del equipo como la suma total de las distancias de los candidatos por pares. Mientras que otros trabajos han propuesto versiones de maximización de este problema, optimizando para funciones de *fuerza de colaboración* del equipo representadas por la función de densidad del (sub)grafo. También existe una generalización de la restricción (28) en (Gajewar y Das Sarma, 2012), que exige una cantidad mínima de miembros para cubrir una habilidad. Esto requiere una tarea generalizada $T \subseteq \mathcal{A} \times \mathbb{N}$ como entrada, mientras que la restricción generalizada

es la siguiente $|\mathcal{X}' \cap S(a)| \geq k_a, \forall (a, k_a) \in T$, donde $S(a) = \{i \in \mathcal{X} | a \in X_i\}$ es el conjunto de candidatos que poseen la habilidad a , $X_i \subseteq \mathcal{A}$ es el conjunto de habilidades del candidato i , y $\mathcal{A}$ es el universo de habilidades. Por ejemplo, una tarea generalizada $T = \{(a_1, k_1), (a_2, k_2), (a_3, k_3)\}$ dice que el equipo a formar debe tener al menos k_1 miembros que posean la habilidad a_1 , k_2 miembros que posean la habilidad a_2 , y k_3 miembros que posean la habilidad a_3 . Note que un miembro puede poseer más de una habilidad requerida, y contaría por cada una de éstas.

3.2.2. Habilidades ponderadas

Esta clasificación toma en cuenta el nivel de experiencia de cada candidato representándola con número entero o real, a diferencia de representarla simplemente de manera binaria.

Un caso del problema consiste en un grafo ponderado no dirigido, con nodos como candidatos y aristas como sus vínculos, y una tarea definida como un conjunto de habilidades. Cada arista tiene un peso que se corresponde con la fuerza de colaboración de sus candidatos asociados¹. Adicionalmente, cada candidato posee un conjunto de habilidades, cada una con un *valor de nivel de habilidad* asociado. El objetivo es seleccionar un subconjunto de candidatos que maximice tanto la función *fuerza de colaboración* como la función *nivel de habilidades*. Esta vez, las habilidades de la tarea pueden o no estar cubiertas por el conjunto de habilidades de los miembros del equipo.

En la literatura se suelen encontrar dos modelos prototípicos para esta clase. En primer lugar, se presenta un modelo de suma lineal representado por la expresión (29): Dado un grafo social $G = (\mathcal{X}, E)$ y una tarea $T \subseteq \mathcal{A}$,

$$\underset{\mathcal{X}' \subseteq \mathcal{X}}{\text{maximizar}} (\alpha f(\mathcal{X}') + (1 - \alpha)g(\mathcal{X}')), \quad (29)$$

con $\alpha \in [0, 1]$ un parámetro de compensación definido por el usuario, $f(\mathcal{X}')$ una función *fuerza de colaboración* del equipo, $g(\mathcal{X}')$ una función *nivel de habilidad* (más adelante se discuten estas funciones). Tenga en cuenta que la restricción de cobertura

¹Recordar que el peso de una arista también puede representar costo de comunicación, pero por simplicidad se decide representar la versión de maximización del problema.

de habilidades—como la restricción (28)—puede omitirse ya que una función de nivel de habilidad es parte de la función objetivo.

En segundo lugar, una versión multiobjetivo del problema está dada por la expresión (30): Dado un grafo social $G = (\mathcal{X}, E)$ y una tarea generalizada $T \subseteq \mathcal{A} \times \mathbb{N}$,

$$\underset{\mathcal{X}' \subseteq \mathcal{X}}{\text{maximizar}} (f(\mathcal{X}'), g(\mathcal{X}')), \quad (30)$$

sujeto a la restricción (31)

$$|\mathcal{X}' \cap S(a)| \geq k_a \quad \forall (a, k_a) \in T, \quad (31)$$

con $S(a) = \{i \in \mathcal{X} | a \in X_i\}$ el conjunto soporte (o *support set*) de los candidatos que tienen la habilidad a (sin importar su nivel de experiencia), y k_a la cantidad mínima requerida de candidatos con la habilidad a . En este modelo, $f(\mathcal{X}')$ y $g(\mathcal{X}')$ también son una función *fuerza de colaboración* y una función *nivel de habilidad*, respectivamente.

Como se ha mencionado anteriormente, la función de fuerza de colaboración juega el papel de una contraparte de maximización de las funciones de distancia de comunicación, y se define comúnmente como la densidad del subgrafo inducido $G[\mathcal{X}']$. Es decir, $f(\mathcal{X}') = \sum_{e \in E_{G[\mathcal{X}]}} \frac{w(e)}{|\mathcal{X}'|}$, donde $w(e)$ es la fuerza de colaboración entre los candidatos de los extremos de la arista e . Por otra parte, la función de nivel de habilidad $g(\mathcal{X}')$ suele definirse para representar cuánta experiencia hay en $\mathcal{X}'$ respecto a las habilidades de la tarea en T . Por ejemplo, Dorn y Dustdar (2010) definen una función de nivel de habilidad como $g(\mathcal{X}') = \sum_{a \in T} (\max_{i \in \mathcal{X}'} (w_{ia}) \cdot w_a)$, donde w_{ia} es el nivel de experiencia del candidato i sobre la habilidad a , y w_a es la importancia de la experiencia a . Juárez y Brizuela (2018), por su parte, proponen $g(\mathcal{X}') = \sum_{i \in \mathcal{X}'} \sum_{a \in T} \frac{w_{ia}}{|\mathcal{X}'|}$ que es la relación entre el total de la experiencia presente y el tamaño del equipo.

3.2.3. Habilidades probabilísticas

Estas formulaciones tratan el problema de la formación de equipos cuando hay incertidumbre en los niveles de habilidad de los candidatos, considerándolos como probabilísticos.

Un caso del problema requiere un grafo no dirigido ponderado que consiste en can-

didatos como nodos y aristas como sus relaciones, y una tarea definida por un conjunto de habilidades. Cada candidato posee un conjunto de habilidades, en el que cada nivel de habilidad se representa con una variable probabilística y una distribución. El objetivo es identificar un subconjunto de candidatos que maximice una función probabilística de nivel de habilidad. Nuevamente, las habilidades de la tarea pueden o no estar cubiertas por el conjunto de habilidades de los miembros del equipo.

La formulación matemática (adaptada de (Liemhetcharat y Veloso, 2012, 2014)) está dada por la expresión (32). Dado un grafo social $G = (\mathcal{X}, E)$ y una tarea $T \subseteq \mathcal{A}$,

$$\underset{\mathcal{X}' \subseteq \mathcal{X}}{\text{maximizar}} \sum_{a \in T} (\mu_{\mathcal{X}'a} + \sigma_{\mathcal{X}'a} \cdot \Phi^{-1}(\alpha)), \quad (32)$$

y las ecuaciones (33) y (34)

$$\mu_{\mathcal{X}'a} = \frac{1}{\binom{|\mathcal{X}'|}{2}} \sum_{i, i' \in \mathcal{X}'} \phi(d(i, i')) \cdot (\mu_{ia} + \mu_{i'a}), \quad (33)$$

$$\sigma_{\mathcal{X}'a}^2 = \frac{1}{\binom{|\mathcal{X}'|}{2}^2} \sum_{i, i' \in \mathcal{X}'} \phi(d(i, i'))^2 \cdot (\sigma_{ia}^2 + \sigma_{i'a}^2), \quad (34)$$

donde el nivel de habilidad $w_{ia} \sim \mathcal{N}(\mu_{ia}, \sigma_{ia}^2)$ es una variable de distribución normal del candidato i con respecto a la habilidad a , con μ_{ia} el rendimiento medio de su habilidad y σ_{ia}^2 su varianza. Adicionalmente, $\Phi^{-1}(\alpha)$ es la función *probit* (o la inversa de la función de distribución acumulativa de la distribución Normal estándar) y $\alpha \in [0, 1]$ un factor de riesgo. Si el riesgo se establece alto $\alpha > 0,5$ (bajo $\alpha < 0,5$), entonces los niveles de habilidad con una varianza más alta obtendrán puntuaciones más altas (más bajas). Finalmente, $\phi(d(i, i'))$ es una función monotónicamente decreciente, como $\frac{1}{d(i, i')}$ o $\exp\left(-\frac{d(i, i') \ln 2}{h}\right)$, de la distancia (más corta) $d(i, i')$ entre los agentes i e i' sobre el grafo G . Obsérvese que esta formulación maneja conjuntamente la función de colaboración ($f(\mathcal{X}')$) y la función de nivel de habilidad ($g(\mathcal{X}')$), como muestran las ecuaciones (33) y (34).

3.3. Discusión

La presente taxonomía ayuda a clasificar y estudiar los PFE mediante su estructura subyacente, fuera de las particularidades del contexto que regularmente los distingue.

Con esto se pueden identificar más cláramente a qué tipo de problema combinatorio se asemeja. Esta pequeña contribución además permite trasladar el conocimiento de otros problemas combinatorios al entorno del PFE.

Cabe aclarar que esta taxonomía no pretende ser completa. Constantemente se proponen más y más problemas combinatorios que involucran la formación de equipos, operando en condiciones o ambientes laborales y sociales particulares. Esto eventualmente puede conllevar la formulación de un modelo que no encaje en la taxonomía propuesta. También es importante aclarar que las divisiones de esta clasificación no son absolutas, algunas formulaciones comparten elementos de distintas categorías, como puede ser un PFE basado en asignación de un equipo y equipo afín.

En busca de un modelo más general, la taxonomía propuesta ayuda a identificar aquellos modelos que pudieran incorporar elementos que destaquen la interacción de sus integrantes de manera más nativa, como lo son los PFE basados en comunidad en general o los PFE basados en asignación con equipo afín.

Capítulo 4. Propuesta de una medida de interacción no simétrica

La mayoría de los PFEs están formulados alrededor del concepto de efectividad de equipo, pero la comprensión sobre la misma aún es limitada (Juárez *et al.*, 2021). En el Capítulo 2 se señala a la seguridad psicológica como predictor de efectividad de equipo. Sin embargo, la seguridad psicológica ha sido poco o nada explorada en el contexto de los PFEs. Esto se debe posiblemente a que esta información es relativamente reciente y ajena a las áreas donde se estudia el PFE. Por otro lado, la seguridad psicológica se considera una característica en el equipo, y no es algo que se pueda medir *a priori* ni de forma individual. Por lo que se necesita alguna forma de inferirla indirectamente, y que al mismo tiempo se pueda calcular con la información que se suele considerar como entrada de los modelos.

Recuerde que la seguridad psicológica en un equipo se define como la creencia compartida por los miembros del equipo de que el equipo representa un espacio seguro para expresarse libremente, sin miedo a represalias o consecuencias negativas hacia su imagen o su carrera (Edmondson, 1999). Esta característica puede manifestarse mediante ciertos patrones de interacción entre los miembros del equipo. Por ejemplo, aquellos equipos que presentan un patrón más uniforme en cuanto al tiempo y el turno para hablar de cada miembro, tienden a ser aquellos con mayor seguridad psicológica (Rozovsky, 2015).

El patrón de interacción, en cambio, es algo que se puede inferir con las herramientas y los enfoques existentes. Por ejemplo, el PFE se trasladó al contexto de las redes sociales, donde tanto los candidatos como sus interacciones están representados de forma nativa. Esta tendencia ofrece un marco adecuado para el análisis de posibles patrones de interacción debido a su estructura social inherente, y es también uno de los principales focos de atención en la literatura (Juárez y Brizuela, 2018; Selvarajah *et al.*, 2019; Chen *et al.*, 2020; Kou *et al.*, 2020). De acuerdo con la taxonomía propuesta en el Capítulo 3, esta variante le corresponde a los modelos de PFE basados en comunidad.

No obstante que los PFE basados en comunidad son apropiados para estudiar los patrones de interacción, los enfoques actuales tienen una limitación en común: las in-

teracciones simétricas. Esto implica que el peso de interacción del par de candidatos $A-B$ sería el mismo si va de A hacia B que si va de B hacia A . Sin embargo, las interacciones sociales no son necesariamente simétricas (ni transitivas) (Solanas *et al.*, 2006). Los ejemplos cotidianos de interacciones no simétricas van desde el supervisor-subordinado en los laboratorios de I+D (Allen *et al.*, 1984), hasta el seguidor-seguido en Twitter (Rishika y Ramaprasad, 2019). Reconocer la no simetría de las interacciones humanas es crucial para revelar patrones más complejos.

En este capítulo se propone una medida no simétrica para ponderar la interacción entre candidatos. También se amplía este concepto clasificando cada vínculo como fuerte o débil. Se espera que este método¹ pueda revelar patrones de interacción que definan estructuras de vínculos más complejos.

4.1. Preliminares

En esta sección se presenta literatura relevante adicional para contexto de la importancia de la propuesta.

4.1.1. Modelos de interacción en la literatura de formación de equipos

En las ciencias de la computación, las formulaciones actuales de los PFE basados en comunidad² definen la efectividad del equipo (o parte de ésta) como una función de las interacciones de los individuos entre sí. Estas interacciones corresponden a ocurrencias de comunicación o colaboración, por ejemplo, intercambios de correo electrónico (Van Staden y Van der Poel, 2016), o participación en proyectos (Lappas *et al.*, 2009; Kargar y An, 2011). En este sentido, esta familia de PFE ofrece un enfoque adecuado para emplear herramientas sociométricas y revelar posibles patrones de interacción.

Lappas *et al.* (2009) fueron los primeros en introducir el PFE en el contexto de las redes sociales. En su trabajo proponen usar la distancia de Jaccard—que cuantifica la disimilitud de dos conjuntos—para definir el coste de comunicación de dos individuos cualesquiera en relación con sus proyectos de equipo anteriores. Sea P_X el conjunto de proyectos en los que participa un individuo X . Entonces, la distancia de Jaccard de los

¹Se emplea la palabra método para referir a la medida no simétrica y a la clasificación en conjunto.

²A partir de ahora, se usa simplemente el término PFE para referir a esta variante.

individuos A y B se define como $1 - \frac{|P_A \cap P_B|}{|P_A \cup P_B|}$. Esta métrica se convirtió en una de las más populares para modelar el coste de la comunicación por parejas en la literatura sobre la formación de equipos (Chen *et al.*, 2020; Majumder *et al.*, 2012; Kargar *et al.*, 2012, 2013). Li y Shan (2010), por otro lado, modelaron el coste de comunicación de dos personas A y B simplemente como $\frac{1}{|P_A \cap P_B|}$. Sin embargo, esta medida de ponderación presenta algunas desventajas en comparación con la distancia de Jaccard. En primer lugar, contiene menos información³; en segundo lugar, no cumple con la desigualdad del triángulo. No obstante, está presente en otros trabajos como en (Jelić y Ševerdija, 2018).

Otros modelos ponderan las interacciones como una fuerza en lugar de un costo. Dorn y Dustdar (2010) modelaron la fuerza de la colaboración simplemente contando las interacciones entre dos individuos. Donde cualquier interacción directa entre A y B aumenta su peso de colaboración en uno, mientras que las interacciones indirectas contribuyen con un valor menor. En lugar de medir las interacciones granulares, Gajewar y Das Sarma (2012) se centraron en las colaboraciones de proyectos en su conjunto para la medición del peso. En particular, propusieron utilizar simplemente $|P_A \cap P_B|$ como fuerza de colaboración—que es intuitivamente mejor para representar colaboración que costo o distancia. Esta forma de medir la interacción se utiliza ampliamente en los PFE cuya función objetivo es maximizar la densidad de colaboración como en (Gajewar y Das Sarma, 2012; Rangapuram *et al.*, 2013; Juárez y Brizuela, 2018).

Como se revisa en esta subsección, independientemente de si la medida de interacción está modelando la distancia de comunicación o la fuerza de colaboración, los dos principales modelos de interacción son el simple y el de Jaccard. La Tabla 4 muestra una comparativa de los trabajos de investigación según sus modelos de interacción. Recuerde que los modelos de interacción son también los componentes básicos de las funciones objetivo de los modelos de optimización de los PFE basados en comunidad. En consecuencia, las funciones objetivo basadas en modelos de interacción simétricos tendrán un impacto en la composición de los equipos resultantes. Estudios como (Gutiérrez *et al.*, 2016) y (Farasat y Nikolaev, 2016) ya han introducido modelos de optimización para la formación de equipos que consideran interacciones no simétricas.

³No considera el número combinado de proyectos desempeñados $|P_A \cup P_B|$.

Sin embargo, ninguno de los dos trabajos está dedicado particularmente al cálculo de las interacciones no simétricas, éstas sólo se dieron como entrada para sus modelos.

Tabla 4. Comparativa de los diferentes trabajos de PFEs clasificados por medida de interacción.

Referencia	Medida de interacción	
	Simple	Jaccard
Li and Shan (Li y Shan, 2010)	✓	
Gajewar and Das Sarma (Gajewar y Das Sarma, 2012)	✓	
Rangapuram et al. (Rangapuram et al., 2013)	✓	
Juárez and Brizuela (Juárez y Brizuela, 2018)	✓	
Lappas et al. (Lappas et al., 2009)		✓
Majumder et al. (Majumder et al., 2012)		✓
Chen et al. (Chen et al., 2020)		✓
Kargar et al. (Kargar et al., 2012)		✓
Kargar et al. (Kargar et al., 2013)		✓

4.1.2. El flujo de información en las redes sociales

Muchas organizaciones orientadas al trabajo en equipo se enfrentan a la necesidad de adquirir y difundir información para prevalecer, especialmente en STEM⁴. La comunicación eficaz en toda la organización es crucial para el rendimiento. Es poco probable que un equipo—mucho menos un solo individuo—posea todos los conocimientos y habilidades necesarios para una tarea compleja. Por ello, los equipos y las organizaciones deben comprender de dónde procede la información fiable y cómo difundirla eficazmente.

4.1.2.1. La fuerza de los vínculos débiles

En el trabajo seminal de Granovetter (1973) se planteó la hipótesis de que cuanto más fuerte es la relación entre dos personas, mayor es la coincidencia entre sus círculos sociales. En otras palabras, si *A* y *B* son muy amigos, lo más probable es que ambos tengan amigos en común. Granovetter, también propuso dividir las relaciones interpersonales en dos grupos: amigos íntimos, representados por *vínculos fuertes*, y conocidos, representados por *vínculos débiles*. Así, si *A* y *B* tienen un vínculo fuerte mientras *B* y *C* también tienen un vínculo fuerte, la hipótesis de Granovetter predice que también existirá un vínculo entre *A* y *C*, ya sea fuerte o débil. Esto se conoce comúnmente como principio de *cierre triádico* (Easley y Kleinberg, 2010).

⁴De las siglas en inglés para ciencia, tecnología, ingeniería y matemáticas.

Además de un efecto local, el principio de *cierre triádico* exhibe el papel de los vínculos débiles en la mejora del flujo de información en una red (Friedkin, 1982). Si $A-B$ es un puente—un único segmento de red que conecta dos componentes—entre dos comunidades sociales, entonces, por el *cierre triádico*, $A-B$ debe ser un vínculo débil. Esto se deriva de que las comunidades sociales son característicamente densas en vínculos fuertes, por lo tanto si A forma parte de una comunidad y B forma parte de otra comunidad, entonces $A-B$ no puede ser un puente y un vínculo fuerte simultáneamente. Consecuentemente, se dice que los vínculos débiles son promotores importantes de la amplia difusión de la información en una red social, mientras que los vínculos fuertes sirven un propósito de difusión a nivel más local.

Intuitivamente, para promover mejores patrones de comunicación y flujo de información en un equipo, sería deseable que los miembros del equipo tuvieran vínculos fuertes hacia el interior del equipo, y vínculos débiles con otros colegas de su red profesional en el exterior del equipo. Esto último tiene un peso adicional por lo que se describe en la siguiente subsección.

4.1.2.2. El portero tecnológico

Allen *et al.* (1984) acuñó el término *technological gatekeeper* o portero tecnológico para referirse a una persona de una organización a la que sus compañeros suelen consultar para pedirle consejo técnico. Parte de su estudio consistió en observar los patrones de comunicación de los trabajadores en varios laboratorios de I+D⁵. El análisis de los datos arrojó que algunos individuos se comunicaban con casi todos los demás dentro de la organización, mientras que otros no lo hacían. Adicionalmente, estas “estrellas” sociométricas mantenían un contacto con múltiples fuentes de información fuera de la organización. Básicamente, su rol no escrito estaba en absorber la información de fuera y ponerla en el contexto de las necesidades de la organización. A su vez, Burt (2004) afirma que los individuos situados en el límite de distintos grupos pueden actuar como intermediarios de buenas ideas, ya que tienen la ventaja competitiva de un acceso temprano a la información y una perspectiva más amplia. Así, en lugar de que la información relevante fluya casualmente hacia la organización, es más

⁵Abreviación común para referir a departamentos o proyectos dedicados a la investigación y el desarrollo.

probable que ésta llegue y se difunda a través de sus porteros.

Pero la importancia de los porteros tecnológicos va más allá de su función en las organizaciones, ya que probablemente también desempeñan un papel en la estructura de las redes. Seguramente estos individuos son también responsables de acortar las distancias en una red, presumiblemente a través de vínculos débiles (según la hipótesis de Granovetter). Por lo tanto, al eliminarlos de la red, sus comunidades incidentes inevitablemente se distanciarían unas de otras.

