

# SEGURIDAD DE LA INFORMACIÓN Y ANÁLISIS DE REDES SOCIALES EN EL DEPÓSITO

## INFORMATION SECURITY AND SOCIAL NETWORKING ANALYSIS ON THE DEPOSIT

Juan David Céspedes Zapata<sup>a</sup>  
juancepedes@usantotomas.edu.co

---

### Resumen

El área de seguridad de la información que pertenece al DEPÓSITO tiene como obligación realizar monitoreos con el fin de preservar los activos de información de la empresa, de tal modo que pueda salvaguardar su integridad, confidencialidad y disponibilidad. Dentro de los monitoreos que son propios de la gestión del área de seguridad de la información existe uno dedicado a la verificación de las actividades de los usuarios con los sistemas funcionales y en particular con las bases de datos. Debido a que estas bases son vulnerables, por el alto volumen de transaccionalidad que se maneja en el día a día, esta monografía muestra el potencial que tiene el Social Network Analysis para detectar el grado de vulnerabilidad y optimizar de esta forma los procesos de la seguridad de la información en el DEPÓSITO.

**Palabras clave:** Seguridad de la Información, Activos de Información, Integridad confidencialidad, Disponibilidad, Social Network .

### Abstract

The area of information security that belong to the DEPOSIT have the duty to develop monitoring activities with the main objective of preserve the information assets of the enterprise ,so that, can safeguard their integrity,confidentiality and availability. Within the monitoring that are specific to the management of the area exists one dedicated to the verification of the activities of the users with functional systems in particular, with data bases. Due to these bases are vulnerable, cause the high volume of transactionality that hobble day by day, this monography show the potential that has Social Network anylisis to detect the degree of vulnerability and optimize in this form the information of security process on the deposit.

**Keywords:** Information Security,Information assets, integrity,confidentiality ,availability,Social Network.

---

<sup>a</sup>Estudiante de estadística Universidad Santo Tomás Bogotá

# Índice

|                                                                         |           |
|-------------------------------------------------------------------------|-----------|
| <b>1. Introducción</b>                                                  | <b>3</b>  |
| <b>2. Planteamiento del Problema General</b>                            | <b>4</b>  |
| <b>3. Objetivos de la Investigación</b>                                 | <b>5</b>  |
| 3.1. Objetivo General . . . . .                                         | 5         |
| 3.2. Objetivos Específicos . . . . .                                    | 5         |
| <b>4. Marco Teórico</b>                                                 | <b>6</b>  |
| 4.1. Teoría Clásica de Social Network Analysis . . . . .                | 6         |
| 4.1.1. Dos Conjuntos de actores . . . . .                               | 7         |
| 4.1.2. Relaciones Múltiples . . . . .                                   | 8         |
| 4.1.3. Relación simple . . . . .                                        | 8         |
| 4.1.4. Relación Multirelacional . . . . .                               | 9         |
| 4.1.5. Tipos especiales de grafos . . . . .                             | 9         |
| 4.1.6. Medidas dentro de un Grafo . . . . .                             | 10        |
| 4.1.7. Grafo Direccionado . . . . .                                     | 12        |
| 4.1.8. Indegree and Outdegree . . . . .                                 | 13        |
| 4.1.9. Centralidad y prestigio . . . . .                                | 14        |
| 4.1.10. Medidas descriptivas . . . . .                                  | 15        |
| 4.2. Teoría no clásica Parte 1: Redes Ponderadas . . . . .              | 19        |
| 4.2.1. Coeficiente de agrupamiento en redes ponderadas . . . . .        | 20        |
| 4.2.2. Shortest Paths . . . . .                                         | 23        |
| 4.2.3. Centralidad de nodo en redes ponderadas . . . . .                | 26        |
| 4.3. Teoría no clásica Parte 2 :Redes Ponderadas de dos modos . . . . . | 30        |
| 4.3.1. ¿Cómo Analizar las redes de dos modos? . . . . .                 | 31        |
| 4.3.2. Proyección . . . . .                                             | 32        |
| 4.3.3. Redes binarias de dos modos . . . . .                            | 32        |
| 4.3.4. Redes Ponderadas de dos modos . . . . .                          | 33        |
| 4.3.5. Caminos más cortos . . . . .                                     | 35        |
| 4.3.6. Centralidad de nodo en redes de dos modos . . . . .              | 37        |
| 4.3.7. Cercanía e intermediación . . . . .                              | 38        |
| 4.3.8. Clustering . . . . .                                             | 39        |
| <b>5. Marco Metodológico</b>                                            | <b>39</b> |
| <b>6. Hallazgos en el desarrollo de cada etapa del proyecto</b>         | <b>41</b> |