4.2. Propuesta de un modelo de interacción no simétrico

En este capítulo, proponemos un nuevo modelo de interacción que reconoce las relaciones no simétricas. Particularmente, se define una función de interacción de fuerza de colaboración de un candidato A respecto a un candidato B como el cociente entre el número de proyectos realizados conjuntamente por A y B , y el número total de proyectos realizados por B . Este modelo se define formalmente por medio de la ecuación (35).

$$w_{\rightarrow}(A, B) = \frac{|P_A \cap P_B|}{|P_B|} \quad (35)$$

con P_A (P_B) el conjunto de proyectos en los que participa A (B). Esta ecuación está definida como una fuerza de colaboración, pero se puede transformar a una distancia de comunicación como $1 - w_{\rightarrow}(A, B)$. Se dice que este modelo es no simétrico porque $w_{\rightarrow}(A, B) \neq w_{\rightarrow}(B, A)$, a menos que $|P_A| = |P_B|$ o $|P_A \cap P_B| = 0$. Como comparativa, se definen los otros modelos de interacción más comunes. La interacción *simple* se modela de la siguiente manera por medio de la ecuación (36)

$$w_s(A, B) = |P_A \cap P_B|. \quad (36)$$

Y la interacción de *Jaccard* se modela por medio de la ecuación (37)

$$w_J(A, B) = \frac{|P_A \cap P_B|}{|P_A \cup P_B|}. \quad (37)$$

La Figura 6 muestra un ejemplo de las diferencias entre la medida propuesta y las otras medidas de interacción. Note que la propuesta en la ecuación (35) no requiere información adicional a los otros dos modelos de interacción de la ecuaciones (36) y

(37). Note además que la principal distinción entre la medida propuesta y las otras medidas recae justamente en la no simetría.

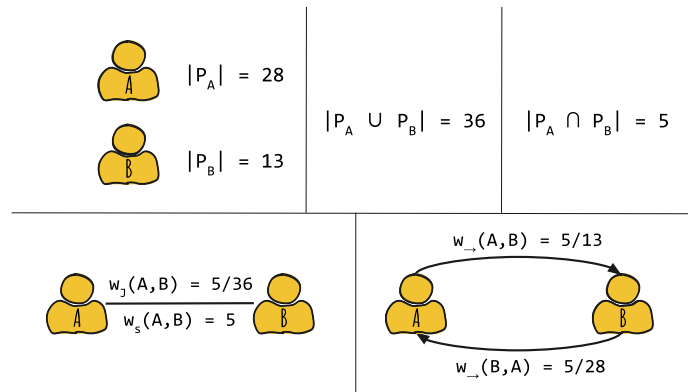


Figura 6. Un ejemplo de los modelos de interacción. Arriba izquierda: número individual de proyectos desempeñados por los individuos A y B. Arriba centro: número total de proyectos distintos desempeñados por A y B. Arriba derecha: número de proyectos donde A y B coinciden. Abajo: izquierda: $w_S(A, B)$ = conteo de colaboraciones de A y B, y $w_J(A, B)$ = índice de Jaccard entre A y B. Abajo derecha: $w_{\rightarrow}(A, B)$ = fuerza de vínculo de A sobre B, y $w_{\rightarrow}(B, A)$ = fuerza de vínculo de B sobre A.

También se propone discernir entre vínculos fuertes y débiles. Dado un umbral $\lambda \in (0, 1)$, si $w_{\rightarrow}(A, B) > \lambda$, se considera que el individuo A ejerce una influencia “fuerte” sobre el individuo B, o que A tiene un vínculo fuerte hacia B; en caso contrario, si $w_{\rightarrow}(A, B) \leq \lambda$ la influencia de A sobre B es “débil”, o A tiene un vínculo débil hacia B (más adelante se describe cómo estimar el valor de lambda). Esta distinción revela una nueva capa de información para el análisis de patrones más complejos por pares de candidatos. En particular, se pueden observar relaciones *totalmente fuertes* (los colaboradores cercanos), las relaciones *parcialmente fuertes* (el maestro–discípulo), y las relaciones *totalmente débiles* (los colaboradores esporádicos). La Figura 7 muestra una ilustración de estas estructuras diádicas.

4.3. Metodología

En esta sección se describen los diferentes conjuntos de datos que se usan para la experimentación y se describe además cómo se construyen los grafos sociales a partir de éstos.

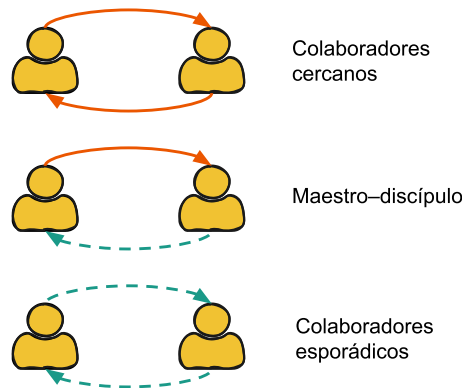


Figura 7. Tres estructuras diádicas. Arriba: los colaboradores cercanos, caracterizados por vínculos fuertes recíprocos o un *vínculo completamente fuerte*. Centro: el maestro-discípulo, caracterizados por un vínculo fuerte y un vínculo débil en direcciones opuestas, también llamado como *vínculo parcialmente fuerte*. Abajo: los colaboradores esporádicos, caracterizados por vínculos débiles recíprocos o un *vínculo completamente débil*. Un vínculo fuerte (débil) está representado mediante una línea continua (discontinua).

4.3.1. Datasets

El uso de conjuntos de datos disponibles públicamente sobre redes de expertos se ha convertido en una práctica habitual en la literatura sobre PFE. Se eligen tres conjuntos de datos ampliamente utilizados para construir los grafos de redes sociales, así como para probar y analizar la propuesta de las aristas dirigidas. En las siguientes subsecciones, se describen brevemente los conjuntos de datos seleccionados y los detalles sobre cómo se delimitaron los mismos.

4.3.1.1. DBLP

El DBLP es un sistema de información bibliográfica de publicaciones de revistas o congresos de Ciencias de la Computación (<https://dblp.org/>). Entre otra información, el DBLP alberga datos sobre los autores investigadores y sus artículos. Para este estudio se limitan los datos a las publicaciones de las siguientes conferencias importantes: SIGMOD, VLDB, ICDE, ICDT, EDBT, PODS, WWW, KDD, SDM, PKDD, ICDM, WSDM, ICML, COLT, UAI, IJCAI, NIPS⁶, CVPR, SODA, FOCS, STOC, STACS, ICALP, ESA. Sólo se consideran aquellos autores que tienen al menos tres publicaciones en cualquier combinación de estas conferencias.

⁶Ahora se denomina NeurIPS, pero en DBLP se sigue etiquetado como en su antiguo acrónimo.

4.3.1.2. GitHub

Es uno de los mayores sistemas de control de versiones para el seguimiento de los cambios en proyectos de código colaborativo (<https://github.com/>). Se restringen los datos a los cien mil repositorios con más de dos “estrellas”—equivalente al “me gusta” comunmente usado en otras redes sociales—recibidas entre enero y mayo de 2016. Se extrajeron los 30 principales usuarios colaboradores de cada uno de esos repositorios. A continuación, se mantienen sólo aquellos usuarios que han colaborado en tres o más de estos repositorios.

4.3.1.3. IMDb

La IMDb es un servicio en línea que recoge información de películas y series de televisión (<https://www.imdb.com/>). Se restringen los datos a las producciones cinematográficas de 2000 a 2015, omitiendo las películas etiquetadas como cortometrajes. De cada película, se extrajo la información sobre su reparto y equipo, incluidos los directores y productores, a los que se hace referencia simplemente como cineastas. A continuación, del conjunto de cineastas, sólo se tomaron los que figuran en la producción de al menos tres películas.

4.3.2. Construyendo el grafo

Como se menciona anteriormente, las redes sociales se representan típicamente utilizando una estructura de grafos. Recordando la noción y la notación estándar de estas estructuras para el PFE como originalmente se plantean en (Lappas *et al.*, 2009). Sea $G = (\mathcal{X}, E)$ un grafo ponderado que codifica una red social de expertos. El conjunto de nodos $\mathcal{X} = \{1, 2, \dots, n\}$ representa los índices de cada uno de los n candidatos. El conjunto de aristas $E = \{(i, i')\}$, representa las colaboraciones por pares de los candidatos $i, i' \in \mathcal{X}$. Además, suele existir una función $w : E \rightarrow \mathbb{R}$ que asigna un peso no negativo de costo de comunicación (o fuerza de colaboración) a cada arista. En este caso, se considera la fuerza de colaboración a partir de las funciones de interacción $w_s(\cdot)$, $w_J(\cdot)$ y $w_{\rightarrow}(\cdot)$ de las ecuaciones (36), (37) y (35), respectivamente.

A partir de los conjuntos de datos, se construyen tres grafos G_{dblp} , G_{git} y G_{imdb} . Los

nodos de cada grafo corresponden a cada uno de los autores, desarrolladores y cineastas, respectivamente. Aquellos candidatos que no tengan un vínculo se removieron del grafo, ya que sólo interesa estudiar las interacciones. Más aún, sólo se consideran vinculados a aquellos candidatos que comparten tres o más colaboraciones. Es decir, cuyo $|P_A \cap P_B| \geq 3, \forall (A, B) \in E$. La elección de este umbral es arbitraria, pero sustentada en la intuición de que los vínculos se construyen con algo más que un par de colaboraciones por coincidencia.

4.3.2.1. Clasificando vínculos por su fuerza

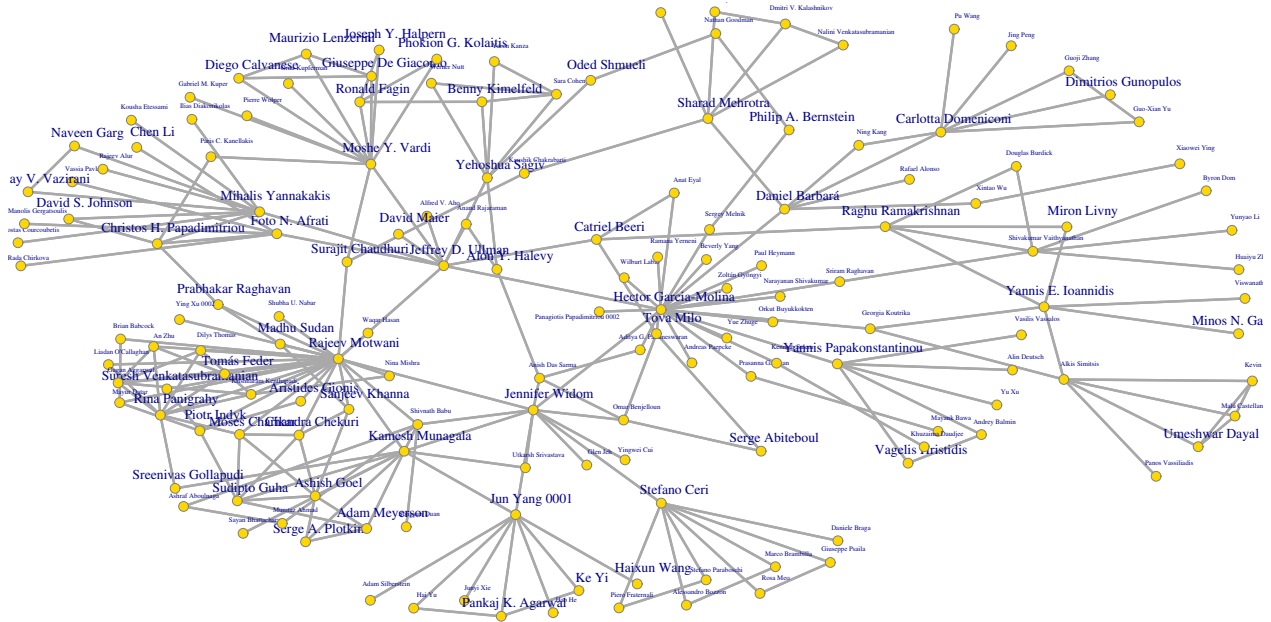
Determinar si un vínculo es fuerte o débil es un factor clave en este estudio. Para esta tarea, se utiliza el método de agrupación conocido como k -medias (Lloyd, 1982). En particular, por cada grafo, tomando en cuenta el conjunto de aristas E y la función de interacción $w_{\rightarrow}$, se usa k -medias con $k = 2$ para clasificar las aristas en dos categorías: fuerte o débil. El objetivo del algoritmo es optimizar la expresión (38).

$$\arg \min_S \sum_{i=1}^k \sum_{e \in S_i} (w_{\rightarrow}(e) - \mu_i)^2, \quad (38)$$

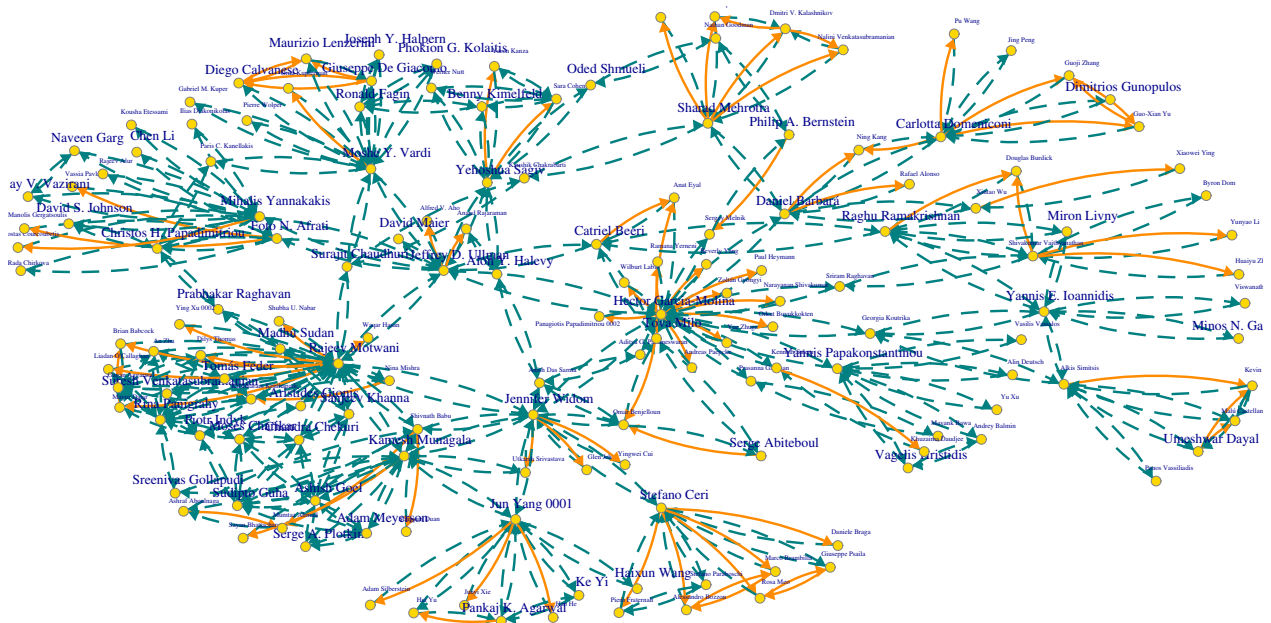
donde $S = \{S_1, \dots, S_k\}$ es una partición del conjunto de aristas E , $\mu_i = \sum_{e \in S_i} \frac{w_{\rightarrow}(e)}{|S_i|}$ es el centroide del subconjunto de aristas S_i con $i = \{1, \dots, k\}$, y k número de clases (que para este caso $k = 2$). La salida es el conjunto de aristas E , donde cada arista se etiqueta según una de las dos clases: fuerte o débil.

La Figura 8 ilustra la diferencia entre un grafo DBLP no dirigido y un grafo DBLP que implementa la propuesta de aristas dirigidas y su clasificación por su fuerza de vínculo. Por supuesto sólo se muestra una parte del grafo DBLP, particularmente es un subgrafo de la red en torno al Profesor Héctor García-Molina, situado en el centro. Note que en la Figura 8b se puede visualizar inmediatamente las relaciones llamadas colaboradores cercanos, maestro-discípulo y colaboradores esporádicos⁷, a diferencia de lo que se puede ver en la Figura 8a. Aunque no se muestran aquí, se observan patrones similares en los gráficos de los otros dos conjuntos de datos.

⁷Algunos términos como maestro-discípulo pueden parecer más adecuados para el DBLP que para los otros dos conjuntos de datos. Esto se debe a la facilidad de lectura, ya que la forma técnica más general de referirse a ellos sería *totalmente débil* para los colaboradores esporádicos, *parcialmente fuerte* para maestro-discípulo, y *totalmente fuerte* para los colaboradores cercanos, y una combinación de éstos para las estructuras triádicas (más adelante en la sección 4.4.4).



(a) Subgrafo no dirigido. Las aristas representan un vínculo entre pares de nodos.



(b) Subgrafo dirigido con diferenciación de la fuerza de los vínculos. Las aristas continuas (en naranja) representan vínculos fuertes y las aristas discontinuas (en cian) representan vínculos débiles.

Figura 8. Dos representaciones (dirigida, no-dirigida) de un subgrafo DBLP de la red en torno al Profesor Héctor García-Molina.

4.4. Análisis y resultados

4.4.1. Vínculos fuertes y débiles

El tiempo es un recurso limitado, y los vínculos fuertes exigen un mayor tiempo de mantenimiento que los vínculos débiles (Easley y Kleinberg, 2010). Es por eso que las redes sociales suelen presentar un número relativamente escaso de vínculos fuertes en comparación con su número de vínculos débiles (Huberman *et al.*, 2008; Marlow *et al.*, 2009).

Como se menciona en la sección anterior, utilizando k -medias (Lloyd, 1982), se clasifican las aristas de cada grafo social como fuertes o débiles, en función de su fuerza de interacción dada la función propuesta $w_{\rightarrow}(\cdot)$. Para comparar, de igual forma se clasifican las aristas de cada grafo pero esta vez en función del índice de Jaccard y del índice simple, de forma independiente. En particular, sustituyendo $w_{\rightarrow}(\cdot)$ por $w_s(\cdot)$ y $w_j(\cdot)$ en la ecuación (38), correspondientemente. La Tabla 5 muestra las proporciones resultantes de vínculos fuertes y débiles por cada conjunto de datos, por cada modelo de fuerza de interacción. En general, se observa que cada clasificación produjo mayor número de vínculos débiles que fuertes. Esto es consistente con lo que generalmente se espera de un conjunto de vínculos en una red social. Por lo tanto, se puede decir que k -medias puede discriminar adecuadamente los vínculos fuertes y los vínculos débiles.

Tabla 5. Resumen en números de los grafos por conjunto de datos. Las clasificaciones de fuerza de vínculo se calcularon con k -medias, por cada conjunto de datos, por cada modelo de interacción, de manera independiente. Se usa el conjunto de aristas no dirigidas E para los modelos simétricos (simple y Jaccard) y $\vec{E}$ para la propuesta.

G	$\mathcal{X}$	E	$\vec{E}$	Fuerza del vínculo					
				Simple		Jaccard		Propuesta	
				fuerte	débil	fuerte	débil	fuerte	débil
DBLP	6985	8958	17916	17 %	83 %	24 %	76 %	35 %	65 %
GitHub	29802	153593	307186	5 %	95 %	24 %	76 %	39 %	61 %
IMDb	22506	35547	71094	11 %	89 %	24 %	76 %	36 %	64 %

Está claro que la medida propuesta ($w_{\rightarrow}(\cdot)$) produce valores de fuerza de interacción no simétricos, a diferencia del índice de Jaccard y el índice simple. Sin embargo, la simetría y la no simetría en estos índices también tiene un impacto en la discriminación de los vínculos. En la Tabla 5 se observa que la medida propuesta produjo una mayor proporción de vínculos fuertes comparado con los del índice de Jaccard y los del

índice simple, en los tres conjuntos de datos. Esto puede explicarse porque algunos de los vínculos de la medida de Jaccard (o del simple) pueden clasificarse como débiles debido a un número muy pequeño de colaboraciones ($P_A \cap P_B$) o a un número muy grande de proyectos ($P_A \cup P_B$), sin tomar en cuenta el número relativo de proyectos de cada individuo. En otras palabras, las medidas simétricas ignoran la contribución individual para calcular la fuerza de los vínculos.

Por ejemplo, considerando el caso donde $|P_A| = 3$, $|P_B| = 10$, $|P_A \cap P_B| = 3$ y $|P_A \cup P_B| = 10$. Esto conduce a una fuerza de vínculo de $w_s(A, B) = 3$ y $w_j(A, B) = 0.3$ para las medidas simple y de Jaccard, respectivamente. Suponiendo que los umbrales correspondientes para considerar un vínculo como fuerte son superiores a 3 y 0.3; esto clasificaría en ambos casos a la arista $A-B$ como un vínculo débil. Sin embargo, nótese que en todos los proyectos en los que ha participado A , también ha participado B , por lo que se puede decir que B tiene una fuerte influencia sobre A . La medida propuesta, en cambio, produce un $w_{\rightarrow}(A, B) = 0.3$ y un $w_{\rightarrow}(B, A) = 1$ que puede interpretarse como un vínculo débil de A hacia B y un vínculo fuerte de B hacia A , o cualitativamente como una estructura maestro-discípulo. Estos resultados muestran que la medida no simétrica propuesta para determinar la fuerza de los vínculos habilita una capa adicional de información, de otro modo inalcanzable por las medidas simétricas existentes.

La Figura 9 muestra una gráfica con la proporción de vínculos *totalmente fuertes* (fuertes-fuertes), *parcialmente fuertes* (fuertes-débiles), y *totalmente débiles* (débiles-débiles) de los grafos DBLP, GitHub e IMDb, usando el método propuesto. En la gráfica se puede observar que los vínculos parcialmente fuertes representan más de un tercio de todas las estructuras diádicas en cualquiera de los conjuntos de datos. Esto añade evidencia a lo que se menciona anteriormente acerca de que las medidas no simétricas proporcionan más información sobre la fuerza de los vínculos.

Más aún, al revisar los vínculos correspondientes a las relaciones *totalmente fuertes*, se descubre que 0 % (0 de 939) de éstos son vínculos débiles con respecto al índice de Jaccard en cuanto al conjunto de datos del DBLP, 4.8 % (1498 de 30707) en GitHub, y 1.7 % en IMDb. En consecuencia, sólo un 3 % (110 de 3567) de los vínculos en una relación *totalmente débil* son vínculos fuertes con respecto al índice de Jaccard en el DBLP, y un notable 0 % (0 de 63206 y 0 de 17134) en los conjuntos de datos de GitHub y de IMDb. Estos porcentajes pueden interpretarse como que, cualquier vínculo débil

(fuerte) en una relación totalmente débil (fuerte), probablemente también sea débil (fuerte) en lo que respecta al índice de Jaccard. En general, estos resultados destacan la consistencia de la propuesta.

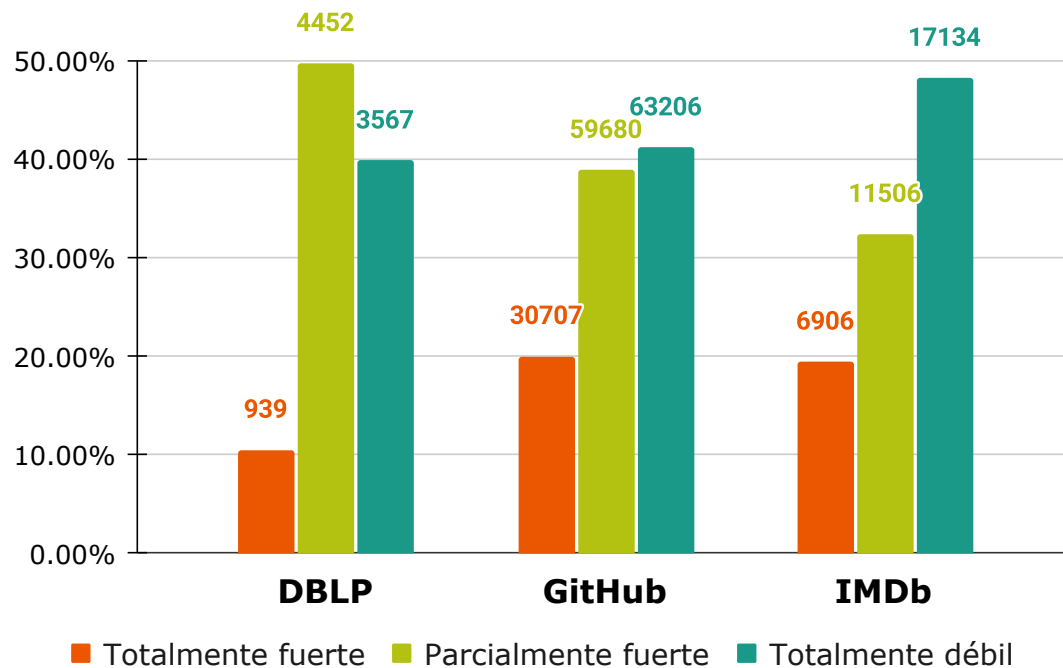


Figura 9. Proporción de vínculos *completamente fuertes* (fuerte–fuerte), *parcialmente fuertes* (fuerte–débil), *completamente débiles* (débil–débil) en los grafos DBLP, GitHub e IMDb.

4.4.2. Puentes locales

A diferencia de otros tipos de redes—como las espaciales o las biológicas—las redes sociales dependen en gran medida de sus vínculos débiles para la conectividad. Como se menciona anteriormente, la teoría de los vínculos débiles de Granovetter predice que los puentes locales en una red social consisten principalmente en vínculos débiles (Granovetter, 1973). Un puente local sería cualquier arista que conecte a dos individuos A y B de forma que la intersección de su vecindario es cero. Formalmente, la *intersección de vecindario* en un grafo está dada por la expresión $\frac{|N_A \cap N_B|}{|N_A \cup N_B \setminus \{A, B\}|}$, siendo N_X el conjunto de nodos adyacentes al individuo X .

La Figura 10 muestra una gráfica con la proporción de *puentes locales totalmente fuertes*, *puentes locales parcialmente fuertes* y *puentes locales totalmente débiles* generados con el método propuesto por cada uno de los grafos sociales. Particularmente, se identifican aquellos pares de candidatos cuya intersección de vecindario es igual

a cero, sin contar aquellos pares aislados (es decir, A, B cuyo $N_A \cup N_B = \{A, B\}$) puesto que se consideran atípicos en una red social. A partir de esta gráfica, también se puede observar que los puentes locales consisten predominantemente en vínculos de relaciones totalmente débiles: el 84.2 % (1101 de 1307) en DBLP, el 86.6 % (2114 de 2441) en GitHub y el 89.2 % (2014 de 2257) en IMDb. En consecuencia, los vínculos totalmente fuertes contribuyen con menos del 1.2 % de los puentes locales, mientras que los vínculos parcialmente fuertes suponen entre el 10 % y el 15 % en cualquiera de los tres grafos sociales. Estos resultados siguen el comportamiento previsto por la teoría de los vínculos débiles en las redes sociales (Granovetter, 1973).

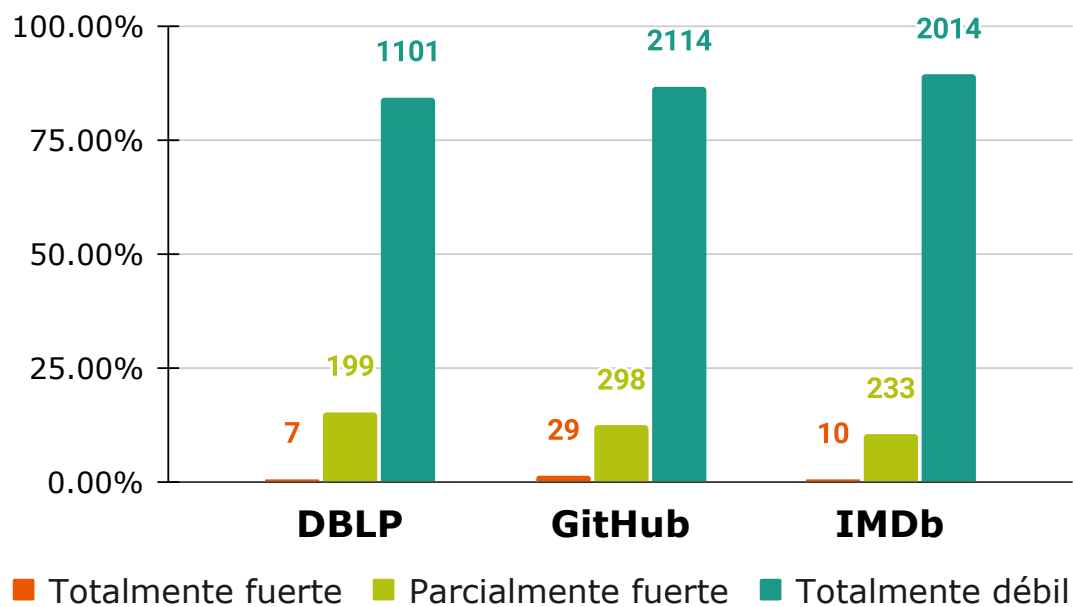


Figura 10. Proporción de los puentes locales *completamente fuertes*, *parcialmente fuertes* y *completamente fuertes* en los grafos DBLP, GitHub e IMDb.

Adicionalmente, si se combinan los resultados de la Figura 9 y la Figura 10 se puede observar que 0.74 % (7 de 939) de los vínculos totalmente fuertes, 4.4 % (199 de 4452) de los vínculos parcialmente fuertes, y 30.8 % (1101 de 3567) de los vínculos totalmente débiles, son puentes locales en relación con el grafo social DBLP. Del mismo modo, en el conjunto de datos de GitHub, 0.09 % (29 de 30707), 0.49 % (298 de 59680) y 3.3 % (2114 de 63206) de los vínculos totalmente fuertes, parcialmente fuertes y totalmente débiles, respectivamente, son puentes locales; mientras que, en IMDb, 0.14 % (10 de 6906), 2.0 % (233 de 11506) y 11.8 % (2014 de 17134) de los vínculos totalmente fuertes, parcialmente fuertes y totalmente débiles, respectiva-

mente, son puentes locales. Estos porcentajes indican que hay una mayor probabilidad de encontrar un puente local eligiendo al azar cualquier enlace totalmente débil, que la probabilidad combinada de encontrar un puente local eligiendo al azar cualquier enlace parcialmente fuerte o cualquier enlace totalmente fuerte.

En resumen, los puentes locales desempeñan un papel importante en la estructura de una red social, ya que conectan comunidades que, de otro modo, estarían más alejadas. Con el método propuesto es posible identificar en qué subconjunto de aristas es más probable encontrar los puentes locales. Esto no puede lograrse directamente con las medidas simétricas debido a lo que se explica en la subsección anterior: que las aristas no dirigidas pueden clasificarse como fuertes o débiles, pero la distinción es más sutil cuando se consideran ambas direcciones.

En lo que respecta a la formación de equipos, puede ser ventajoso que los miembros del equipo tengan vínculos *totalmente fuertes* entre sí y que al mismo tiempo tengan vínculos *totalmente débiles* con sus colegas o compañeros en su red exterior. Esto sirve para dos propósitos: uno es estimular un mejor ambiente de comunicación entre los miembros del equipo, y el otro es maximizar (la probabilidad de) los puentes locales conectados al equipo para promover el flujo de información clave desde el exterior.

4.4.3. Rumbo a la identificación de porteros tecnológicos

Como se menciona en la sección 4.1.2.2, un portero tecnológico es una persona que canaliza información técnica de fuentes externas y la pone en el contexto de su organización. Estos porteros suelen estar situados en el centro de la red de comunicación de una organización, lo que los convierte en una pieza clave de la estructura de su red. Más aún, Allen *et al.* (1984) observaron que aquellos equipos con un portero tecnológico como miembro, mostraron un mejor desempeño en sus actividades que aquellos equipos sin un portero tecnológico.

Para identificar a los porteros de una organización se requiere encuestar al personal sobre la frecuencia y relevancia de la comunicación entre sus compañeros. Sin embargo, este método no es escalable. Suponga por ejemplo, que se quiere identificar a los posibles porteros tecnológicos en una red social para incorporar alguno a un

equipo. Encuestar una red social muy grande no siempre es práctico ni posible. Como alternativa, se puede definir un portero en un grafo como un nodo con un valor alto en alguna medida de centralidad, como la centralidad de intermediación. La *centralidad de intermediación* de un nodo es el número de caminos más cortos que pasan por éste, considerando todos los pares de caminos más cortos del grafo. Sin embargo, este cálculo es relativamente costoso. El algoritmo de Brandes para la centralidad de intermediación tiene una complejidad de tiempo de $O(|V||E| + |V|^2 \log|V|)$ para grafos ponderados (o $O(|V||E|)$ para grafos no ponderados) y una complejidad de espacio de $O(|V| + |E|)$ (Brandes, 2001). Por lo tanto, cualquier análisis que utilice este método sigue estando limitado a grafos sociales con hasta unas pocas decenas de miles de nodos.

Tabla 6. Tabla de correlaciones de centralidad de intermediación y recuento de aristas por tipo de fuerza de vínculo.

Tipo de fuerza de vínculo	DBLP	GitHub	IMDb
Jaccard débil	0.42	0.26	0.45
Jaccard fuerte	-0.33	-0.32	-0.44
Propuesta completamente débil	0.49	0.38	0.56
Propuesta parcialmente fuerte	-0.04	0.00	-0.13
Propuesta completamente fuerte	-0.24	-0.33	-0.40
Propuesta débil (saliente)	0.44	0.17	0.49
Propuesta fuerte (saliente)	-0.04	0.02	-0.24
Propuesta débil (entrante)	0.39	0.38	0.45
Propuesta fuerte (entrante)	-0.27	-0.43	-0.42

En cambio, a continuación se muestra que es posible identificar porteros potenciales mediante los recuentos de tipos de vínculos (débiles o fuertes) de cada nodo. Note que un recuento de vínculos (aristas) por tipo, entre todos los nodos, toma sólo $O(|E|)$ u.t.⁸, y además es una tarea fácilmente paralelizable. La Tabla 6 muestra la correlación de Spearman entre los valores centralidad de intermediación de un nodo y su recuento de tipos de vínculos incidentes. A partir de esta tabla, se puede ver que el recuento de relaciones *totalmente débiles* tiene la mayor correlación con la centralidad de intermediación en los tres conjuntos de datos. Esto no es una coincidencia ya que parte de este resultado se explica por la tendencia de los puentes locales a ser vínculos totalmente débiles. Una vez más, los vínculos totalmente débiles parecen desempeñar un papel importante en la red. Pero además de eso, también se puede observar cuántas más variables hay para analizar si se emplea la medida no simétri-

⁸Unidades de tiempo.

ca propuesta en lugar de una medida simétrica. En particular, la medida de Jaccard sólo puede proporcionar dos variables para el recuento de las aristas según el tipo de fuerza de los vínculos, mientras que la medida propuesta proporciona hasta siete variables diferentes para la misma tarea. Puede darse el caso de que una combinación lineal (o no lineal) de estas variables tenga una mejor correlación con la centralidad de intermediación, o quizás con otra definición más adecuada de portero tecnológico. Esto se puede hacer por ejemplo mediante un análisis de correlación múltiple.

4.4.4. Análisis de estructuras 3-clique

El método propuesto también demuestra su utilidad para el análisis de estructuras triádicas (o 3-cliques). Si se clasifica cada posible 3-clique en función de la combinación de los vínculos fuertes y débiles de sus miembros, se distinguen siete familias. Cada familia de 3-cliques está determinada por el número total de vínculos fuertes (o débiles) que contiene: la familia de cero vínculos fuertes, la familia de un vínculo fuerte, y así sucesivamente hasta la familia de seis vínculos fuertes. Adicionalmente, cada familia de 3-cliques puede tener una o cuatro estructuras dependiendo de la configuración de los vínculos fuertes y débiles. Por ejemplo, sólo hay una configuración posible de vínculos fuertes en una familia de seis vínculos fuertes, mientras que en una familia de dos vínculos fuertes, hay cuatro configuraciones posibles. En total existen siete familias y 16 configuraciones estructurales posibles como se muestra en la Figura 11. Estas configuraciones pueden tener un mayor o menor sentido en la práctica según el contexto en el que residan estas estructuras, pero para fines de este trabajo se consideran meramente como estructuras abstractas formadas a partir de la combinación de las configuraciones de sus vínculos.

Algunas configuraciones son más comunes que otras. La Figura 12 muestra dos configuraciones, una es una configuración común en el mundo real (el gran maestro–maestro–discípulo) mientras que la otra es virtualmente imposible (un ciclo maestro–discípulo). En la configuración gran maestro–maestro–discípulo se observa una fuerte influencia del gran maestro sobre el maestro y el discípulo, y a su vez una fuerte influencia del maestro sobre el discípulo. Mientras que en la configuración del ciclo maestro–discípulo, se observa un ciclo de fuerte influencia en una sola dirección y un ciclo de débil influencia en la dirección opuesta, lo que tiene poco sentido en un

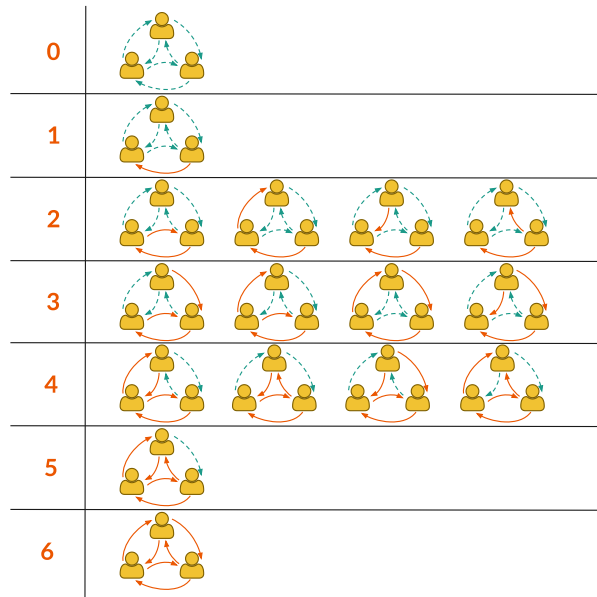


Figura 11. Las siete familias y las 16 configuraciones posibles de 3-cliques considerando una distinción de los vínculos como fuertes (línea continua color naranja) y débiles (línea punteada color cian). En el lado izquierdo se indica el número de vínculos fuerte en la estructura; en el lado derecho, cada posible estructura según la combinación de las configuraciones de sus vínculos.

entorno de colaboración real. Más aún, de acuerdo al Teorema 4.1, el ciclo maestro-discípulo representa una configuración que no es posible obtener dada la propuesta de interacción y un cierto umbral.

Teorema 4.1 Sea $w_{\rightarrow}$ la función de interacción definida por la ecuación (35) y un umbral $\lambda \in (0, 1)$, donde $w_{\rightarrow}(A, B) > \lambda$ se considera un vínculo fuerte de la persona A hacia la persona B y $w_{\rightarrow}(A, B) \leq \lambda$ se considera un vínculo débil de A hacia B, entonces la configuración del tipo ciclo maestro-discípulo (como se observa en la derecha de la Figura 12) no es factible.

Demostración. Se demuestra directamente. Sean A, B y C tres personas cuyos vínculos están clasificados de la siguiente manera: $A \rightarrow_f B$, $B \rightarrow_f C$, $C \rightarrow_f A$, $A \rightarrow_d C$, $C \rightarrow_d B$ y $B \rightarrow_d A$, donde $X \rightarrow_f Y$ indica un vínculo fuerte de X hacia Y y $X \rightarrow_d Y$ indica un vínculo débil de X hacia Y. Por definición, esto significa que para el par A, B tiene un $w_{\rightarrow}(A, B) = \frac{|P_A \cap P_B|}{|P_B|} > \lambda$ y un $w_{\rightarrow}(B, A) = \frac{|P_A \cap P_B|}{|P_A|} \leq \lambda$. Se despeja $|P_A \cap P_B|$ en ambos casos, quedando como $|P_A \cap P_B| > |P_B|\lambda$ (i) y $|P_A \cap P_B| \leq |P_A|\lambda$ (ii). Se hace lo mismo con los pares B, C y C, A quedando como $|P_B \cap P_C| > |P_C|\lambda$ (iii), $|P_B \cap P_C| \leq |P_B|\lambda$ (iv) y $|P_A \cap P_C| > |P_A|\lambda$ (v), $|P_A \cap P_C| \leq |P_C|\lambda$ (vi), respectivamente. Luego, combinando las desigualdades (ii) y (v), se obtiene $|P_A \cap P_B| \leq |P_A|\lambda < |P_A \cap P_C|$ (vii). De

la misma forma, combinando las desigualdades (i) y (iv), y (iii) y (vi), se obtienen $|P_B \cap P_C| \leq |P_B|\lambda < |P_A \cap P_B|$ (viii) y $|P_A \cap P_C| \leq |P_C|\lambda < |P_B \cap P_C|$ (ix), respectivamente. Simplificando las desigualdades (vii), (viii) y (ix), se tiene que $|P_A \cap P_B| < |P_A \cap P_C|$ (x), $|P_B \cap P_C| < |P_A \cap P_B|$ (xi), y $|P_A \cap P_C| < |P_B \cap P_C|$ (xii), correspondientemente. Finalmente, al combinar las desigualdades (x), (xi) y (xii) se llega a la siguiente expresión $|P_A \cap P_B| < |P_A \cap P_C| < |P_B \cap P_C| < |P_A \cap P_B|$ que representa una desigualdad cíclica no factible. Por lo tanto, no es posible obtener en la práctica una configuración del tipo ciclo maestro-discípulo mediante la función $w_{\rightarrow}$ y un umbral $\lambda \in (0, 1)$. ■

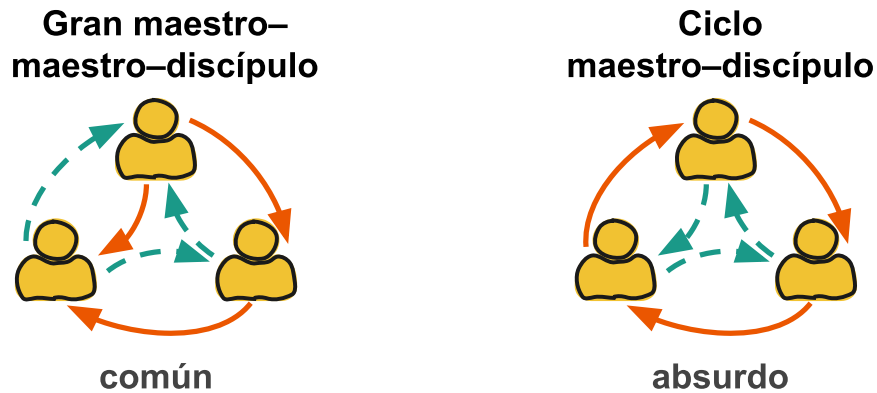


Figura 12. Ejemplo de dos configuraciones de estructuras triádicas. Izquierda: el gran maestro–maestro–discípulo, caracterizado por un nodo (el gran maestro) con dos vínculos parcialmente fuertes hacia dos nodos, estos últimos (el maestro–discípulo) con un vínculo parcialmente fuerte entre ellos. El gran maestro debe presentar dos vínculos fuertes salientes y dos vínculos débiles entrantes. Derecha: el ciclo maestro–discípulo, caracterizado por una relación cíclica de maestros–discípulos. Note que esta última configuración es hipotética, es improbable que exista en la realidad.