|                                                                                                                                                         |           |
|---------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| 6.1. Hallazgos en la etapa de recolección de los datos . . . . .                                                                                        | 41        |
| 6.2. Hallazgos en la etapa de definición de actores . . . . .                                                                                           | 41        |
| 6.2.1. Grafo Select . . . . .                                                                                                                           | 42        |
| 6.2.2. Grafo Delete . . . . .                                                                                                                           | 42        |
| 6.2.3. Grafo Insert . . . . .                                                                                                                           | 42        |
| 6.2.4. Grafo Update . . . . .                                                                                                                           | 42        |
| 6.3. Hallazgos en la etapa Descriptiva . . . . .                                                                                                        | 47        |
| 6.3.1. Outdegree Clásico . . . . .                                                                                                                      | 47        |
| 6.3.2. Outdegree Ponderado(no clásico) . . . . .                                                                                                        | 53        |
| 6.3.3. Densidades del grafo Ponderado y Clásico por su Outdegree . . . . .                                                                              | 59        |
| 6.3.4. Análisis del Indegree Clásico. . . . .                                                                                                           | 59        |
| 6.3.5. Análisis del Indegree Ponderado . . . . .                                                                                                        | 65        |
| 6.3.6. Densidad del Grafo por Indegree . . . . .                                                                                                        | 72        |
| 6.3.7. Conectividad . . . . .                                                                                                                           | 72        |
| 6.3.8. Cercanía . . . . .                                                                                                                               | 72        |
| 6.4. Hallazgos en la etapa de formulacion de criterios para guiar el Monitoreo y generación<br>de alertas tempranas para situaciones atípicas . . . . . | 74        |
| 6.5. Conclusiones . . . . .                                                                                                                             | 75        |
| <b>A. Códigos</b>                                                                                                                                       | <b>75</b> |

## 1. Introducción

EL DEPÓSITO, que tiene domicilio en Bogotá D.C. y sucursales en Medellín, y en Cali, es una sociedad anónima constituida con recursos de los sectores financiero y bursátil del país. La institución se encarga de la custodia, administración, compensación y liquidación de valores en DEPÓSITO de títulos valores de contenido crediticio, de participación, representativos de mercancías e instrumentos financieros que se encuentren inscritos en el Registro Nacional de Valores y Emisores -RNVE-, ya sea que se emitan, negocien o registren localmente o en el exterior.<sup>1</sup>

Esta actividad se ejerce con estricta observancia de las normas cambiarias y tributarias aplicables. EL DEPÓSITO está facultado para recibir títulos físicos por cuenta de los depositantes directos y para inmovilizarlos en bóvedas de alta seguridad permitiendo su circulación desmaterializada a través de registros electrónicos.

Dentro de su estructura organizacional, aparece un área denominada Gerencia de Riesgos y Cumplimiento, que tiene como objetivo responder a la necesidad de prevenir y controlar los riesgos que podrían amenazar el desarrollo normal de la operación del DEPÓSITO como actor vital del mercado de capitales colombiano.

Para gestionar el riesgo operativo, EL DEPÓSITO cuenta con políticas, procedimientos así como manuales que le permiten mantener un perfil de riesgo adecuado en respuesta a las necesidades del mercado de valores. De este modo, periódicamente identifica, mide, controla y monitorea los riesgos relacionados con el recurso humano, los procesos, la tecnología, la infraestructura y los acontecimientos externos. En general, su tarea se centra en:

- Gestión de Continuidad del Negocio
- Gestión de Riesgos
- Gestión de Seguridad Corporativa
- Prevención Lavado de Activos y Financiación del Terrorismo
- Gestión de Sistema de Control Interno
- Gestión de Seguridad de la Información

Este trabajo, surgió de la necesidad permanente del área de Gestión de Seguridad de la Información, la cual es la encargada de implementar las medidas preventivas y reactivas del DEPÓSITO y de los sistemas tecnológicos que permiten resguardar y proteger la información buscando mantener la confidencialidad, la disponibilidad e integridad de la misma.