En cuanto al resto de las configuraciones, La Figura 13 muestra el mapa de calor de las 16 configuraciones de 3-cliques en cada uno de los conjuntos de datos. La intensidad del color azul de una celda representa la proporción de su correspondiente configuración de 3-cliques por conjunto de datos. En otras palabras, cuanto más claro es el azul de una celda, hay menos ocurrencias de esa configuración, mientras que cuanto más oscuro es el azul, hay más ocurrencias.

Adicionalmente, en el mapa de calor, se puede observar que efectivamente algunas configuraciones son más frecuentes que otras. En concreto, de izquierda a derecha, las configuraciones 1, 2, 5, 10, 13 y 16 son más frecuentes que las configuraciones 4, 8, 9, 11 y 12. Las primeras configuraciones representan las siguientes estructuras comunes: (1) los tres colaboradores esporádicos, (2) los dos colaboradores esporádicos y la relación maestro–discípulo, (5) los maestros colaboradores esporádicos y el único discípulo, (10) el gran maestro–maestro–discípulo, (13) el maestro y los dos

discípulos colaboradores cercanos, y (16) los tres colaboradores cercanos. Mientras que las últimas configuraciones representan las siguientes estructuras improbables: (4) el gran maestro–maestro–discípulo que presenta una relación esporádica de gran maestro–discípulo, (8) los colaboradores cercanos y el discípulo con una relación esporádica con el colaborador cercano del maestro, (9) el ciclo maestro–discípulo (que se demostró que no es posible obtener), (11) la persona con dos colaboradores cercanos siendo ellos colaboradores esporádicos, (12) el gran maestro–maestro–discípulo que presenta una relación de gran maestro–discípulo de colaboración cercana. Observe que las estructuras improbables son más difíciles de poner en palabras, ya que son más abstractas que reales. En otras palabras, estas configuraciones se consideran artefactos. Aunque esto no quiere decir que dichas configuraciones no existan, con excepción de la configuración maestro–discípulo, se encontró al menos un ejemplo de cada configuración posible en cada conjunto de datos.

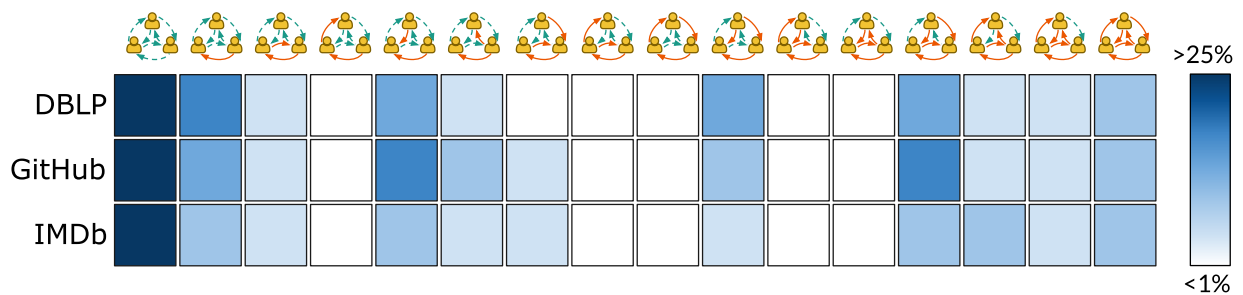


Figura 13. Mapa de calor de las configuraciones de los 3-cliques por conjunto de datos.

4.5. Discusión

En general, la propuesta de interacción no simétrica y su clasificación en cuanto su fuerza de vínculo (fuerte o débil), reveló información importante sobre las redes a diferentes escalas.

A nivel macro, se observó que cada red estudiada tiene un mayor número de vínculos débiles que vínculos fuertes. Además, se observó que los puentes locales se encuentran mayoritariamente en las relaciones totalmente débiles (es decir, relaciones conformadas por dos vínculos débiles). Ambos resultados evidencian una característica oculta en las redes, pero lo más importante es que estas observaciones son consistentes con el trabajo seminal de Granovetter y su teoría de la fuerza de los vínculos débiles Granovetter (1973).

A nivel micro, el método propuesto reveló tres tipos de estructuras diádicas basadas en la fuerza de sus vínculos: los colaboradores cercanos (vínculos fuerte–fuerte), el maestro–discípulo (vínculos fuerte–débil) y los colaboradores esporádicos (vínculos débil–débil). Adicionalmente, se observaron estructuras triádicas más complejas basadas en la configuración de sus vínculos, es decir, su dirección y fuerza. Un análisis de estas estructuras mostró que algunas configuraciones se encontraron en las tres redes, mientras que otras configuraciones son virtualmente inexistentes.

Estos resultados muestran el potencial de la propuesta para habilitar una nueva dimensión en la información, misma que permite visualizar nuevos patrones estructurales y de interacción. En contraste, las medidas simétricas (Jaccard y simple) no solo impiden analizar la interacción entre dos candidatos en ambas vías, sino que representan la interacción como un solo valor agregado. Incluso si se clasifican las aristas mediante su medida simétrica, una arista débil podría deberse meramente a un número muy pequeño de colaboraciones ($P_A \cap P_B$) o a un número muy grande de proyectos combinados ($P_A \cup P_B$), sin tomar en cuenta los números proyectos relativos de cada persona. De esta forma, sería prácticamente imposible determinar estructuras del tipo maestro–discípulo (parcialmente fuerte o fuerte–débil). Cabe resaltar que las estructuras maestro–discípulo representan una proporción mayor al 30 % de las relaciones diádicas mediante la medida no simétrica propuesta, en cualquiera de los tres conjuntos de datos usados. Y si no se puede representar este tipo de relaciones, tampoco se pueden estudiar estructuras más complejas como las estructuras triádicas (o 3-cliques). En conjunto, la inconveniencia de las medidas simétricas tiene también un posible impacto en la búsqueda eficiente de puentes locales o porteros tecnológicos.

Por otra parte, los resultados presentados en este capítulo pueden servir para el diseño de funciones objetivo más adecuadas, o para modelos más generales. También, las estructuras reveladas pueden servir para diseñar algoritmos eficientes, o aportar información útil para la toma de decisiones en escenarios multicriterio. Por ejemplo, puesto que el PFE es intrínsecamente un problema multiobjetivo (como se discute en el Capítulo 2), es probable que la solución consista en un conjunto de equipos igualmente buenos en lugar de un único equipo. Esto requiere un método eficaz para decidir qué equipo se va a formar. Esta situación pone de manifiesto la conveniencia del método propuesto para análisis posteriores a la optimización, en el que el factor

decisivo podría ser alguna estructura de configuración deseable dentro del equipo.

Finalmente, los resultados de este capítulo se obtuvieron probando con conjuntos de datos provenientes de las redes de: coautorías de investigadores en las ciencias de la computación (DBLP), de cineastas según las filmaciones en las que coincidieron (IMDb), y de desarrolladores que colaboraron en los mismos proyectos de software (GitHub). Sin embargo, la metodología propuesta se podría aplicar para analizar equipos (o comunidades) en otros escenarios como en redes de crimen organizado, en redes de negocio, o en redes de trabajo de diversos contextos. El desafío más claro es contar con datos suficientes y de una cierta calidad, y en determinar una granularidad adecuada para definir las interacciones. Las interacciones, a su vez, definen a los conjuntos P que posteriormente se usan para medir la fuerza de interacción mediante la función $w_{\rightarrow}$. Recuerde que P_A es el conjunto de proyectos realizados por el candidato A —donde un proyecto equivale a una publicación, una película, o un proyecto de desarrollo de software, según el conjunto de datos usado—, esto significa que la definición de interacción se debe adaptar al contexto de su dominio. Mientras que, también se puede refinar el conteo de interacciones a escenarios más frecuentes, por ejemplo el número de intervenciones en una plática de pasillo, el número de reuniones de trabajo, o de intercambios por correo electrónico.

Capítulo 5. Estudio comparativo de criterios de optimización desde una perspectiva multi-objetivo: densidad versus distancia

En un estudio preliminar realizado por Juárez y Brizuela (2018), se observó que algunas soluciones encontradas al optimizar para ciertos objetivos estaban, al mismo tiempo, optimizando otros objetivos que no necesariamente guardan una relación de dependencia con los primeros. Este resultado invita a cuestionar la efectividad de optimizar para unos criterios sobre otros, especialmente ante la ausencia de consenso sobre los criterios específicos para formar equipos efectivos.

En el presente capítulo se explora el impacto que tiene optimizar para un par de criterios y posteriormente evaluar para otro(s) par(es) de criterios. Particularmente, el enfoque se centra en comparar dos familias de criterios: los criterios basados en distancia de comunicación y los criterios basados en densidad de colaboración. El interés por estas dos familias se basa en un debate de la literatura del Problema de Formación de Equipo (PFE) sobre la conveniencia de optimizar para criterios de una familia sobre la otra.

5.1. Preliminares

5.1.1. Criterios de optimización para el Problema de Formación de Equipos Basados en Comunidades

Existen varios criterios de optimización que se han estudiado en la literatura para el PFE basado en comunidades. Las dos familias de criterios principales son las que recién mencionadas: los criterios basados en distancia de comunicación y los criterios basados en densidad de colaboración.

Los criterios de ambas familias requieren como entrada un equipo cuyos integrantes forman parte de un grafo social, recordando rápidamente la notación de estas estructuras. Sea $G = (\mathcal{X}, E)$ un grafo ponderado que codifica una red social de expertos. El conjunto de nodos $\mathcal{X} = \{1, 2, \dots, n\}$ representa los índices de cada uno de los n candidatos. El conjunto de aristas $E = \{(A, B)\}$, representa las colaboraciones por

pares de los candidatos $A, B \in \mathcal{X}$. Además, suele existir una función $w : E \rightarrow \mathbb{R}$ que asigna un peso no negativo de fuerza de colaboración (o $c : E \rightarrow \mathbb{R}$ para costo de comunicación) a cada arista. Finalmente, un equipo se denota como $\mathcal{X}' \subseteq \mathcal{X}$, mientras que $G[\mathcal{X}]$ representa el subgrafo inducido por los nodos del equipo $\mathcal{X}'$ sobre el grafo original G .

5.1.1.1. Criterios basados en distancia

Estos criterios toman como medida de interacción un costo de comunicación entre pares de individuos como los que se discuten las secciones 4.1.1 y 4.2.

Algunos de estos criterios se enlistan a continuación:

- $\text{Cc-R}(\mathcal{X})$. El costo de comunicación en un radio de distancia. Se define como el camino más largo entre los caminos más cortos entre cualesquier par de miembros del equipo, es decir, $\text{Cc-R}(\mathcal{X}) = \max_{A, B \in \mathcal{X}} d(A, B)$ (Lappas *et al.*, 2009).
- $\text{Cc-MST}(\mathcal{X})$. El costo de comunicación en el árbol de expansión de peso mínimo. Formalmente, se define como $\text{Cc-MST}(\mathcal{X}) = \sum_{(A, B) \in E_{\text{MST}(G[\mathcal{X}])}} c(A, B)$ (Lappas *et al.*, 2009).
- $\text{Cc-}\Sigma d(\mathcal{X})$. El costo de comunicación entre pares de miembros. Se define como la suma total de las distancia de comunicación entre cualesquier par de miembros del equipo, es decir, $\text{Cc-}\Sigma d = \sum_{A, B \in \mathcal{X}} d(A, B)$ (Kargar *et al.*, 2012).

Recordando que para la distancia de comunicación, $c(A, B)$ es el valor de una arista $A-B$ que representa el costo de comunicación entre la persona A y la persona B . De manera que la distancia $d(A, B)$ es el costo de las aristas en el camino más corto entre la persona A y la persona B . Note que en los criterios arriba mencionados, $d(A, B)$ se calcula sobre el subgrafo inducido por el equipo, y no en el grafo de entrada.

5.1.1.2. Criterios basados en densidad

Estos criterios toman como medida de interacción un peso de colaboración entre pares de individuos como los que se discuten en las secciones 4.1.1 y 4.2.

Los criterios principales de esta clasificación se enlistan a continuación:

- $D(\mathcal{X})$. La densidad de colaboración de un equipo. Se define como la densidad ponderada de un grafo, es decir, $D = \frac{1}{\binom{|\mathcal{X}|}{2}} \sum_{(A,B) \in E_{G[\mathcal{X}]}} w(A, B)$ (Juárez y Brizuela, 2018).
- $sD(\mathcal{X})$. La subdensidad de colaboración de un equipo. Se define como la densidad ponderada de un subgrafo, es decir, $D_{\frac{1}{|\mathcal{X}|}} = \sum_{(A,B) \in E_{G[\mathcal{X}]}} w(A, B)$ (Gajewar y Das Sarma, 2012).

Recordando que para la densidad de colaboración, $w(A, B)$ es el valor de una arista $A-B$ que representa la fuerza de vínculo entre la persona A y la persona B .

5.1.1.3. Debate entre criterios basados en distancia y criterios basados en densidad

Tanto los criterios basados en distancia como los criterios basados en densidad están definidos alrededor de conceptos muy usados para medir grafos en general. En cuanto a la formación de equipos, ambos conjuntos de criterios sirven para modelar qué tan cercanos son los miembros del equipo como equipo, y de una u otra forma qué tan efectivos pueden ser. No obstante, cuando se trata de determinar cuál de las dos familias de criterios usar, la respuesta no es tan clara.

Actualmente, existe un debate abierto en la literatura sobre cuál de estos dos enfoques representa mejor la efectividad de equipo. Esto tiene un impacto directo en los modelos de optimización de PFE puesto que implementan los criterios de uno u otro enfoque como función objetivo.

Por un lado, optimizar para objetivos basados en distancia promueve la conectividad del equipo. Es decir, una solución factible suele tener al menos un camino entre cualquier par de miembros dentro del subgrafo inducido por el equipo. Sin embargo, si las habilidades demandadas por una cierta tarea no las pueden cubrir personas cercanas en el grafo, esto obliga incluir miembros no esenciales para la tarea y que funcionen como puentes entre los miembros esenciales o incluso puede que no exista

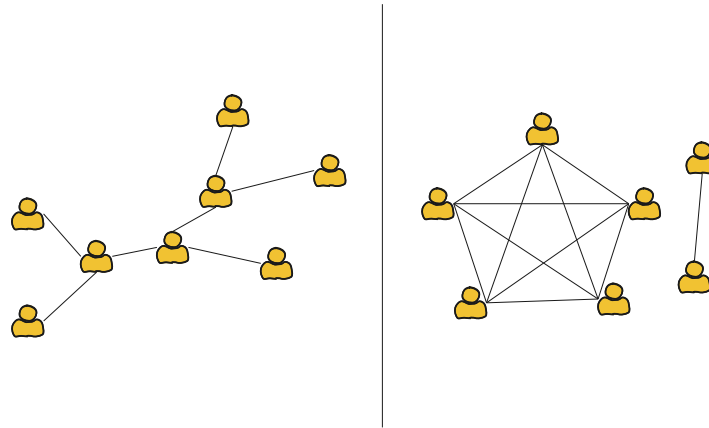


Figura 14. Izquierda: grafo social conectado, no denso. Derecha: grafo social desconectado, relativamente denso.

una solución conectada factible. Más aún, Kargar y An (2011) señalaron una desventaja importante de usar criterios como el diámetro o el costo del árbol de expansión mínimo para el PFE. La crítica está en que pequeñas variaciones al grafo de entrada pueden producir equipos radicalmente distintos.

Por otro lado, optimizar para objetivos basados en densidad promueve la saturación de conexiones entre los miembros del equipo. Dicho de otro modo, la densidad ayuda a promover la cercanía de los miembros del equipo, en promedio. La desventaja principal es que, a diferencia de los objetivos basados en distancia, los objetivos basados en densidad no garantizan la conectividad del equipo. Se puede tener un grafo con un valor de densidad relativamente alto aún cuando algún nodo se encuentre desconectado.

En la Figura 14 se puede observar un ejemplo de las ventajas y desventajas de ambos enfoques. En el lado izquierdo de la figura, se puede ver que existe un camino entre cada par de miembros del equipo, lo que representa que existe un canal de comunicación entre todo el equipo. No obstante, la baja densidad (o dispersión) en el grafo social puede representar una alta latencia en la comunicación. En el lado derecho de la figura, se puede ver un grafo con una densidad relativamente alta, especialmente en la componente con cinco miembros donde la conexión entre cualesquiera de ellos es directa. En contraparte, el equipo está desconectado.

5.1.1.4. Otros criterios de optimización

Existen otros criterios de optimización que modelan aspectos distintos a la comunicación o colaboración de los miembros de un equipo. Algunos de ellos tienen que ver con la experiencia de sus integrantes, su ubicación geográfica, o incluso el tamaño del equipo.

Entre los criterios más populares se encuentran aquellos que consideran el nivel de experiencia de los miembros del equipo. Por ejemplo,

- $Z(\mathcal{X}')$. El nivel total de experiencia relativa del equipo. Se define como la suma total de la experiencia de cada miembro por cada habilidad requerida en la tarea, dividido entre el número de miembros en el equipo. Formalmente, $Z(\mathcal{X}') = \sum_{i \in \mathcal{X}'} \sum_{(a, k_a) \in T} \frac{w_{ia}}{|\mathcal{X}'|}$, donde w_{ia} es el valor de la experiencia del candidato i sobre la habilidad a (Juárez y Brizuela, 2018).
- $C(\mathcal{X})$. El costo monetario total del equipo, es decir, $C(\mathcal{X}') = \sum_{i \in \mathcal{X}'} c_i$, donde c_i es el costo monetario de incluir al candidato i en el equipo (Pérez-Toledano *et al.*, ???).

5.1.2. Problemas de formación de equipo multiobjetivo

Como se ha discutido en capítulos anteriores, la efectividad de equipo es un concepto multiobjetivo por naturaleza, lo que hace del PFE un problema multiobjetivo. Sorprendentemente, el estudio multiobjetivo de este problema sigue siendo relativamente escaso con respecto a las versiones mono-objetivo.

El estudio de los problemas desde una perspectiva multiobjetivo es fundamentalmente diferente de su perspectiva mono-objetivo (Miettinen, 2012; Coello *et al.*, 2007). Esto se debe principalmente a que en la optimización multiobjetivo no existe una única solución óptima cuando los objetivos están en conflicto entre ellos. En su lugar, hay un conjunto de soluciones (en compromiso) que se consideran óptimas, al conjunto de todas estas soluciones se le conoce como conjunto óptimo de Pareto. Por lo tanto, no es suficiente con optimizar de forma independiente las habilidades técnicas, o la colaboración del equipo o la composición de la actitud del equipo.

En cuanto a los trabajos multiobjetivo del PFE, se suelen encontrar todo tipo de objetivos, algunos muy contextuales. Por ejemplo, en (Chen *et al.*, 2020) se desea encontrar un equipo de desarrolladores de software que minimice la distancia de comunicación y las diferencias asociadas a la ubicación geográfica (como región geográfica, lengua nativa y zona horaria). En (Das *et al.*, 2020) se desea encontrar un equipo de estudiantes que minimice la distancia de comunicación y maximice la diversidad (en cuanto a habilidades y perfiles). En (Pérez-Toledano *et al.*, ????) se desea encontrar un equipo de basquetbol que maximice la suma total de los índices de desempeño de los jugadores y al mismo tiempo que minimice el costo monetario del equipo.

A pesar de la diversidad de objetivos en estos trabajos, hay una tendencia general en considerar algún criterio de distancia de comunicación o de fuerza de colaboración (o similar), y otro que tiene que ver con otros criterios mayormente relacionados con habilidades técnicas. Otros trabajos multiobjetivo que abordan estos tipos de objetivos son (Feng *et al.*, 2010; Niveditha *et al.*, 2017; Juárez y Brizuela, 2018). Sin embargo, estos trabajos y los antes mencionados están poco o nada enfocados en estudiar el impacto de usar ciertos criterios sobre otros.

5.2. Propuesta de criterio de densidad basado en interacción no-simétrica y su implementación en un modelo multiobjetivo

Se presenta ahora una propuesta de un criterio de optimización basado en densidad de colaboración. El criterio propuesto se define por medio de la ecuación (39).

$$D_{\rightarrow}(\mathcal{X}) = \frac{1}{2^{\binom{|\mathcal{X}|}{2}}} \sum_{(A,B) \in \tilde{E}_{G[\mathcal{X}]}} w_{\rightarrow}(A, B). \quad (39)$$

Note que este criterio de optimización incorpora la propuesta no simétrica de interacción ($w_{\rightarrow}(A, B)$) introducida en el capítulo anterior. Esta densidad es básicamente la función de densidad ponderada de un grafo dirigido reemplazando $w_{\rightarrow}(A, B)$ como el peso de la arista $(A, B) \in E$.