En la seguridad de la información es importante señalar que su manejo, en gran parte se basa en la tecnología. Ella permite discriminar el grado de confidencialidad, disponibilidad e Integridad. Si esto no sucediera, este material podría ser divulgado, mal utilizado, robado, borrado o sabotado, afectando considerablemente los activos de información. La información es poder y el acceso a esta debe ser controlado debido a las posibilidades estratégicas que ofrece tener acceso a cierta información, ésta se clasifica de acuerdo al manual de políticas internas del DEPÓSITO como:

- Crítica: Es indispensable para la operación de la empresa.
- Sensible: Debe de ser conocida por las personas autorizadas
- Privada: Debe de ser conocida únicamente por el personal de la empresa
- Pública: Esta disponible para cualquier individuo

Este estudio, abordó la problemática desde la seguridad de la información, lo que implica que esta enfocado desde el punto de vista de la arquitectura correlacionada<sup>2</sup> en el *data protection*. De esta forma es

<sup>1</sup> Tomado de página web del DEPÓSITO

<sup>2</sup> ver página

posible estudiar las interrelaciones entre los usuarios y las bases de datos y estimar su grado de vulnerabilidad.

La reducción o eliminación de riesgos asociados a una cierta información es el objeto de la seguridad de la información y la seguridad informática.<sup>3</sup>

Precisamente, dentro de la información de naturaleza crítica se encontraron las bases de datos del DEPÓSITO y un modo de gestionar la protección de estas fue la creación de una herramienta analítica que mostrara de un manera eficiente las diferentes relaciones de los usuarios con las bases de datos. Para lograr este propósito se usó el Social Network Analysis (SNA) con el cual fue posible la recolección de datos, construcción de relaciones o flujos entre los usuarios, las bases de datos y su análisis hasta obtener el conocimiento suficiente que permitió detectar tempranamente situaciones atípicas elevando de esta forma el nivel de seguridad y protección de los activos del DEPÓSITO.

Con el uso de Social Network Analysis (SNA) se optimizó la descripción de las relaciones entre los usuarios y las bases de datos, mejoró la visualización de las mismas, permitió minimizar los tiempos de búsqueda y los estados computacionales<sup>4</sup> y facilitó la reacción de los mecanismos de la seguridad de la información<sup>5</sup>.

Para la aplicación y desarrollo del SNA sobre los activos del DEPÓSITO, particularmente sobre las bases de datos, se recolectaron los datos relacionales proporcionados por el DAM<sup>6</sup> en forma de *edgelist as cases* y luego dado el carácter multirelacional se construyó una matriz de adyacencia multirelacional ponderada bipartita. Con estas construcciones se logró describir la información y mejorar la visualización destacando las interrelaciones más relevantes a través de diferentes herramientas.

Con la matriz de adyacencia multirelacional bipartita, se procedió a aplicar las herramientas descriptivas y de visualización del SNA, obteniendo numerosos resultados que tuvieron que ser analizados hasta descubrir situaciones atípicas las cuales fueron tratadas y contrarrestadas a la luz de lo establecido por el manual de políticas de seguridad de la información del DEPÓSITO.

Finalmente, El desarrollo de ésta monografía buscó aportar desde el punto de vista práctico una aproximación estadística a la mejora de la seguridad de la información en el DEPÓSITO, Su elaboración se convirtió en la oportunidad de aprender nuevos conocimientos y aplicaciones que hayan su fundamento en la carrera profesional de estadística.

## 2. Planteamiento del Problema General

El DEPÓSITO monitorea a través de área de Gestión de la Seguridad Información las potenciales amenazas sobre los diferentes activos de información para garantizar la confiabilidad, disponibilidad e integridad, de los mismos. En ese sentido, para el DEPÓSITO resulta de especial importancia describir

---

<sup>3</sup>Los términos seguridad de la información, seguridad informática y garantía de la información son usados frecuentemente como sinónimos porque todos ellos persiguen una misma finalidad al proteger la confidencialidad, integridad y disponibilidad de la información. Sin embargo, no son exactamente lo mismo existiendo algunas diferencias sutiles. Estas diferencias radican principalmente en el enfoque, las metodologías utilizadas, y las zonas de concentración.

- La seguridad de la información involucra la implementación de estrategias que cubran los procesos en donde la información es el activo primordial. Estas estrategias deben tener como punto primordial el establecimiento de políticas, controles de seguridad, tecnologías y procedimientos para detectar amenazas que puedan explotar vulnerabilidades y que pongan en riesgo dicho activo, es decir, que ayuden a proteger y salvaguardar tanto información como los sistemas que la almacenan y administran.
- La seguridad informática, es el área de la informática que se enfoca en la protección de la infraestructura computacional y todo lo relacionado con esta y, especialmente, la información contenida o circulante.

<sup>4</sup>“Los estados computacionales disminuyen el volumen de instrucciones, registros, memoria y optimiza los procesos de planificación” Tanenbaum, A. Dos procesos e hilos. Sistemas operativos modernos. Tercera Edición. Prentice Hall. página 1076

<sup>5</sup>Batram A., Navegar por la complejidad. disponible en [www.granica.com](http://www.granica.com)

<sup>6</sup>Database Activity Monitoring

y crear criterios de interpretación que permitan visualizar las interacciones entre los diferentes usuarios y sus bases de datos.

Las potenciales amenazas hoy pueden venir de muchas partes. No se puede desconocer que existen diferentes actores<sup>7</sup>, cuyo esquema de comportamiento o nivel de conocimiento informático puede ser empleado de forma destructiva afectando en diferentes grados los activos de información pudiendo inclusive afectar el core del negocio.

Frente a esta realidad, el DEPÓSITO ha generado una serie de procedimientos y sobre ellos sobresale un software denominado DAM, basado en un esquema de capas de control de acceso para impedir que este tipo de daños sucedan y defender la integridad de sus bases de datos, proteger la propiedad intelectual, planes y dinámicas del negocio, las aplicaciones críticas, prevenir ciberataques y amenazas internas entre otras situaciones.

El DAM también se encarga de aplicar los controles propios de las auditorías que con el pleno auge de la revolución digital, son cada vez más exigentes en materia de hacer cumplir las políticas, los derechos y los controles de auditoría. Adicionalmente el DAM, surge como un modo de defenderse de la naturaleza continua de los comportamientos de usuarios y proveer la visibilidad de los accesos, usando una clasificación sofisticada de usuarios no autorizados y bots maliciosos<sup>8</sup>.

La interacción continua de usuarios con las bases de datos de la empresa por medio de actividades SQL<sup>9</sup> monitoreados por el DAM muestran sin embargo, la necesidad de complementar estos análisis con otras herramientas y criterios de interpretabilidad que ayuden a optimizar los procesos de seguridad de información del depósito. Detectada la necesidad y para resolver esta situación, en este trabajo de grado, se recurrió al Social Network Analysis y su aplicación como una contribución al monitoreo de las bases de datos clasificadas como activos críticos<sup>10</sup> y evaluar los posibles vectores de ataque.

Por estas razones, para el desarrollo del trabajo de grado se formuló la siguiente pregunta como guía de investigación:

¿Es posible describir y crear criterios de interpretación de las interrelaciones entre los usuarios y las bases de datos, a través de técnicas de Social Networking Analysis, que permitan elevar el nivel de seguridad de la información de las operaciones que realizan los usuarios sobre las bases de datos del DEPÓSITO?

### 3. Objetivos de la Investigación

#### 3.1. Objetivo General

Describir y crear criterios de interpretación de las interrelaciones entre los usuarios y las bases de datos, a través de técnicas de Social Networking Analysis (SNA), que permitan elevar el nivel de seguridad de la información de las operaciones que realizan los usuarios sobre las bases de datos del DEPÓSITO.