Se decide trabajar con una densidad en vez de una distancia por dos razones principales. La primera razón está basada en el trabajo preliminar de Juárez y Brizuela

(2018). En su trabajo, se observó que en muchos casos al optimizar simultáneamente para el nivel de experiencia total de los miembros de equipo y para la densidad ponderada del equipo, el conjunto de soluciones no dominadas contenía al menos una solución con un “buen” valor de distancia de comunicación. Más aún, en varios casos, se encontraron soluciones con un mejor valor en criterio de distancia que soluciones producidas por una heurística diseñada para ello.

La segunda razón tiene que ver con la conveniencia de implementar la función no simétrica $w_{\rightarrow}(A, B)$ en un modelo de densidad a diferencia de un modelo de distancia. Calcular una densidad ponderada de un grafo dirigido es trivial. En cambio, calcular una distancia considerando aristas dirigidas requiere un algoritmo más sofisticado. Por ejemplo, calcular el árbol de expansión de costo mínimo con aristas dirigidas ya no admite un enfoque voraz (es decir, un algoritmo glotón) como en su versión de aristas no dirigidas (Gibbons, 1985). No obstante, el verdadero inconveniente viene del enfoque basado en distancias en sí mismo. Particularmente, ¿cómo definir un valor de distancia (diámetro o costo de árbol de expansión) de un equipo no conectado como el que se muestra en la derecha de la Figura 14? La alternativa de definir un símbolo como ∞ , N/A, u otros, dice muy poco de qué tan cerca está un equipo desconectado de estar conectado. Este no es el caso con un enfoque basado en densidad, puesto que un equipo conectado o desconectado siempre arroja un valor.

5.2.1. Formulación multiobjetivo

Se proponen dos objetivos a tomar en cuenta a la hora de formar un equipo efectivo: un objetivo referente a la densidad y un objetivo referente a la experiencia. Intuitivamente, uno desearía formar un equipo tan colaborativo y experimentado como sea posible. Sin embargo, ante la nula o baja probabilidad de que exista un equipo que sea el mejor en ambos criterios, el paradigma multiobjetivo ofrece una alternativa. En este sentido, el paradigma multiobjetivo motiva a encontrar al equipo más colaborativo aunque no sea el más experto, al equipo más experimentado aunque no sepa trabajar como una unidad, y a todos los otros equipos intermedios entre estos dos.

Sea $G = (\mathcal{X}, E)$ un grafo social dirigido, donde $\mathcal{X}$ es el conjunto de índices de los candidatos y E es el conjunto de aristas entre pares de candidatos. Sea $w_{\rightarrow}(A, B)$

el peso de colaboración del candidato $A \in \mathcal{X}$ con respecto al candidato $B \in \mathcal{X}$. Sea $\mathcal{A} = \{a_1, a_2, \dots, a_m\}$ el universo de habilidades existentes. Sea $w_{ia} \in \mathbb{R}_{\geq 0}$ la idoneidad del candidato $i \in \mathcal{X}$ con respecto a la habilidad $a \in \mathcal{A}$. Y sea $T = \{(a_1, k_{a_1}), (a_2, k_{a_2}), \dots, (a_m, k_{a_m})\} \subseteq \mathcal{A} \times \mathbb{N}$ una tarea general que señala el número de candiatos k_a que debe haber en el equipo por tipo de habilidad a .

La formulación del PFE multiobjetivo propuesta se define a través de la expresión (40): Dado un grafo social $G = (\mathcal{X}, E)$, y una tarea general $T = \{(a_1, k_{a_1}), (a_2, k_{a_2}), \dots, (a_m, k_{a_m})\}$,

$$\begin{aligned} & \underset{\mathcal{X}' \subseteq \mathcal{X}}{\text{maximizar}} && (D_{\rightarrow}(\mathcal{X}'), Z(\mathcal{X}')), \\ & \text{sujeto a} && |\mathcal{X}' \cap S(a)| \geq k_a, \\ & && \forall (a, k_a) \in T, \end{aligned} \tag{40}$$

donde $S(a)$ es el conjunto de candidatos en $\mathcal{X}$ cuyo $w_{ia} > 0$, $D_{\rightarrow}(\mathcal{X}')$ la densidad propuesta, mientras que $Z(\mathcal{X}')$ es el nivel de experiencia total de los candidatos del equipo (descrito en la subsección 5.1.1.4).

5.3. Metodología

En esta sección se describe el conjunto de datos y su uso para generar el grafo social para el experimento. También se introducen los casos de prueba. Además se describe el algoritmo evolutivo empleado y los detalles correspondientes en cuanto a inicialización y operadores de variación. Finalmente se detalla el diseño del experimento.

5.3.1. Conjunto de datos y el grafo social

El grafo social se construye utilizando los datos del DBLP, un sistema de información bibliográfica de artículos relacionados con ciencias de la computación, introducido en la sección 4.3.1.1. La metodología general para construir este grafo a partir del DBLP se define en (Lappas *et al.*, 2009; Gajewar y Das Sarma, 2012).

Se utilizan los datos XML del DBLP¹ de un respaldo etiquetado con el 2 de marzo del 2015, extrayendo únicamente la información bibliográfica de 16 conferencias seleccio-

¹<http://dblp.uni-tier.de/xml/>

nadas. Cada conferencia se clasifica dentro de una de las siguientes áreas: Inteligencia Artificial (IA), Base de Datos (BD), Minería de Datos (MD) y Teoría (T). Las conferencias y su clasificación van de la siguiente manera: $IA = \{ICML, ECML, COLT, UAI\}$, $BD = \{SIGMOD, VLDB, ICDE, ICDT\}$, $MD = \{WWW, KDD, SDM, PKDD\}$, y $T = \{SODA, FOCS, STOC, STACS\}$. Sea $G = (\mathcal{X}, E)$ el grafo que se desea formar. El conjunto de candidatos $\mathcal{X}$ está definido por aquellos autores expertos en al menos una de las cuatro áreas. Se considera que un autor es experto en una de estas áreas si tiene tres o más publicaciones en alguna conferencia del área correspondiente. Un autor puede ser experto en más de un área. Si un autor $i \in \mathcal{X}$ es experto en IA, entonces se dice que $i \in S(IA)$ e $IA \in X_i$; lo mismo es para BD, MD y T. El conjunto de aristas E está definido por aquellos pares de autores que comparten un cierto número de publicaciones. Particularmente, existe una arista $(i, j) \in E$ entre los autores i, j si son coautores en al menos dos publicaciones en las 16 conferencias seleccionadas. Finalmente, el nivel de experiencia w_{ia} de un candidato $i \in \mathcal{X}$ sobre un área $a \in \{IA, BD, MD, T\}$ está definido por el número total de publicaciones en las conferencias del área correspondiente.

5.3.2. Casos de prueba

Como se menciona anteriormente, un caso de prueba está definido por un grafo y una tarea. Una tarea general $\mathcal{T} = \{(a_1, k_{a_1}), (a_2, k_{a_2}), \dots, (a_m, k_{a_m})\}$ consiste en un conjunto de pares (a, k_a) que define la habilidad a requerida y el número k_a de candidatos que la posean. Adicionalmente, se define el tamaño K de una tarea $\mathcal{T}$ como la suma de candidatos requeridos por habilidad², es decir, $K = \sum_{(a, k_a) \in \mathcal{T}} k_a$.

Para los experimentos, se adoptaron 15 casos de prueba propuestos en (Juárez y Brizuela, 2018). Éstos consisten en cinco tareas por cada tamaño de $K = \{4, 8, 12\}$, como se muestra en el lado izquierdo de la Tabla 7. Adicionalmente, se generaron otros 15 casos de prueba divididos en cinco tareas por cada tamaño de $K = \{14, 16, 18\}$, como se muestra en el lado derecho de la Tabla 7. Ambos conjuntos de tareas se generaron de manera aleatoria.

²Note que el tamaño de una tarea no necesariamente refleja el número de candidatos requeridos en un equipo. Pueden existir candidatos con una o más habilidades.

Tabla 7. Benchmark de casos de prueba generados de manera aleatoria.

Tarea	K	AI	DM	DB	T	Tarea	K	AI	DM	DB	T
Caso 1	4	1	1	0	2	Caso 16	14	0	7	4	3
Caso 2	4	1	1	2	0	Caso 17	14	4	3	3	4
Caso 3	4	2	0	0	2	Caso 18	14	2	1	5	6
Caso 4	4	0	3	1	0	Caso 19	14	3	3	6	2
Caso 5	4	2	1	1	0	Caso 20	14	4	4	6	0
Caso 6	8	3	2	3	0	Caso 21	16	2	4	8	2
Caso 7	8	1	2	4	1	Caso 22	16	4	4	2	6
Caso 8	8	2	2	2	2	Caso 23	16	5	3	2	6
Caso 9	8	2	3	2	1	Caso 24	16	4	4	3	5
Caso 10	8	2	1	5	0	Caso 25	16	5	2	2	7
Caso 11	12	4	2	3	3	Caso 26	18	3	4	5	6
Caso 12	12	3	3	3	3	Caso 27	18	4	9	4	1
Caso 13	12	2	3	5	2	Caso 28	18	4	10	4	0
Caso 14	12	4	0	6	2	Caso 29	18	7	7	0	4
Caso 15	12	1	5	4	2	Caso 30	18	4	3	2	9

5.3.3. Algoritmo evolutivo multi-objetivo

Para resolver el PFE multiobjetivo propuesto en la sección anterior, se emplea una metaheurística conocida como NSGA-II (Deb *et al.*, 2002). El NSGA-II es un algoritmo genético multiobjetivo que, a lo largo de los años, ha demostrado empíricamente su utilidad para resolver una amplia variedad de problemas combinatorios multiobjetivo (Verma *et al.*, 2021). Más aún, múltiples versiones del PFE multiobjetivo también han sido resueltas con este algoritmo (Feng *et al.*, 2010; Ahmed *et al.*, 2013; Niveditha *et al.*, 2017; Juárez y Brizuela, 2018; Pérez-Toledano *et al.*, 2019; Chen *et al.*, 2020; Das *et al.*, 2020).

En general, los algoritmos genéticos son altamente personalizables, el esquema de los algoritmos es simplemente una columna vertebral con marcadores de posición para los operadores. Los componentes de un algoritmo genético típico son en su mayoría inherentes al problema. La representación de los individuos, el procedimiento de inicialización y los operadores de recombinación y mutación que se propusieron, se describen a lo largo de esta sección. El método de selección es el torneo binario basado en la aglomeración, y el reemplazo también es nativo del NSGA-II.

5.3.3.1. Representación de la solución

Para este PFE multiobjetivo, una solución o individuo³ $\mathcal{I}$ se representa de manera directa como la definición de equipo. Dado un grafo $G = (\mathcal{X}, E)$ y una tarea T , una *solución factible* $\mathcal{I}$ es un subconjunto de $\mathcal{X}$ tal que $|\mathcal{I} \cap S(a)| \geq k_a$, para todo $(a, k_a) \in T$.

Como ejemplo, suponga que una organización necesita formar un equipo para una tarea definida por las siguientes habilidades $T = \{(s_1, 1), (s_2, 3), (s_3, 1)\}$. Además, suponga un conjunto de candidatos disponibles $\mathcal{X} = \{a, b, c, d, e, f\}$ cuya red social se ilustra en la Figura 15. Finalmente, suponga que el conjunto de habilidades de cada candidato es como sigue: $X_a = X_d = \{s_2\}$, $X_b = X_c = \{s_1\}$, $X_e = \{s_1, s_2\}$, $X_f = \{s_2, s_3\}$. Por lo tanto, cada uno de los siguientes subconjuntos de $\mathcal{X}$ son soluciones factibles: $\mathcal{I}_1 = \{a, e, f\}$, $\mathcal{I}_2 = \{a, d, e, f\}$, $\mathcal{I}_3 = \{a, b, d, f\}$. Note que el subgrafo inducido por las soluciones no necesariamente están conectados, como es el caso de $G[\mathcal{I}_1]$.

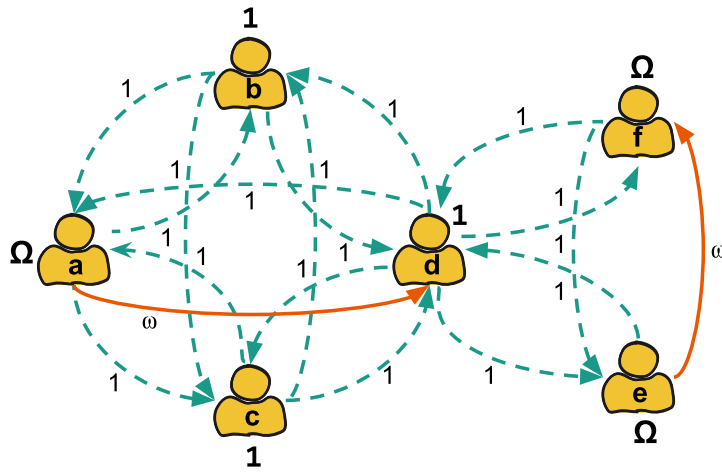


Figura 15. Ejemplo de un grafo de una red social.

La aptitud de una solución $\mathcal{I}$ está dada por el par $(D_{\rightarrow}(\mathcal{I}), Z(\mathcal{I}))$. Para ilustrar esto, se recicla el ejemplo planteado en el párrafo anterior. En la Figura 15, el nivel de experiencia de un candidato i con respecto a la tarea (i.e., $\sum_{(a, k_a) \in T} w_{ia}$) se encuentra representado con un valor en negritas junto a su nodo correspondiente. Del mismo

³No confundir la palabra individuo en el contexto evolutivo con la palabra persona en el contexto de miembro de equipo. Se usa la palabra persona o candidato para referirse a integrantes o potenciales integrantes en un equipo, y por otro lado se usa la palabra individuo para referirse a una solución del problema.

modo, el peso de colaboración de una arista dirigida se encuentra determinado por el valor junto a ésta. Como parte de este ejemplo, considere además que Ω y ω son un par de constantes arbitrariamente grandes. Así, las aptitudes de las soluciones $\mathcal{I}_1$, $\mathcal{I}_2$ e $\mathcal{I}_3$ son $(\frac{\omega+1}{6}, \Omega)$, $(\frac{\omega+2}{6}, \frac{3\Omega+1}{4})$, $(\frac{\omega+7}{12}, \frac{\Omega+1}{2})$, respectivamente. Note que $\mathcal{I}_1$ e $\mathcal{I}_2$ son soluciones no-dominadas entre ellas, mientras que $\mathcal{I}_3$ es dominada por ambas.

5.3.3.2. Inicialización de la población

Los algoritmos genéticos requieren una población inicial que sirva como semilla del proceso evolutivo. Generalmente, una rutina que inicialice dicha población debe generar cada individuo de forma aleatoria (Eiben y Smith, 2003). Mejor aún si los individuos generados también pertenecen al espacio de soluciones factibles.

El Algoritmo 1 muestra el pseudocódigo utilizado para generar N individuos factibles de forma aleatoria. La función $\text{ReservoirSampling}(S, k)$, en la línea 5, devuelve un subconjunto S' de tamaño k muestreado aleatoriamente de un conjunto S . Así, en las líneas 4 a 6, se asegura que cada individuo generado cumple los requisitos de tener al menos k_α miembros en el equipo por cada habilidad α requerida en la tarea T .

Algoritmo 1: Inicialización

input : un conjunto $\mathcal{X}$ de n candidatos, una tarea $\mathcal{T}$ de m habilidades diferentes, y un número entero N
output: un conjunto P de $|P| = N$ individuos factibles.

```

1  $P \leftarrow \emptyset$ ;
2 while  $|P| < N$  do
3    $\mathcal{I} \leftarrow \emptyset$ ;
4   for  $a_j \in A(T)$  do
5      $\mathcal{I} \leftarrow \mathcal{I} \cup \text{ReservoirSampling}(S(a_j), k_{a_j})$ ;
6   end
7    $P \leftarrow P \cup \{\mathcal{I}\}$ ;
8 end
9 return  $P$ ;

```

5.3.3.3. Operadores genéticos

En esta subsección se presentan dos operadores de variación: uno de recombinación y otro de mutación. En ambos operadores se garantiza que los individuos sean soluciones factibles.

El Algoritmo 2 muestra el pseudocódigo del operador de recombinación utilizado. Básicamente, por cada tupla $(a, k_a) \in T$, con una probabilidad $p = 0.5$, el individuo O_1 recibe todos aquellos miembros de P_1 que cuenten con la habilidad a (e.i., $S_{P_1}(a)$), lo mismo pasa con O_2 y los miembros correspondientes de P_2 ; en contraparte, con una probabilidad $q = (1 - p)$, O_1 recibe los miembros de P_2 que cuenten con a , lo mismo es para O_2 con respecto a P_1 . Tome en cuenta que este método tiene una probabilidad de que los individuos de la descendencia O_1 y O_2 sean idénticos a sus padres P_1 y P_2 . No obstante, la probabilidad de que esto ocurra disminuye exponencialmente con respecto al número distinto de habilidades requeridas—estrictamente, la probabilidad de clonación es de $\frac{1}{2^{m-1}}$ con $m = |T|$.

Algoritmo 2: Recombinación

```

input : dos individuos,  $P_1$  y  $P_2$ , y una tarea  $T$ 
output: dos individuos,  $O_1$  y  $O_2$ 

1  $O_1 \leftarrow \emptyset$ ;
2  $O_2 \leftarrow \emptyset$ ;
3 for  $a_j \in A(T)$  do
4   if CoinFlip() = Heads then
5      $O_1 \leftarrow O_1 \cup S_{P_1}(a_j)$ ;
6      $O_2 \leftarrow O_2 \cup S_{P_2}(a_j)$ ;
7   else
8      $O_1 \leftarrow O_1 \cup S_{P_2}(a_j)$ ;
9      $O_2 \leftarrow O_2 \cup S_{P_1}(a_j)$ ;
10  end
11 end
12 return  $O_1, O_2$ ;

```

El Algoritmo 3 muestra el pseudocódigo con el operador de mutación que se utiliza. Su funcionamiento es el siguiente. Primero, se crea un nuevo individuo I' a partir de un solo miembro elegido de forma aleatoria del equipo de entrada I . Después, si hay un déficit en cuanto al número de personas requeridas por cada habilidad, se toma un miembro aleatorio de entre aquellos miembros de I que cuenten con alguna de las habilidades faltantes, y se agrega a I' . Esta última operación se repite hasta que no exista déficit de habilidades en I' con respecto a la tarea T .

Note que ambos operadores garantizan la factibilidad de los individuos resultantes.

Algoritmo 3: Mutación

input : un individuo I y una tarea $\mathcal{T}$
output: un individuo I'

```

1  $\mathcal{I}' \leftarrow \mathcal{I} \setminus \text{ReservoirSampling}(\mathcal{I}, 1);$ 
2  $D \leftarrow \text{SkillDeficit}(\mathcal{I}', T);$ 
3 while  $|D| > 0$  do
4    $a_j \leftarrow \text{random skill from } D;$ 
5    $\mathcal{I}' \leftarrow \mathcal{I}' \cup \text{ReservoirSampling}(S(a_j), 1);$ 
6    $D \leftarrow \text{SkillDeficit}(\mathcal{I}', T);$ 
7 end
8 return  $I'$ ;

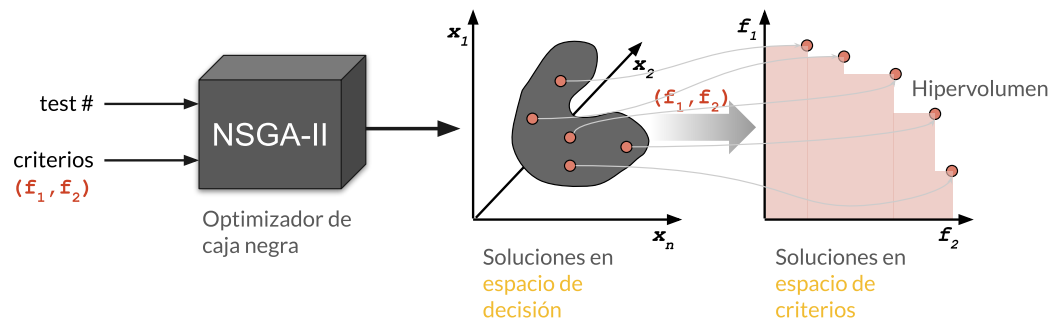
```

5.3.4. Diseño de experimento

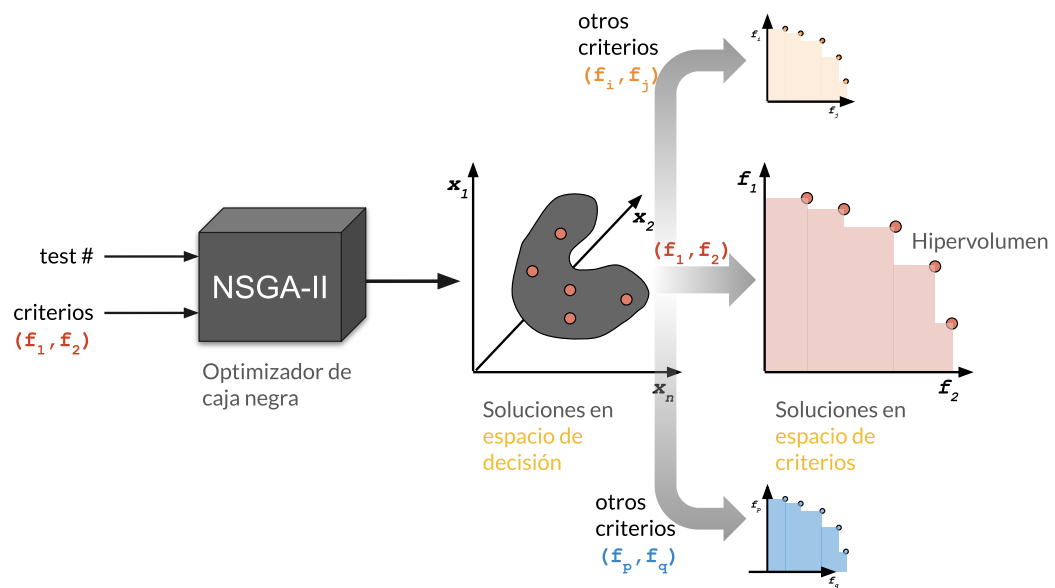
Cuando se habla de evaluar conjuntos de soluciones no dominadas, el hipervolumen es probablemente uno de los indicadores más adecuados. Particularmente, el hipervolumen es uno de los pocos indicadores conocidos que refleja, con un sólo valor, la cercanía de un conjunto de soluciones con el frente óptimo de Pareto y su distribución en éste (Li y Yao, 2019).