#### 3.2. Objetivos Específicos

1. Recolectar datos a través del software diseñado por IBM conocido como DAM, el cual muestra las diferentes operaciones realizadas por los usuarios con las bases de datos del DEPÓSITO registradas por segundo.
2. Describir a través de un grafo el conjunto de relaciones directas (ties) que se dan entre los usuarios y las bases de datos del DEPÓSITO.

---

<sup>7</sup>ver Página

<sup>8</sup>DAM-Database Activity Monitoring. IBM. Uso privativo de la empresa.

<sup>9</sup>Select:seleccionar, Delete: Borrar, Update: Actualizar, Insert: Insertar

<sup>10</sup>de acuerdo con la matriz de clasificación de activos del DEPÓSITO

3. Medir a través de relaciones, la importancia de los usuarios y las bases de datos a través de medidas de centralidad y prestigio.
4. Identificar y analizar relaciones recurrentes para descubrir situaciones atípicas.

## 4. Marco Teórico

Para el desarrollo del tema propuesto, y comprensión del SNA sobre los activos críticos del depósito, se recurrió a la obra de Wassermann (1994)<sup>11</sup> y Tore Opsahl (2009)<sup>12</sup>. De sus obras se recogieron conceptos centrales para definición de una red, descripción, medición, y análisis de sus comportamientos. El análisis de redes sociales tiene sus raíces teóricas en el trabajo de sociólogos tempranos como George Simmel y Émile Durkheim, que escribió sobre la importancia de estudiar los patrones de relaciones que conectan los actores sociales. Los científicos sociales han utilizado el concepto de “redes sociales” desde principios del siglo 20 para connotar un complejo conjunto de relaciones entre los miembros de los sistemas sociales en todas las escalas, desde interpersonal hasta el internacional. En la década de 1930 se comenzó a usar el término de forma sistemática para denotar los patrones de conexiones, que abarcan conceptos tradicionalmente utilizados por el público y los utilizados por los científicos sociales. Estudiosos como Ronald Burt, Kathleen Carley, Mark Granovetter, David Krackhardt, Edward Laumann, Anatol Rapoport, Barry Wellman, Douglas R. White, y Harrison White expandieron el uso de análisis de redes sociales sistemáticas<sup>13</sup>. Hoy sus aplicaciones han encontrado diversos espacios académicos, y se ha convertido en una herramienta para el seguimiento y explicación de diversas actividades como la lucha contra el lavado de dinero y el terrorismo.

Basados en los teóricos escogidos, se hizo un breve repaso de los conceptos más importantes que sustentan la descripción de una red, su medición e interpretación. El análisis de redes sociales así enfocado, permitió el estudio de los agentes en la estructura de la red, a través de un exhaustivo análisis de las medidas de centralidad y prestigio, permitió observar las relaciones de poder, de protagonismo, confianza, etc., así como la detección de comunidades, grupos, etc. debido a la visualización de clústeres específicos. Sin duda, este tipo de análisis contribuye a la seguridad de la información del DEPÓSITO.

### 4.1. Teoría Clásica de Social Network Analysis

El presente trabajo, fundó su enfoque teórico y metodológico en las ideas clásicas expresadas por Wasserman(1994) en su libro Social Network Analysis. Methods and Applications. En esa misma dirección el SNA, resulta ser un proceso de investigación de las estructuras a través del uso de la red y los grafos. Se caracterizan estas redes por usar términos de nodos ( actores individuales dentro de la red ) y los lazos o bordes (relaciones o interacciones) que los conectan<sup>14</sup>.

La Teoría de social Networking Analysis tiene su fundamento en la Teoría matemática de los grafos y en la teoría estadística. Tiene por objeto el estudio de los actores en terminos de nodos y los lazos o interacciones que los conectan.

De acuerdo con Wasserman(1994). Se debe identificar un conjunto de  $g$  actores, este conjunto definido como  $\eta$ , donde  $\eta$  a partir de ahora se referirá a la colección de nodos dentro de un grafo es decir:

$$\eta = \{\eta_1, \eta_2, \dots, \eta_g\}$$

Hablando de las relaciones entre actores, es decir, para  $\eta_i$  y  $\eta_j$ , se definirá el concepto de fuerza de interacción como qué tan frecuentemente interactúa  $\eta_i$  con  $\eta_j$ .

<sup>11</sup>Wasserman, S. Faust, K. Social Network Analysis. Methods and Applications. 1994.

<sup>12</sup>Tore Opsahl. Structure and Evolution of Weighted Networks. University of London (Queen Mary College), London, UK. Thesis 2009, 104-122.

<sup>13</sup>Linton Freeman, The Development of Social Network Analysis. Vancouver: Empirical Press, 2006.

<sup>14</sup>Wasserman, S. Faust, K. Social Network Analysis. Methods and Applications. 1994.

Si la relación es direccional. El valor la relación de actores  $\eta_i$  con  $\eta_j$  no será el mismo de  $\eta_j$  con  $\eta_i$ , si la relación esta presente se considerará para un conjunto especial de pares que serán denotados con la letra  $\iota$ .

Si bajo un conjunto de pares está bajo consideración la relación  $(\eta_i, \eta_j)$  y adicionalmente la relación esta presente se denotará como:

$$\eta_i \rightarrow \eta_j$$

los elementos de los pares ordenados  $\iota$  tienen una relación que es persistente, se define el conjunto

$$\iota = \{\iota_1, \iota_2, \dots, \iota_g\}$$

Estas relaciones pueden ser mostradas en un gráfico mediante líneas rectas que posteriormente se analizarán. Con base en estos conceptos se puede definir el grafo como un conjunto de nodos y un conjunto de interacciones o líneas llamadas *ties* matemáticamente definida por los dos conjuntos como:

$$G := (\eta, \iota)$$

para Wasserman(1994) las relaciones tienen diferentes características y en ese sentido se pueden tipificar de la siguiente manera:

- las relaciones dentro de un grafo se definirán de tipo dicotómico.
- Cuando las relaciones no son direccionales es imposible distinguir entre la relación de  $(\eta_i, \eta_j)$  respecto de  $(\eta_j, \eta_i)$ , se definirá una relación indirecta como la existencia de reciprocidad, es decir:

$$\eta_i \leftrightarrow \eta_j$$

- Cuando las relaciones son directas existen  $g(g-1)$  Posibles ordenaciones de pares mientras que cuando las relaciones no son direccionadas es decir los pares no son ordenados existen  $\frac{g(g-1)}{2}$  posibles ordenaciones de pares.

#### 4.1.1. Dos Conjuntos de actores

En una red medir las interacciones dentro de un grupo particular de actores puede no resultar tan interesante como mirar la interacción con un segundo grupo, en este sentido.

$$\begin{aligned}\eta &= \{\eta_1, \eta_2, \dots, \eta_g\} \\ \mu &= \{\mu_1, \mu_2, \dots, \mu_k\}\end{aligned}$$

Para ello se debe considerar el conjunto  $\eta$ , refiriéndose al primer conjunto de actores y  $\mu$  refiriéndose al segundo conjunto. El conjunto  $\eta$  contiene  $g$  actores y el segundo conjunto  $\mu$  contiene  $k$  actores.

Los elementos del segundo conjunto  $\mu, \mu = \{\mu_1, \mu_2, \dots, \mu_k\}$  donde  $\mu_i$  es un elemento típico de este conjunto puede formar  $\binom{k}{2}$  parejas con los demás actores en  $\mu$

Al analizar grupos de actores, se pueden analizar dos tipos de pares, se dice que si los actores pertenecen al mismo conjunto forman parte del conjunto de pares homogéneo mientras que si pertenecen a diferentes conjuntos se dice son heterogéneos.