Un esquema de optimización y evaluación típico para un problema multiobjetivo está representado en la Figura 16a. Primero, un optimizador multi-objetivo—el NSGA-II en este caso—recibe como entrada un caso de prueba y un conjunto de funciones objetivo a optimizar. El optimizador devuelve como salida un conjunto de soluciones no dominadas que aproximan el frente Pareto con respecto a los objetivos optimizados. Finalmente, se evalúa el hipervolumen resultante de las soluciones no dominadas proyectadas en el espacio de criterios.

Sin embargo, la pregunta importante en este capítulo es la siguiente: ¿es posible optimizar para ciertos objetivos y al mismo tiempo obtener buenos resultados en otros objetivos? Un primer paso en esa dirección sería optimizar una tupla de objetivos dada, y posteriormente evaluar el conjunto resultante con cada tupla de objetivos considerados de interés. La Figura 16b muestra un esquema de la metodología planteada. En la imagen se optimiza para los objetivos (f_{x_1}, f_{y_1}) y posteriormente se evalúa el conjunto de soluciones resultante calculando el hipervolumen en cada espacio de criterios (f_{x_1}, f_{y_1}) , (f_{x_2}, f_{y_2}) , y (f_{x_3}, f_{y_3}) . Este proceso se repite optimizando para (f_{x_2}, f_{y_2}) y evaluando para cada tupla de criterios, y repitiendo todo una vez más con (f_{x_3}, f_{y_3}) .



(a) Metodología típica para evaluar conjuntos no dominados en optimización multiobjetivo.



(b) Propuesta de metodología para evaluar conjuntos no dominados en optimización multiobjetivo.

Figura 16. Ejemplo de metodologías para evaluar conjuntos de soluciones no dominadas.

Finalmente, se obtiene una matriz comparativa de hipervolúmenes: optimizado para (f_{x_i}, f_{y_i}) y evaluado para (f_{x_j}, f_{y_j}) . Se usará el término comparación cruzada para referir a este esquema.

5.3.4.1. Detalles adicionales

En el párrafo anterior se describe el experimento a un nivel muy general. A continuación se cubren los detalles adicionales.

Al conjunto no dominado resultante de optimizar para una tupla de criterios (f_{x_i}, f_{y_i}) se le denomina $ND_{(f_{x_i}, f_{y_i})}$. Por cada tupla de criterios, se realizan quince ejecuciones independientes del NSGA-II. De manera que para (f_{x_i}, f_{y_i}) , se tienen quince conjuntos no dominados $ND_{(f_{x_i}, f_{y_i})}^{(1)}, ND_{(f_{x_i}, f_{y_i})}^{(2)}, \dots, ND_{(f_{x_i}, f_{y_i})}^{(15)}$. A partir de estos quince conjuntos no dominados, se producen dos conjuntos: un conjunto con el consolidado de soluciones no dominadas llamado $CND_{(f_{x_i}, f_{y_i})}$ y un conjunto con la unión de los conjuntos de soluciones no dominadas llamado $UND_{(f_{x_i}, f_{y_i})}$. La Figura 17 muestra un esquema de esto. Posteriormente, el bloque anterior se repite quince veces como se muestra en la Figura 18. De esta manera, para la tupla de criterios (f_{x_i}, f_{y_i}) , se tienen quince conjuntos consolidados $CND_{(f_{x_i}, f_{y_i})}^{(1)}, CND_{(f_{x_i}, f_{y_i})}^{(2)}, \dots, CND_{(f_{x_i}, f_{y_i})}^{(15)}$, y quince conjuntos de uniones $UND_{(f_{x_i}, f_{y_i})}^{(1)}, UND_{(f_{x_i}, f_{y_i})}^{(2)}, \dots, UND_{(f_{x_i}, f_{y_i})}^{(15)}$.

La experimentación se divide en dos esquemas llamados Experimento **A** y Experimento **B**. Para el Experimento **A** se emplean únicamente los conjuntos consolidados de soluciones no dominadas CND , y para el Experimento **B** se emplean únicamente los conjuntos con la unión de las soluciones no dominadas UND . Para fines prácticos, la descripción se centra sobre los conjuntos CND , aunque la idea es la misma con los conjuntos UND .

Se denomina $Hv_{(f_{x_i}, f_{y_i}) \vdash (f_{x_j}, f_{y_j})}$ al hipervolumen resultante de optimizar para la tupla de criterios (f_{x_i}, f_{y_i}) y evaluar para la tupla de criterios (f_{x_j}, f_{y_j}) . Así, por cada conjunto consolidado $CND_{(f_{x_i}, f_{y_i})}^{(1 \dots 15)}$, por cada (f_{x_j}, f_{y_j}) , se tienen quince hipervolúmenes correspondientes $Hv_{(f_{x_i}, f_{y_i}) \vdash (f_{x_j}, f_{y_j})}^{(1)}, Hv_{(f_{x_i}, f_{y_i}) \vdash (f_{x_j}, f_{y_j})}^{(2)}, \dots, Hv_{(f_{x_i}, f_{y_i}) \vdash (f_{x_j}, f_{y_j})}^{(15)}$. Estos hipervolúmenes se promedian quedando como $\bar{H}v_{(f_{x_i}, f_{y_i}) \vdash (f_{x_j}, f_{y_j})}$, por cada (f_{x_j}, f_{y_j}) . De esta manera, para un caso de prueba cualquiera, se pueden comparar los hipervolúmenes $\bar{H}v_{(f_{x_1}, f_{y_1}) \vdash (f_{x_j}, f_{y_j})}, \bar{H}v_{(f_{x_2}, f_{y_2}) \vdash (f_{x_j}, f_{y_j})}$, hasta $\bar{H}v_{(f_{x_k}, f_{y_k}) \vdash (f_{x_j}, f_{y_j})}$, y determinar cuál de las tuplas

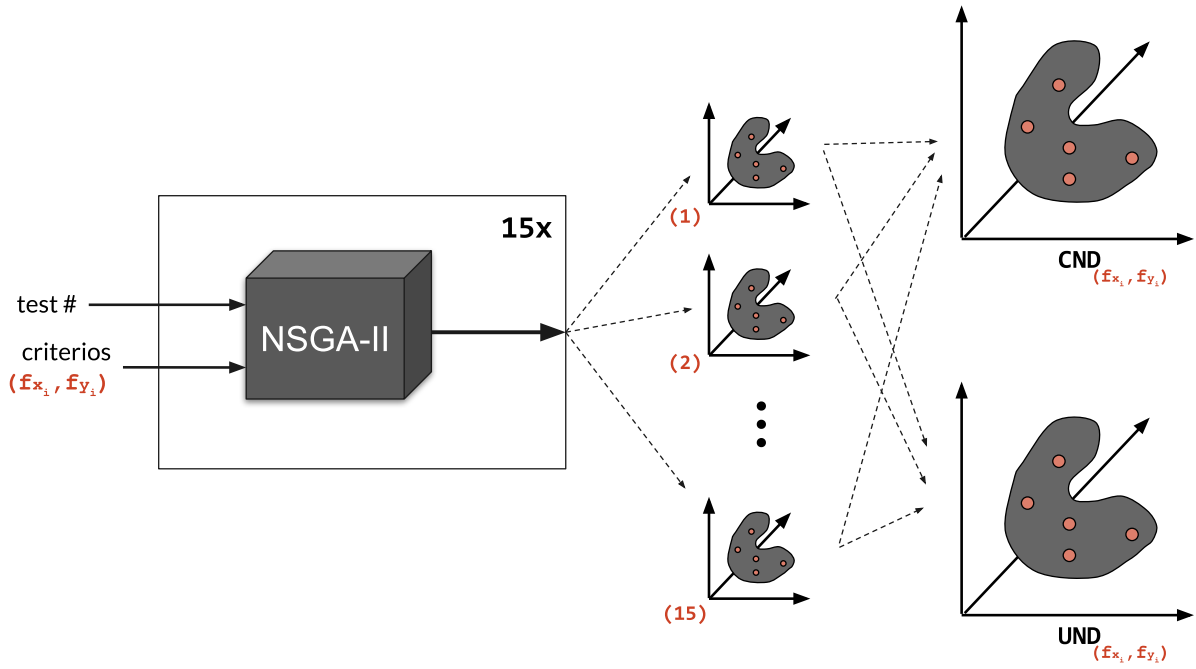


Figura 17. Esquema parcial del experimento. Se realizan quince ejecuciones independientes del NSGA-II, resultando en quince frentes no dominados. Estos frentes se combinan de manera distinta para formar los conjuntos de soluciones *CND* y *UND*.

$(f_{x_1}, f_{y_1}), (f_{x_2}, f_{y_2}), \dots$, o (f_{x_k}, f_{y_k}) es la más efectiva con respecto a la tupla (f_{x_j}, f_{y_j}) .

Posteriormente, por cada tupla (f_{x_j}, f_{y_j}) , se calcula el ranquin fraccional (también conocido como “ranquin 1,2.5,2.5,4”) de la lista de hipervolumenes $\bar{H}v_{(f_{x_1}, f_{y_1}) \vdash (f_{x_j}, f_{y_j})}, \bar{H}v_{(f_{x_2}, f_{y_2}) \vdash (f_{x_j}, f_{y_j})}, \dots, \bar{H}v_{(f_{x_k}, f_{y_k}) \vdash (f_{x_j}, f_{y_j})}$. De manera que, por cada caso de prueba q , se denomina como $R_{(f_{x_i}, f_{y_i}) \vdash (f_{x_j}, f_{y_j})}^{(q)}$ al ranquin del hipervolumen $\bar{H}v_{(f_{x_i}, f_{y_i}) \vdash (f_{x_j}, f_{y_j})}$ del caso de prueba correspondiente.

Finalmente, se denomina como $\bar{R}_{(f_{x_i}, f_{y_i}) \vdash (f_{x_j}, f_{y_j})}$ al ranquin promedio de los ránquines $R_{(f_{x_i}, f_{y_i}) \vdash (f_{x_j}, f_{y_j})}^{(1)}, R_{(f_{x_i}, f_{y_i}) \vdash (f_{x_j}, f_{y_j})}^{(2)}, \dots, R_{(f_{x_i}, f_{y_i}) \vdash (f_{x_j}, f_{y_j})}^{(30)}$ correspondientes a cada caso de prueba, con respecto a cada tupla de criterios de optimización (f_{x_i}, f_{y_i}) y a cada tupla de criterios de evaluación (f_{x_j}, f_{y_j}) . En conjunto, esto da una matriz de promedios de los ránquines de desempeño de calidad en cuanto a hipervolumen como se representa en la Tabla 8. Note que la diagonal de la tabla implica optimizar para una cierta tupla de criterios y evaluar para esa misma tupla de criterios.

Como se mencionó antes, este procedimiento se realiza para ambos experimentos **A** y **B**. Uno considerando los conjuntos *CND* y el otro considerando los conjuntos *UND*.

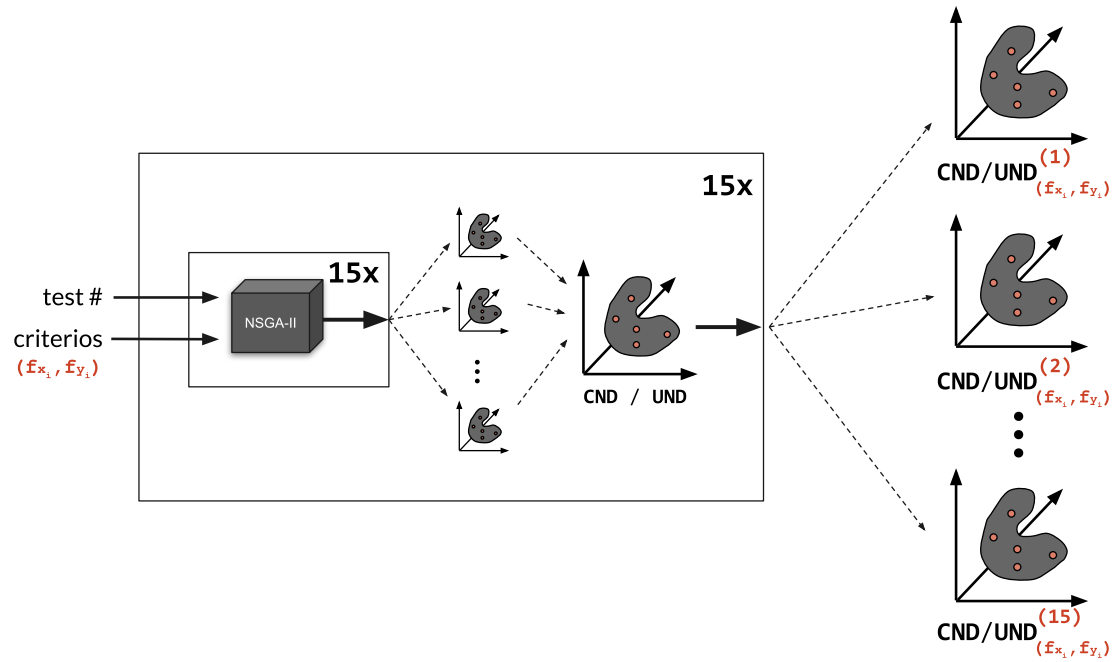


Figura 18. Esquema parcial del experimento. Se repite el bloque de la Figura 17 quince veces, resultando en quince CND y quince UND.

Tabla 8. Tabla de ejemplo con los ránquines promedio de los hipervolumenes generados al optimizar para una tupla de criterios y evaluar para otra tupla de criterios, de los 30 casos de prueba.

		optimizado			
		(f_{x_1}, f_{y_1})	(f_{x_2}, f_{y_2})	$\dots$	(f_{x_k}, f_{y_k})
evaluado	(f_{x_1}, f_{y_1})	$\bar{R}_{(f_{x_1}, f_{y_1}) \vdash (f_{x_1}, f_{y_1})}$	$\bar{R}_{(f_{x_2}, f_{y_2}) \vdash (f_{x_1}, f_{y_1})}$	$\dots$	$\bar{R}_{(f_{x_k}, f_{y_k}) \vdash (f_{x_1}, f_{y_1})}$
	(f_{x_2}, f_{y_2})	$\bar{R}_{(f_{x_1}, f_{y_1}) \vdash (f_{x_2}, f_{y_2})}$	$\bar{R}_{(f_{x_2}, f_{y_2}) \vdash (f_{x_2}, f_{y_2})}$	$\dots$	$\bar{R}_{(f_{x_k}, f_{y_k}) \vdash (f_{x_2}, f_{y_2})}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	(f_{x_k}, f_{y_k})	$\bar{R}_{(f_{x_1}, f_{y_1}) \vdash (f_{x_k}, f_{y_k})}$	$\bar{R}_{(f_{x_2}, f_{y_2}) \vdash (f_{x_k}, f_{y_k})}$	$\dots$	$\bar{R}_{(f_{x_k}, f_{y_k}) \vdash (f_{x_k}, f_{y_k})}$

5.4. Análisis y resultados

Para este estudio se consideró el siguiente conjunto de tuplas de criterios $\mathcal{C} = \{Cc-R, Cc-R_J, Cc-MST, Cc-MST_J, Cc-\Sigma d, Cc-\Sigma d_J, D, D_J, sD, sD_J, D_{\rightarrow}, (D, D_{\rightarrow})\} \times \{Z\}$. El primer conjunto de criterios representa los criterios de comunicación como se muestra en la Tabla 9 (introducidos en las subsecciones 5.1.1.1 y 5.1.1.2), mientras que $Z(\cdot)$ representa el criterio de habilidad técnica (introducido en la subsección 5.1.1.4).

Tabla 9. Tabla de criterios a utilizar según su clasificación y su modelo de interacción.

Interacción	Distancia	Densidad
Simple	$Cc-R(\cdot)$	$D_J(\cdot)$
	$Cc-MST(\cdot)$	$sD_J(\cdot)$
	$Cc-\Sigma d(\cdot)$	
Jaccard	$Cc-R_J(\cdot)$	$D_J(\cdot)$
	$Cc-MST_J(\cdot)$	$sD_J(\cdot)$
	$Cc-\Sigma d_J(\cdot)$	
Propuesta	—	$D_{\rightarrow}(\cdot)$

Por cada caso de prueba, se realizaron 225 (15×15) ejecuciones independientes del NSGA-II, optimizando por cada tupla en $\mathcal{C}$ (i.e., $30 \times 225 \times 12 = 81 \times 10^3$ ejecuciones totales). Los parámetros de ejecución del NSGA-II se muestran en la Tabla 10. Estos parámetros fueron retomados de una experimentación anterior en (Juárez y Brizuela, 2018).

Tabla 10. Tabla de parámetros de ejecución usados para el NSGA-II.

Parametro	Valor
Tamaño de la población	100
Tamaño del torneo	2
Número de generaciones	200
Probabilidad de recombinación	0.95
Probabilidad de mutación	0.08

5.4.1. Experimento A: considerando frente consolidado

En la Tabla 11 se muestra un resumen de los resultados del experimento que considera el frente consolidado de soluciones no dominadas como se describe en la sección 5.3.4.1. Las cifras en la tabla representan el ranquin promedio de optimizar para cierta tupla y evaluar para cierta otra tupla en los 30 casos de prueba.

Los resultados muestran principalmente que optimizar para un objetivo basado en

densidad junto con el objetivo de habilidad técnica, da mejores resultados, en promedio, que optimizar para un objetivo basado en distancia junto con el mismo objetivo de habilidad técnica.

Esto es más notorio tomando como referencia la diagonal de la tabla. Note que la diagonal de la tabla implica optimizar para una tupla de objetivos y evaluar para esa misma tupla de objetivos. Intuitivamente, uno esperaría ver ránquines promedio cercanos al primer o segundo puesto. Esto es así para los objetivos basados en densidad, pero no es así para los objetivos basados en distancia. Por otro lado, en promedio, ningún objetivo basado en distancia estuvo dentro de los primeros seis lugares (el 50 % superior). En cambio, los objetivos basados en densidad, con excepción de la subdensidades, estuvieron consistentemente dentro de los primeros seis lugares de los ránquines promedio.

Tabla 11. Ránquines promedio de los hipervolumenes generados al optimizar para la tupla $(Z, \cdot)$ de la horizontal y evaluar para $(Z, \cdot)$ de la vertical, de los 30 casos de prueba, para el Experimento **A** considerando el frente consolidado. En verde están aquellos ránquines que encabezan la primera mitad de su renglón y en rojo aquellos ránquines en la peor mitad.

	Cc-R _j	Cc-R	Cc-MST _j	Cc-MST	Cc-Σd _j	Cc-Σd	D _j	D	sD _j	sD	D _→	(D,D _→)
Cc-R _j	8.32	7.75	7.63	8.43	7.92	8.38	4.78	4.12	7.07	6.20	4.03	3.37
Cc-R	8.38	7.45	7.77	8.53	8.32	8.45	4.98	3.95	7.13	6.13	3.63	3.27
Cc-MST _j	8.62	8.12	7.37	8.53	7.88	8.55	5.08	3.98	6.80	5.90	3.93	3.23
Cc-MST	8.42	7.62	7.73	8.53	8.18	8.42	5.05	3.95	7.07	6.10	3.57	3.37
Cc-Σd _j	9.28	7.82	7.33	8.50	7.88	8.25	5.05	4.38	6.53	6.10	3.83	3.03
Cc-Σd	9.02	7.08	7.57	8.37	8.42	7.92	5.45	4.35	6.83	6.10	3.70	3.20
D _j	9.10	9.97	9.23	9.97	9.43	8.80	1.23	4.87	3.70	5.97	3.23	2.50
D	9.00	9.40	8.97	9.57	8.97	8.90	3.87	1.80	6.97	4.13	4.00	2.43
sD _j	9.17	9.93	9.33	9.60	9.63	9.10	1.30	5.50	1.97	5.40	3.87	3.20
sD	9.40	9.30	9.00	9.57	9.13	9.27	4.00	2.37	6.43	2.30	4.57	2.67

5.4.2. Experimento B: sin considerar el frente consolidado

En la Tabla 12 se muestra un resumen de los resultados del experimento sin considerar el frente consolidado de soluciones no dominadas como se describe en la sección 5.3.4.1. Al igual que en la tabla anterior, las cifras en esta tabla representan el ranquin promedio de optimizar para cierta tupla y evaluar para cierta otra tupla en los 30 casos de prueba.