Es importante decir que al agregar estos dos conjuntos de actores  $\mu$  y  $\eta$  surgen particularmente en el caso relacional directo cuatro tipos de relaciones:

- el *sender* pertenece a  $\eta$  y el *receiver* pertenece a  $\eta$
- el *sender* pertenece a  $\mu$  y el *receiver* pertenece a  $\mu$
- el *sender* pertenece a  $\eta$  y el *receiver* pertenece a  $\mu$
- el *sender* pertenece a  $\mu$  y el *receiver* pertenece a  $\eta$

En la matriz sociométrica<sup>15</sup> vale la pena destacar que los actores originadores o *senders* se ven representados por las filas mientras que los *receivers* hallan su lugar en las columnas, si la relación es definida por actores en conjuntos de diferente tamaño la matriz tendrá como resultado un arreglo rectangular.

#### 4.1.2. Relaciones Múltiples

Habiendo definido anteriormente que  $\iota$  corresponde a una relación entre  $\eta_i, \eta_j$  donde  $\eta_i$  corresponde a un actor. Se define el concepto de Grafo  $G$  como un modelo que contiene dos fuentes de información: un conjunto de (nodos)actores y un conjunto de líneas(lazos o relaciones). En un grafo cada línea representa la relación desordenada de un par de diferentes actores.

$$\begin{aligned}\iota_1 &= (\eta_1, \eta_2) \\ \iota_2 &= (\eta_2, \eta_1) \\ \iota_3 &= (\eta_3, \eta_1) \\ \bigcup_1^3 \iota_i &\subseteq \iota \\ \bigcup_1^3 \eta_i &\subseteq \eta \\ \iota &\in G \\ \eta &\in G\end{aligned}$$

Es necesario aclarar, que en el grafo se puede tener más de una relación en una base de datos de redes sociales, incluso se pueden tener relaciones directas e indirectas al mismo tiempo. Ejemplos de estos tipos de relaciones son.

#### 4.1.3. Relación simple

Para este caso es necesario suponer que se tiene un conjunto  $\eta = \{\eta_1, \eta_2, \dots, \eta_g\}$  se definirá la matrix sociométrica  $X$ , como una relacion simple direccional si esta contiene el número de pares ordenados que los actores en  $\eta$  Considerese  $x_{ij}$  como el valor de la conexión del  $i$ ésimo actor contra el  $j$ ésimo actor, se define  $X$  como la sociomatrix asociada. Dado que existe un único conjunto la matriz tendrá dimensión  $g \times g$  así es posible visualizar numerosas situaciones dentro de este tipo de relaciones como lo son:

- los elementos en la diagonal de la matriz  $X$  donde  $i = j$  representan las auto elecciones (*self choices*) de los actores por relacionarse consigo mismos.
- los valores que puede tomar  $x_{ij}$  en una matriz sociometrica son 2 posibles 1 si existe una relación ,0 si no existe. Sin embargo en escenarios donde puede tomar más valores estos pueden ser fácilmente convertidos a estas 2 categorias una de las habilidades de la notación sociometrica es la capacidad de tratar con relaciones valuadas.

Como las relaciones valuadas múltiples son un caso de la relación multirelacional se procederá a tratar el siguiente caso tomando nota que este último es de particular interés para el trabajo aquí desarrollado.

<sup>15</sup>Matriz sociométrica: Matriz de unos y ceros donde se encuentran representadas las relaciones entre dos conjuntos de actores

#### 4.1.4. Relación Multirelacional

Suponga que se tienen  $R$  relaciones Medidas en un conjunto simple de actores del siguiente modo:

$$R = \{X_1, \dots, X_R\}$$

$X_R$  representa una matriz sociométrica se define un  $x_{ij}$  como la fuerza de la interacción (medida en unos y ceros) del  $i$ ésimo actor al  $j$ ésimo actor en la  $r$ ésima relación. Estas medidas son posicionadas en un conjunto de sociomatrices donde cada una de las sociomatrices son etiquetadas, es decir clasificadas de acuerdo con la relación múltiple de tal modo que queden idénticamente cada una de tamaño  $g \times g$ . De hecho, es posible una visualización de estas como capas de una matriz tridimensional de tamaño  $g \times g \times R$ .

Como se puede apreciar y posteriormente se evidenciara en el desarrollo del trabajo de grado, se puede representar la información de un modo fácil mediante una matriz sociométrica debido a que sin importar el número de capas siempre se mantendrán los siguientes tres principios:

1. Para una relación directa donde se define un *sender* y un *receiver* la matriz sociométrica no será simétrica es decir  $x_{ij} \neq x_{ji}$ .
2. Para una relación indirecta será el valor de  $x_{ij} = x_{ji}$ .
3. Para una relación individual será representada por los elementos  $x_{ii}$  de la matriz sociométrica.