Los resultados muestran nuevamente una tendencia que favorece, en promedio, a los objetivos basados en densidad sobre los objetivos basados en distancia, cuando se optimizan en conjunto con el objetivo de habilidad técnica.

Esta vez se puede notar un aumento generalizado en el promedio de los ránquines en la diagonal correspondiente a los objetivos basados en densidad, pero ninguno con un ranquin peor al quinto puesto. En cuanto a los ránquines promedio de la diagonal de los objetivos basados en distancia mejoraron mientras que otros empeoraron, pero en general se mantuvieron relativamente en la misma zona—en la parte baja del ranquin.

En cambio, algunos de los ránquines promedio de la diagonal de los objetivos basados en distancia mejoraron mientras que otros empeoraron, pero en general se mantuvieron relativamente iguales—en la parte baja del ranquin.

Tabla 12. Ránquines promedio de los hipervolumenes generados al optimizar para la tupla $(Z, \cdot)$ de la horizontal y evaluar para $(Z, \cdot)$ de la vertical, de los 30 casos de prueba, para el Experimento **B** sin considerar el frente consolidado. En verde están aquellos ránquines que encabezan la primera mitad de su renglón y en rojo aquellos ránquines en la peor mitad.

	Cc-R _j	Cc-R	Cc-MST _j	Cc-MST	Cc-Σd _j	Cc-Σd	D _j	D	sD _j	sD	D _→	(D,D _→)
Cc-R _j	8.20	8.28	7.48	8.53	7.95	8.43	5.37	3.83	6.15	4.78	4.33	4.65
Cc-R	8.40	8.15	7.55	8.73	7.82	8.47	5.57	3.90	6.08	4.75	4.43	4.15
Cc-MST _j	8.33	8.28	7.43	8.60	7.73	8.70	5.77	3.93	5.82	4.78	4.47	4.15
Cc-MST	8.43	8.18	7.47	8.73	7.87	8.47	5.70	3.77	5.92	4.78	4.50	4.18
Cc-Σd _j	8.67	8.18	7.63	8.77	7.77	9.03	5.20	3.57	6.18	4.75	4.13	4.12
Cc-Σd	8.47	8.12	7.95	8.93	7.92	8.93	5.23	3.40	6.18	4.68	4.07	4.12
D _j	8.57	8.00	8.43	8.83	8.17	8.07	4.87	4.10	4.47	4.03	5.50	4.97
D	9.17	8.63	8.73	9.57	8.37	8.87	3.90	2.63	6.53	3.33	4.93	3.33
sD _j	8.63	8.07	8.47	8.73	8.13	7.97	4.77	4.30	4.53	4.03	5.27	5.10
sD	9.67	9.33	9.10	9.73	8.70	8.97	3.33	2.67	5.93	2.77	4.73	3.07

5.4.3. Tiempo de ejecución

Además del desempeño en cuanto a calidad de soluciones, es importante analizar el desempeño en cuanto al tiempo de ejecución.

En la Figura 19 se muestra el tiempo (ms) promedio de ejecución de las 225 corridas del NSGA-II, por objetivo distinto, por caso de prueba. En color rojo, se muestran los tiempos correspondientes a los objetivos basados en distancia (Jaccard y simple). En color cian, se muestran los objetivos basados en densidad (Jaccard y simple), mientras que en magenta, se muestran el objetivo propuesto y una combinación entre el objetivo propuesto y la densidad simple.

A simple vista, se puede observar que el proceso evolutivo es más lento cuando se optimizan objetivos como el costo del diámetro (Cc-R) y el costo de la suma de distancias (Cc-Σd). Esto tiene sentido puesto que para evaluar Cc-R y Cc-Σd se necesita

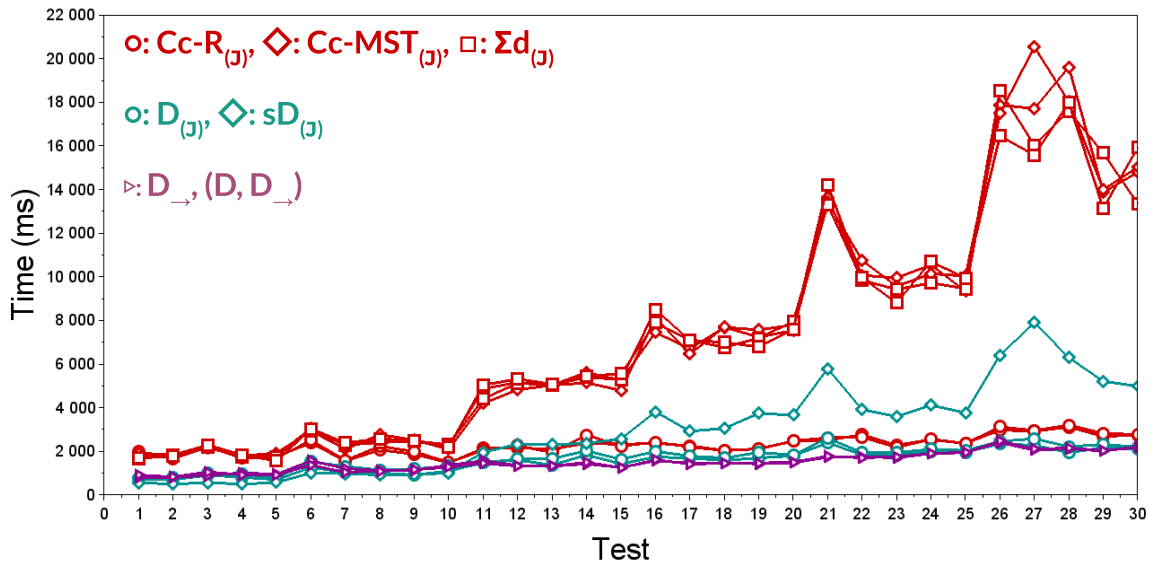


Figura 19. Tiempos de ejecución promedio del NSGA-II por cada caso de prueba, por cada criterio de optimización empleado.

resolver un problema conocido como *all pairs shortest paths* del subgrafo inducido por el equipo, que tiene un tiempo de ejecución de $O(|\mathcal{X}'|^3)$ u.t. donde $\mathcal{X}'$ es el conjunto de nodos que forman el equipo. En cuanto al otro objetivo basado en distancia, el costo del árbol de expansión mínimo (Cc-MST), se ve que no es tan tardado como sus símiles, pero ligeramente más lento que algunos objetivos basados en densidad. Recuerde que el algoritmo para calcular un MST requiere $O(|\mathcal{X}'| \log |E_{G[\mathcal{X}']}|)$ u.t.

En cuanto al tiempo de ejecución correspondiente a los objetivos basados en densidad, ocurre algo interesante. Tanto la densidad (D) y como la subdensidad (sD) se calculan en $O(|E_{G[\mathcal{X}']}|)$ u.t. Sin embargo, a partir del caso de prueba #11 se observa una diferencia notoria entre estos dos objetivos durante el proceso evolutivo, donde sD eleva su tiempo de ejecución con respecto a D. Esto se debe a un efecto producido por la definición de cada objetivo, en donde el denominador de D ejerce más presión a un tamaño de equipo más reducido que con respecto a sD. En otras palabras, en promedio, el objetivo D tiene una tendencia a producir soluciones factibles de equipos con menos número de integrantes que los equipos producidos por sD. Este efecto también se reportó anteriormente en (Juárez y Brizuela, 2018).

Finalmente, el tiempo de ejecución del objetivo propuesto, la densidad dirigida

($D_{\rightarrow}$), se muestra bastante similar al tiempo de ejecución de la densidad simple. La complejidad computacional de calcular $D_{\rightarrow}$ es igual a las otras densidades, es decir, $O(|E_{G[\mathcal{X}']}|)$ u.t. Y al igual que la densidad simple D , la densidad propuesta $D_{\rightarrow}$ también favorece la formación de equipos de tamaño eficiente.

5.5. Discusión

Los resultados anteriores favorecen a la densidad propuesta ($D_{\rightarrow}$) y la densidad simple (D) en cuanto a calidad de soluciones, tiempo de ejecución y tamaño del equipo. Sin embargo, no hay que olvidar que estos resultados se obtienen de un proceso de optimización multi-objetivo, donde los criterios de densidad o distancia se optimizan conjuntamente con un criterio de nivel de experiencia técnica del equipo. Además, es importante recalcar que los resultados fueron producidos mediante un método de optimización estocástica, el NSGA-II. Esto último conlleva un par de aclaraciones particulares: (1) el NSGA-II, como se dijo, no es un método exacto sino que existe una cierta variabilidad en cuanto a los resultados que produce en cada ejecución, y (2) incluso la no sintonización de los parámetros pudiese favorecer al esquema ganador. Una forma de tratar con el primer punto es repetir los experimentos un cierto número de veces y analizar los resultados con base en su promedio (similar a lo que se propone en la subsección 5.3.4.1). Para el segundo punto se requiere probar con otro método de optimización, idealmente con un método exacto aunque esto puede suponer otro tipo de problemas como tiempos de ejecución considerablemente prolongados o la dificultad para generar soluciones en ciertas regiones del frente óptimo de Pareto cuando este presente una forma no convexa.

Tampoco se descarta que algunos casos de prueba pudieran favorecer más un esquema de criterios que otro. Por lo tanto, una correlación de las definiciones de los casos de prueba pudiera explicar que un criterio presente mejores resultados que otro. Sin embargo, esto requeriría de un análisis mucho más profundo de los casos de prueba. Para obtener mayor evidencia de esto, se puede conducir un experimento similar usando otros conjuntos de datos como entrada, como el grafo social de GitHub y el grafo social de IMDb. Por supuesto, esto también implica diseñar casos de prueba específicos con las habilidades de cada dominio.

No obstante de las limitaciones anteriores, existen dos explicaciones adicionales que suman al caso en favor de las densidades. La primer razón tiene que ver con el siguiente principio: conforme se aumenta la densidad de colaboración en un equipo, potencialmente se disminuye su distancia de comunicación (como se muestra en la Figura 20).

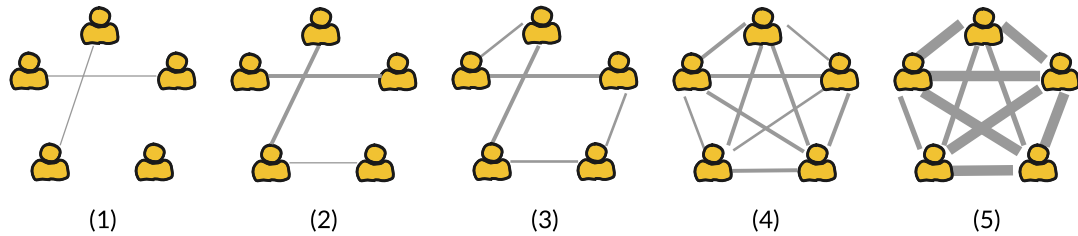


Figura 20. Ejemplo de conectividad y densidad en un grafo. A mayor densidad, potencialmente mayor conectividad.

La segunda razón está asociada con lo que se comenta en la subsección 5.1.1.3, que para equipos (grafos) desconectados, no hay una definición concreta para los criterios basados en distancia ($Cc-R$, $Cc-MST$ y $Cc-\Sigma d$). Esto significa que los criterios basados en distancia, a diferencia de los criterios basados en densidad, no proveen suficiente información a los algoritmos de mejora iterativa sobre qué tan buena es una solución hasta contar con un equipo conectado. Esto pone en desventaja a los criterios basados en distancia ante algoritmos evolutivos (como el NSGA-II), que dependen de inicializar soluciones de manera aleatoria—que lo más probable es que resulte en equipos inicialmente desconectados.

Colectivamente, esta propuesta y sus resultados dan una pequeña idea de las características de un modelo de PFE más general. Particularmente, un modelo más general puede ser aquel con un enfoque de problema multiobjetivo que considere algún criterio de interacción, especialmente uno basado en densidad de grafo (o similar). Éste puede beneficiarse de implementar la propuesta de aristas no simétricas del capítulo anterior.

Capítulo 6. Conclusiones

En este trabajo de tesis se presentó una revisión de la literatura donde se describen las componentes de los problemas de formación de equipo, y se detalla cómo se obtienen los valores de entrada de los modelos. Como uno de los objetivos principales, se realizó una taxonomía de los problemas de formación de equipo donde se destaca la estructura de los modelos. También se propuso una medida no simétrica de interacción entre personas para representar la fuerza de su vínculo, y con ella se propuso clasificar los enlaces como fuertes o débiles. Este concepto permitió establecer un vínculo con el área de la sociometría, particularmente con la teoría de Granovetter (1973) conocida como la fuerza de los enlaces débiles. Asimismo, se propuso una medida de densidad ponderada basada en la propuesta de interacción no simétrica. Finalmente, se realizó un estudio comparativo entre criterios de optimización basados en densidad de colaboración y basados en distancia de comunicación, desde una perspectiva multiobjetivo en conjunto con un objetivo de nivel de experiencia del equipo.

6.1. Conclusiones generales

En conjunto, los estudios presentados en esta tesis dejan cuatro conclusiones principales:

- La taxonomía propuesta invita primeramente a reflexionar sobre los problemas de formación de equipo desde el punto de vista de su estructura, sin importar el tipo de equipo que se quiere formar y del contexto en el que reside. Ésta ofrece el marco faltante para estudiar los elementos esenciales de cada modelo, y así relacionarlos con otros problemas combinatorios bien conocidos. Básicamente, se identifican dos grandes familias: los PFE basados en asignación y los PFE basados en comunidad. Por un lado los PFE basados en asignación se relacionan con la estructura de los problemas de asignación, mientras que los PFE basados en comunidad se relacionan con la estructura de los problemas de cubrimiento de conjunto. Identificar este tipo de relaciones permite además tomar lo que se conoce previamente de un problema y aplicarlo a los PFEs.
- Un análisis de la literatura de investigación de equipos sugiere que los patrones de interacción tienen un rol importante en su eficacia. Por lo tanto la componente

que representa la interacción de equipo no debe faltar en las formulaciones de los PFEs. En los modelos basados en asignación, la componente que modela la interacción entre pares de candidatos asemeja a un caso particular del problema conocido como QAP si se considera un flujo unitario. En los modelos basados en comunidad, la componente que modela la interacción del equipo es nativa de esta familia de problemas, y se representa con alguna medida de cercanía entre los nodos de un grafo como puede ser distancia o densidad.

- Los modelos de interacción no simétrica entre pares de personas tiene ventajas sobre los modelos simétricos. Los modelos de interacción no simétrica revelan patrones y estructuras de interacción que no son visibles para los modelos de interacción simétrica. Estos patrones y estructuras pueden ser útiles para los análisis y la toma de decisión posteriores a la optimización, o bien, pueden ser útiles para diseñar algoritmos especializados que exploten patrones particulares de los grafos sociales.
- Se encuentra evidencia a favor de los objetivos basados en densidad de colaboración sobre los objetivos basados en distancia de comunicación. Considerando una formulación multi-objetivo del Problema de Formación de Equipo en donde se optimiza para dos objetivos: uno de habilidad técnica y otro objetivo basado en interacción (i.e., distancia o densidad). En general, en un análisis cruzado, las formulaciones con objetivos basados en densidad mostraron un mejor desempeño que aquellas con objetivos basados en distancia.

6.2. Limitaciones

Las limitaciones de este trabajo de tesis son muchas y muy amplias. Estudiar el Problema de Formación de Equipos no sólo es encontrar la forma de resolver un problema de optimización de la manera más eficiente posible. Una enorme parte del problema es definir el modelo que se debe optimizar. Como se menciona en el Capítulo 2, esto depende fuertemente del contexto en el que opera el equipo (remoto o en sitio, en escritorio o en campo), la naturaleza de la tarea que tiene que realizar (física o cognitiva) y el entorno social. Tratar de generalizar conocimiento a partir de esta situación es en sí una tarea complicada.

En general, los modelos propuestos en esta tesis así como los modelos de la literatura parten de la suposición de que éstos representan, cuando menos, alguna parte de la efectividad de equipo. Esto aún cuando las áreas especializadas como la psicología social o la psicología organizacional no tienen una definición concreta de efectividad de equipo. Más aún, incluso suponiendo que los modelos de optimización están alineados con la efectividad de equipo, éstos dependen de los valores de las variables de entrada que representan factores humanos como habilidades técnicas, comunicación interpersonal, o rasgos de personalidad. El problema aquí es que no se cuenta con herramientas que puedan medir estas variables de forma precisa o directa. Es decir, en muchos casos, no se tiene certeza de la calidad de los valores de entrada de los modelos de optimización. En esos casos, invariablemente de la cercanía del modelo con la efectividad de equipo, el equipo resultante seguramente no será el “mejor”.

Finalmente, aún en lo general, no hay que perder de vista que en la práctica, el Problema de Formación de Equipo probablemente no se visualiza como un modelo de optimización sino como un modelo predictivo. Y el paradigma de predicción requiere otro tipo de consideraciones adicionales como la estandarización y la calidad en los datos de entrada, el porcentaje de confianza en la predicción y el manejo del riesgo en la predicción.

En lo particular, existen limitaciones que son específicas de cada capítulo. En cuanto al Capítulo 4, Granovetter planteó la hipótesis de que la fuerza del vínculo es una combinación (probablemente lineal) de cuatro variables: tiempo, intensidad emocional, confianza mutua y servicios recíprocos (Granovetter, 1973). Nuestra propuesta, sin embargo, es mucho menos detallada que la fuerza del vínculo de Granovetter y no debe interpretarse como un sustituto de la misma. Simplemente se carece de las herramientas y la información necesarias para evaluar variables como la intensidad emocional o la confianza mutua. Otros estudios sobre la fuerza de los vínculos han intentado identificar las variables de Granovetter en el contexto de diferentes redes sociales, pero los resultados resultan ser específicos de un dominio o red social, o incluso poco prácticos de evaluar (Liberatore y Quijano-Sanchez, 2017). Alternativamente, nuestra propuesta estima la fuerza de los vínculos en el contexto de lo que se ha hecho en la literatura del PFE y la información disponible de los conjuntos de datos.

En cuanto al Capítulo 5, el estudio se limitó a uno de los tres conjuntos de datos que

se usaron en el Capítulo 4. Para usar los dos conjuntos de datos restantes se necesita definir cómo dotar cada candidato de un conjunto de habilidades correspondientes a los proyectos en los que participaron. De la mano con lo anterior está definir un conjunto de casos de prueba que también requieren definir ese universo de habilidades. Otra limitante importante está en el uso de un único método de optimización para el experimento, el NSGA-II. Aún cuando en la discusión del Capítulo 5 se describen aspectos adicionales que favorecen a los objetivos basados en densidad sobre los objetivos basados en distancia, no se sabe con certeza si el NSGA-II, por alguna razón, favoreció también a la densidad sobre la distancia.

6.3. Trabajo a futuro

Queda mucho trabajo por hacer, tanto que resultaría muy ingenuo tratar de describirlo en una sola sección. Se acota la descripción de trabajo a futuro a aquello directamente relacionado con las limitantes mencionadas en la sección anterior.

En cuanto al Capítulo 4, se sugiere diseñar un experimento para comparar nuestra propuesta de fuerza de enlace con otras propuestas de la literatura, aún cuando estas últimas sean específicas de la red social. Otra dirección de trabajo a futuro consiste en observar una red social, comunidad, o equipo, y la evolución de su(s) estructura(s) a lo largo del tiempo usando el modelo de interacción propuesto. Esto también tiene una aplicación en cuanto a la predicción de nuevos enlaces.

Para el estudio realizado en el Capítulo 5, se sugiere realizar las siguientes extensiones para generalizar los resultados. Primero, además del DBLP, experimentar con los conjuntos de datos del IMDb y GitHub como lo realizado en el Capítulo 4. Esto requiere además definir casos de prueba específicos para cada dataset. Segundo, probar con otra metaheurística además del NSGA-II. Algunas opciones pueden ser el MOEA/D (Zhang y Li, 2007) o alguna de sus variantes, o un enfoque de partículas como el MO-PSO (Coello y Lechuga, 2002).

Literatura citada

- Abdelsalam, H. M. (2009). Multi-objective team forming optimization for integrated product development projects. En: *Foundations of Computational Intelligence Volume 3*. Springer, pp. 461–478.
- Adams, J. A. et al. (2011). Coalition formation for task allocation: Theory and algorithms. *Autonomous Agents and Multi-Agent Systems*, **22**(2): 225–248.
- Ahmed, F., Deb, K., y Jindal, A. (2013). Multi-objective optimization and decision making approaches to cricket team selection. *Applied Soft Computing*, **13**(1): 402–414.
- Allen, T. J. et al. (1984). *Managing the flow of technology: Technology transfer and the dissemination of technological information within the R&D organization*, Vol. 1. The MIT Press.
- Anagnostopoulos, A., Becchetti, L., Castillo, C., Gionis, A., y Leonardi, S. (2010). Power in unity: forming teams in large-scale community systems. En: *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, pp. 599–608.
- Anagnostopoulos, A., Becchetti, L., Castillo, C., Gionis, A., y Leonardi, S. (2012). On-line team formation in social networks. En: *Proceedings of the 21st international conference on World Wide Web*. ACM, pp. 839–848.
- Arkin, E. M. y Hassin, R. (2000). Minimum-diameter covering problems. *Networks: An International Journal*, **36**(3): 147–155.
- Awal, G. K. y Bharadwaj, K. (2014). Team formation in social networks based on collective intelligence—an evolutionary approach. *Applied Intelligence*, **41**(2): 627–648.
- Bahargam, S., Golshan, B., Lappas, T., y Terzi, E. (2019). A team-formation algorithm for faultline minimization. *Expert Systems with Applications*, **119**: 441–455.
- Baykasoğlu, A., Dereli, T., y Daş, S. (2007). Project team selection using fuzzy optimization approach. *Cybernetics and Systems: An International Journal*, **38**(2): 155–185.
- Boon, B. H. y Sierksma, G. (2003). Team formation: Matching quality supply and quality demand. *European Journal of Operational Research*, **148**(2): 277–292.
- Bradley, B. H., Postlethwaite, B. E., Klotz, A. C., Hamdani, M. R., y Brown, K. G. (2012). Reaping the benefits of task conflict in teams: The critical role of team psychological safety climate. *Journal of Applied Psychology*, **97**(1): 151.
- Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, **25**(2): 163–177.
- Buckingham, M. y Clifton, D. O. (2001). *Now, discover your strengths*. Simon and Schuster.
- Burt, R. S. (2004). Structural holes and good ideas. *American journal of sociology*, **110**(2): 349–399.
- Campion, M. A., Medsker, G. J., y Higgs, A. C. (1993). Relations between work group characteristics and effectiveness: Implications for designing effective work groups. *Personnel psychology*, **46**(4): 823–847.

- Cattell, R. B., Eber, H. W., y Tatsuoka, M. M. (1970). *Handbook for the sixteen personality factor questionnaire (16 PF): In clinical, educational, industrial, and research psychology, for use with all forms of the test*. Institute for Personality and Ability Testing Champaign, IL.
- Chen, L., Ye, Y., Zheng, A., Xie, F., Zheng, Z., y Lyu, M. R. (2020). Incorporating geographical location for team formation in social coding sites. *World Wide Web*, **23**(1): 153–174.
- Chen, S.-J. y Lin, L. (2004). Modeling team member characteristics for the formation of a multifunctional team in concurrent engineering. *IEEE Transactions on Engineering Management*, **51**(2): 111–124.
- Coello, C. A. C., Lamont, G. B., y Van Veldhuizen, D. A. (2007). *Evolutionary algorithms for solving multi-objective problems*, Vol. 5. Springer.
- Coello, C. C. y Lechuga, M. S. (2002). Mopso: A proposal for multiple objective particle swarm optimization. En: *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No. 02TH8600)*. IEEE, Vol. 2, pp. 1051–1056.
- Cohen, S. G. y Bailey, D. E. (1997). What makes teams work: Group effectiveness research from the shop floor to the executive suite. *Journal of management*, **23**(3): 239–290.
- Das, A., Gómez-Zarà, D., y Contractor, N. (2020). Forming diverse teams based on members' social networks: A genetic algorithm approach. En: *International Conference on Complex Networks and Their Applications*. Springer, pp. 346–357.
- Deb, K., Pratap, A., Agarwal, S., y Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, **6**(2): 182–197.
- Demirović, E., Schwind, N., Okimoto, T., y Inoue, K. (2018). Recoverable team formation: Building teams resilient to change. En: *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, pp. 1362–1370.
- Dereli, T., Baykasoğlu, A., y Daş, G. S. (2007). Fuzzy quality-team formation for value added auditing: A case study. *Journal of Engineering and Technology Management*, **24**(4): 366–394.
- Dorn, C. y Dustdar, S. (2010). Composing near-optimal expert teams: a trade-off between skills and connectivity. En: *OTM Confederated International Conferences. On the Move to Meaningful Internet Systems*. Springer, pp. 472–489.
- Dubois, D. y Prade, H. (1978). Operations on fuzzy numbers. *International Journal of systems science*, **9**(6): 613–626.
- Duhigg, C. (2016). What google learned from its quest to build the perfect team. *The New York Times Magazine*, **26**: 2016.
- Durfee, E. H., Boerkoel Jr, J. C., y Sleight, J. (2014). Using hybrid scheduling for the semi-autonomous formation of expert teams. *Future Generation Computer Systems*, **31**: 200–212.

- Easley, D. y Kleinberg, J. (2010). *Networks, crowds, and markets*, Vol. 8. Cambridge university press Cambridge.
- Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative science quarterly*, **44**(2): 350–383.
- Eiben, A. E. y Smith, J. E. (2003). *Introduction to evolutionary computing*, Vol. 53. Springer.
- Farasat, A. y Nikolaev, A. G. (2016). Social structure optimization in team formation. *Computers & Operations Research*, **74**: 127–142.
- Fatima, S. y Wooldridge, M. (2019). Computing optimal coalition structures in polynomial time. *Autonomous Agents and Multi-Agent Systems*, **33**(1): 35–83.
- Feng, B., Jiang, Z.-Z., Fan, Z.-P., y Fu, N. (2010). A method for member selection of cross-functional teams using the individual and collaborative performances. *European Journal of Operational Research*, **203**(3): 652–661.
- Fitzpatrick, E. L. y Askin, R. G. (2005). Forming effective worker teams with multi-functional skill requirements. *Computers & Industrial Engineering*, **48**(3): 593–608.
- Foushee, H. C. y Helmreich, R. L. (1988). Group interaction and flight crew performance. En: *Human factors in aviation*. Elsevier, pp. 189–227.
- Friedkin, N. E. (1982). Information flow through strong and weak ties in intraorganizational social networks. *Social networks*, **3**(4): 273–285.
- Gajewar, A. y Das Sarma, A. (2012). Multi-skill collaborative teams based on densest subgraphs. En: *Proceedings of the 2012 SIAM International Conference on Data Mining*. SIAM, pp. 165–176.
- Garey, M. R. y Johnson, D. S. (1979). *Computers and intractability*, Vol. 174. freeman San Francisco.
- Garey, M. R. y Johnson, D. S. (1990). *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co. New York, NY, USA.
- Gibbons, A. (1985). *Algorithmic graph theory*. Cambridge university press.
- Gladstein, D. L. (1984). Groups in context: A model of task group effectiveness. *Administrative science quarterly*, pp. 499–517.
- Glover, F., Hultz, J., y Klingman, D. (1979). Improved computer-based planning techniques. part ii. *Interfaces*, **9**(4): 12–20.
- Gold, N. (2004). *Teamwork: Multi-disciplinary perspectives*. Springer.
- Golshan, B., Lappas, T., y Terzi, E. (2014). Profit-maximizing cluster hires. En: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 1196–1205.
- Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology*, **78**(6): 1360–1380.

- Gutiérrez, J. H., Astudillo, C. A., Ballesteros-Pérez, P., Mora-Melià, D., y Candia-Véjar, A. (2016). The multiple team formation problem using sociometry. *Computers & Operations Research*, **75**: 150–162.
- Guzzo, R. A. y Dickson, M. W. (1996). Teams in organizations: Recent research on performance and effectiveness. *Annual review of psychology*, **47**(1): 307–338.
- Hackman, J. R. (2011). *Collaborative intelligence: Using teams to solve hard problems*. Berrett-Koehler Publishers.
- Herrera, F., Herrera-Viedma, E., y Martínez, L. (2000). A fusion approach for managing multi-granularity linguistic term sets in decision making. *Fuzzy sets and systems*, **114**(1): 43–58.
- Hlaotittinun, O., Bonjour, E., y Dulmet, M. (2007). A team building approach for competency development. En: *Industrial Engineering and Engineering Management, 2007 IEEE International Conference on*. IEEE, pp. 1004–1008.
- Hlaotittinun, O., Bonjour, E., y Dulmet, M. (2008). A multidisciplinary team building method based on competency modelling in design project management. *International Journal of Management Science and Engineering Management*, **3**(3): 163–175.
- Huang, J.-B., Sun, X., Zhou, Y., y Sun, H.-L. (2017). A team formation model with personnel work hours and project workload quantified. *The Computer Journal*, **60**(9): 1382–1394.
- Huberman, B. A., Romero, D. M., y Wu, F. (2008). Social networks that matter: Twitter under the microscope. *arXiv preprint arXiv:0812.1045*.
- Jackson, S. E., Brett, J. F., Sessa, V. I., Cooper, D. M., Julin, J. A., y Peyronnin, K. (1991). Some differences make a difference: Individual dissimilarity and group heterogeneity as correlates of recruitment, promotions, and turnover. *Journal of applied psychology*, **76**(5): 675.
- Jelić, S. y Ševerdija, D. (2018). Government formation problem. *Central European Journal of Operations Research*, **26**(3): 659–672.
- John, O. P. y Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, **2**(1999): 102–138.
- Juárez, J. y Brizuela, C. A. (2018). A multi-objective formulation of the team formation problem in social networks: preliminary results. En: *Proceedings of the Genetic and Evolutionary Computation Conference*. ACM, pp. 261–268.
- Juárez, J., Santos, C., y Brizuela, C. A. (2021). A comprehensive review and a taxonomy proposal of team formation problems. *ACM Computing Surveys (CSUR)*, **54**(7): 1–33.
- Jung, C. (2016). *Psychological types*. Routledge.
- Kargar, M. y An, A. (2011). Discovering top-k teams of experts with/without a leader in social networks. En: *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, pp. 985–994.

- Kargar, M., An, A., y Zihayat, M. (2012). Efficient bi-objective team formation in social networks. En: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 483–498.
- Kargar, M., Zihayat, M., y An, A. (2013). Finding affordable and collaborative teams from a network of experts. En: *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM, pp. 587–595.
- Katzenbach, J. R. y Smith, D. K. (2015). *The wisdom of teams: Creating the high-performance organization*. Harvard Business Review Press.
- Kellerer, H., Pferschy, U., y Pisinger, D. (2013). *Knapsack Problems*. Springer Science & Business Media.
- Korpi, E. y Ala-Risku, T. (2008). Life cycle costing: a review of published case studies. *Managerial Auditing Journal*, **23**(3): 240–261.
- Kostopoulos, K. C. y Bozionelos, N. (2011). Team exploratory and exploitative learning: Psychological safety, task conflict, and team performance. *Group & Organization Management*, **36**(3): 385–415.
- Kou, Y., Shen, D., Snell, Q., Li, D., Nie, T., Yu, G., y Ma, S. (2020). Efficient team formation in social networks based on constrained pattern graph. En: *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, pp. 889–900.
- Kozlowski, S. W. y Bell, B. S. (2003). Work groups and teams in organizations. *Handbook of psychology*, pp. 333–375.
- Kozlowski, S. W. y Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological science in the public interest*, **7**(3): 77–124.
- Kozlowski, S. W., Chao, G. T., Chang, C.-H. D., y Fernandez, R. (2016). Using big data to advance the science of team effectiveness.
- Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, **2**(1-2): 83–97.
- Laguna, M., Kelly, J. P., González-Velarde, J., y Glover, F. (1995). Tabu search for the multilevel generalized assignment problem. *European journal of operational research*, **82**(1): 176–189.
- Lappas, T., Liu, K., y Terzi, E. (2009). Finding a team of experts in social networks. En: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 467–476.
- Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al. (2009). Life in the network: the coming age of computational social science. *Science (New York, NY)*, **323**(5915): 721.
- Levandowsky, M. y Winter, D. (1971). Distance between sets. *Nature*, **234**(5323): 34.
- Li, C.-T. y Shan, M.-K. (2010). Team formation for generalized tasks in expertise social networks. En: *2010 IEEE Second International Conference on Social Computing*. IEEE, pp. 9–16.

- Li, M. y Yao, X. (2019). Quality evaluation of solution sets in multiobjective optimisation: A survey. *ACM Computing Surveys (CSUR)*, **52**(2): 1–38.
- Liberatore, F. y Quijano-Sanchez, L. (2017). What do we really need to compute the tie strength? an empirical study applied to social networks. *Computer Communications*, **110**: 59–74.
- Liemhetcharat, S. y Veloso, M. (2012). Modeling and learning synergy for team formation with heterogeneous agents. En: *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, pp. 365–374.
- Liemhetcharat, S. y Veloso, M. (2014). Weighted synergy graphs for effective team formation with heterogeneous ad hoc agents. *Artificial Intelligence*, **208**: 41–65.
- Liu, Q., Luo, T., Tang, R., y Bressan, S. (2015). An efficient and truthful pricing mechanism for team formation in crowdsourcing markets. En: *2015 IEEE International Conference on Communications (ICC)*. IEEE, pp. 567–572.
- Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, **28**(2): 129–137.
- Loiola, E. M., De Abreu, N. M. M., Boaventura-Netto, P. O., Hahn, P., y Querido, T. (2007). A survey for the quadratic assignment problem. *European journal of operational research*, **176**(2): 657–690.
- Majumder, A., Datta, S., y Naidu, K. (2012). Capacitated team formation problem on social networks. En: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 1005–1013.
- Marlow, C., Byron, L., Lento, T., y Rosenn, I. (2009). Maintained Relationships on Facebook.
- Miettinen, K. (2012). *Nonlinear multiobjective optimization*, Vol. 12. Springer Science & Business Media.
- Morgan Jr, B. B., Glickman, A. S., Woodard, E. A., Blaiwes, A. S., y Salas, E. (1986). Measurement of team behaviors in a navy environment. Reporte técnico, Battelle Columbus Labs Research Triangle Park Nc.
- Myers, I. B., McCaulley, M. H., Quenk, N. L., y Hammer, A. L. (1998). *MBTI manual: A guide to the development and use of the Myers-Briggs Type Indicator*, Vol. 3. Consulting Psychologists Press Palo Alto, CA.
- Niveditha, M., Swetha, G., Poornima, U., y Senthilkumar, R. (2017). A genetic approach for tri-objective optimization in team formation. En: *2016 Eighth International Conference on Advanced Computing (ICoAC)*. IEEE, pp. 123–130.
- Okimoto, T., Schwind, N., Clement, M., Ribeiro, T., Inoue, K., y Marquis, P. (2015). How to form a task-oriented robust team. En: *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, pp. 395–403.

- Okimoto, T., Ribeiro, T., Bouchabou, D., y Inoue, K. (2016). Mission oriented robust multi-team formation and its application to robot rescue simulation. En: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, IJCAI'16, p. 454–460.
- Onnela, J.-P., Waber, B. N., Pentland, A., Schnorf, S., y Lazer, D. (2014). Using sociometers to quantify social interaction patterns. *Scientific reports*, **4**: 5604.
- Papadimitriou, C. H. y Steiglitz, K. (1998). *Combinatorial optimization: algorithms and complexity*. Courier Corporation.
- Pentico, D. W. (2007). Assignment problems: A golden anniversary survey. *European Journal of Operational Research*, **176**(2): 774–793.
- Pentland, A. (2012). The new science of building great teams. *Harvard business review*, **90**(4): 60–69.
- Pérez-Toledano, M. A., Rodriguez, F. J., García-Rubio, J., y Ibanez, Sergio José, j. v. n. y. p. (????). Players' selection for basketball teams, through performance index rating, using multiobjective evolutionary algorithms.
- Prasad, B. (1996). *Concurrent engineering fundamentals*, Vol. 1. Prentice Hall Englewood Cliffs, NJ.
- Rangapuram, S. S., Bühler, T., y Hein, M. (2013). Towards realistic team formation in social networks based on densest subgraphs. En: *Proceedings of the 22nd international conference on World Wide Web*. ACM, pp. 1077–1088.
- Recardo, R. J., Wade, D., Mention, C. A., y Jolly, J. A. (1996). Chapter 8 - team assessment instruments. En: R. J. Recardo, D. Wade, C. A. Mention, y J. A. Jolly (eds.), *Teams*. Gulf Professional Publishing, Boston, pp. 211 – 249.
- Reich, G. y Widmayer, P. (1989). Beyond steiner's problem: A vlsi oriented generalization. En: *International Workshop on Graph-theoretic Concepts in Computer Science*. Springer, pp. 196–210.
- Rishika, R. y Ramaprasad, J. (2019). The effects of asymmetric social ties, structural embeddedness, and tie strength on online content contribution behavior. *Management Science*, **65**(7): 3398–3422.
- Rostami, P. y Neshati, M. (2019). T-shaped grouping: Expert finding models to agile software teams retrieval. *Expert Systems with Applications*, **118**: 231–245.
- Rozovsky, J. (2015). The five keys to a successful google team. *Retrieved from*.
- Saaty, T. L. (1990). How to make a decision: the analytic hierarchy process. *European journal of operational research*, **48**(1): 9–26.
- Salas, E., Cooke, N. J., y Rosen, M. A. (2008). On teams, teamwork, and team performance: Discoveries and developments. *Human factors*, **50**(3): 540–547.
- Salas, E., Reyes, D. L., y McDaniel, S. H. (2018). The science of teamwork: Progress, reflections, and the road ahead. *American Psychologist*, **73**(4): 593.

- Selvarajah, K., Zadeh, P. M., Kargar, M., y Kobti, Z. (2019). Identifying a team of experts in social networks using a cultural algorithm. *Procedia Computer Science*, **151**: 477–484.
- Shoda, Y., Mischel, W., y Wright, J. C. (1994). Intraindividual stability in the organization and patterning of behavior: Incorporating psychological situations into the idiographic analysis of personality. *Journal of personality and social psychology*, **67**(4): 674.
- Sohlenius, G. (1992). Concurrent engineering. *CIRP annals*, **41**(2): 645–655.
- Solanas, A., Salafranca, L., Riba, C., Sierra, V., y Leiva, D. (2006). Quantifying social asymmetric structures. *Behavior Research Methods*, **38**(3): 390–399.
- Solansky, S. T. (2011). Team identification: a determining factor of performance. *Journal of Managerial Psychology*, **26**(3): 247–258.
- Strnad, D. y Guid, N. (2010). A fuzzy-genetic decision support system for project team formation. *Applied soft computing*, **10**(4): 1178–1187.
- Swezey, R. W. y Salas, E. E. (1992). *Teams: Their training and performance..* Ablex Publishing.
- Tavana, M., Azizi, F., Azizi, F., y Behzadian, M. (2013). A fuzzy inference system with application to player selection and team formation in multi-player sports. *Sport Management Review*, **16**(1): 97–110.
- Tseng, T.-L. B., Huang, C.-C., Chu, H.-W., y Gung, R. R. (2004). Novel approach to multi-functional project team formation. *International Journal of Project Management*, **22**(2): 147–159.
- Tuckman, B. W. (1965). Developmental sequence in small groups. *Psychological bulletin*, **63**(6): 384.
- Van de Water, T., Van de Water, H., y Bukman, C. (2007). A balanced team generating model. *European Journal of Operational Research*, **180**(2): 885–906.
- Van Leekwijck, W. y Kerre, E. E. (1999). Defuzzification: criteria and classification. *Fuzzy sets and systems*, **108**(2): 159–178.
- Van Staden, W. J. y Van der Poel, E. (2016). Team formation in digital forensics. En: *2016 Information Security for South Africa (ISSA)*. IEEE, pp. 91–97.
- Verma, S., Pant, M., y Snasel, V. (2021). A comprehensive review on nsga-ii for multi-objective combinatorial optimization problems. *IEEE Access*, **9**: 57757–57791.
- Wang, X., Zhao, Z., y Ng, W. (2015). A comparative study of team formation in social networks. En: *International conference on database systems for advanced applications*. Springer, pp. 389–404.
- Wang, Z., Yan, H.-S., y Ma, X.-D. (2003). A quantitative approach to the organisation of cross-functional teams in concurrent engineering. *The International Journal of Advanced Manufacturing Technology*, **21**(10-11): 879–888.
- West, M. A. (2012). *Effective teamwork: Practical lessons from organizational research*. John Wiley & Sons.

- Wi, H., Oh, S., Mun, J., y Jung, M. (2009). A team formation model based on knowledge and collaboration. *Expert Systems with Applications*, **36**(5): 9121–9134.
- Witte, E. H. y Davis, J. H. (1996). Understanding group behavior, vol. 2: Small group processes and interpersonal relations.
- Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., y Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *science*, **330**(6004): 686–688.
- Xu, Y., Ma, J., y Guo, X. (2012). Modeling researchers' characteristics for the formation of research team. En: *Preoceedings of PACIS*. pp. 252–259.
- Yaakob, S. B. y Kawata, S. (1999). Workers' placement in an industrial environment. *Fuzzy sets and systems*, **106**(3): 289–297.
- Zadeh, L. (1965). Fuzzy sets. *Information and Control*, **8**(3): 338 – 353.
- Zakarian, A. y Kusiak, A. (1999). Forming teams: an analytical approach. *IIE transactions*, **31**(1): 85–97.
- Zhang, L. y Zhang, X. (2013). Multi-objective team formation optimization for new product development. *Computers & Industrial Engineering*, **64**(3): 804–811.
- Zhang, Q. y Li, H. (2007). Moea/d: A multiobjective evolutionary algorithm based on decomposition. *IEEE Transactions on evolutionary computation*, **11**(6): 712–731.
- Zhou, Y., Huang, J.-B., Jia, X.-L., y Sun, H.-L. (2017). On participation constrained team formation. *Journal of Computer Science and Technology*, **32**(1): 139–154.