#### 4.1.5. Tipos especiales de grafos

Es de particular interés en la teoría de SNA identificar los tipos de grafos con el fin de ponerse en contexto y clasificar perfectamente el grafo que la situación demanda. Se define entonces, aunque no con todo el rigor matemático, los diferentes tipos de grafos:

1. Grafo Isomórfico: Dos Grafos son isomorficos si en el mapeo uno a uno de los nodos (actores) del primer grafo al segundo se preserva la adyacencia (relación) entre los nodos. Un mapeo uno a uno significa que para cada nodo en  $G_2$  es mapeado a uno y únicamente un nodo en  $G_1$ .

$$\begin{aligned} G_1, G_2 \text{ Grafos} \\ \eta_i \in G_1 \\ \eta_k \in G_2 \\ \phi(\eta_i) = \eta_k \quad \phi^{-1}(\eta_k) = \eta_i \end{aligned}$$

teniendo  $\eta_i$  y  $\eta_k$  sus mismas propiedades teoricas es decir el mismo degree, excentricidad, etc. siendo definido  $\phi$  como la función de mapeo del nodo en otro grafo

2. Grafo Árbol : un grafo árbol corresponde a un grafo que es acíclico es decir sin caminatas que inicien y terminen en el mismo nodo, adicionalmente contienen el mínimo número de líneas para estar conectado. En un árbol todas las líneas actúan como puente y el número total de líneas contiene el número total de nodos menos uno es decir  $g - 1$
3. Grafo Complementario: Un grafo es considerado complementario de otro si posee los mismos actores pero cuyas interacciones son complementarias o no están presentes respecto del grafo original.

4. Grafo Bipartito: Corresponde a una partición según la cual se puedan localizar cada uno de los nodos en diferentes conjuntos de tal forma que las relaciones queden descritas como la relación entre los actores del conjunto  $\eta_1$  con  $\eta_2$ , pero sin la existencia de relaciones entre los actores de un mismo conjunto. un Two mode Network con cada uno de los actores perteneciendo a dos diferentes conjuntos relacionando los individuos del primer conjunto con el segundo pueden ser representados a través de un grafo bipartito. En el caso del trabajo este tipo de grafo se adapta ya que en las bases de datos que se poseen no existe una interacción entre tablas ni entre individuos (usuarios) sino únicamente del conjunto de usuarios hacia el conjunto de las tablas.

#### 4.1.6. Medidas dentro de un Grafo

**Grado Nodal *Nodal Degree*** El grado de un nodo denotado como  $d(\eta_i)$  corresponde al número de líneas que son incidentes con el nodo, equivalentemente se puede definir como el número de nodos que son adyacentes (relacionados) al nodo y puede variar su valor desde 0 hasta  $g - 1$ , donde  $g$  corresponde al número de elementos menos él que se relacionan con el nodo anteriormente citado, un nodo con un grado de 0 es llamado aislado *isolate*.

El degree es muy fácil de computar y puede medir el grado de sociabilidad de un nodo respecto de los demás, conociendo esto se puede determinar la media del nodo como:

$$\bar{d} = \frac{\sum_{i=1}^g d(\eta_i)}{g} = \frac{2L}{g}$$

Donde  $g$  representa el número de actores en el mismo conjunto y  $L$  el número de líneas relacionadas con el nodo  $\eta_i$

**Variabilidad** La variabilidad del grado nodal mide en cierta medida que tan desigualmente los actores en una red están respecto de recibir o iniciar nuevos lazos

si el grado nodal para todos los nodos es igual se dice que el grafo es  $d$ -regular donde todos los  $d(n_i) = d$  siendo  $d$  una constante se podría entender como el máximo grado de uniformidad. La varianza del grado nodal es representada como:

$$S_d^2 = \frac{\sum_{i=1}^g (d(n_i) - \bar{d})^2}{g}$$

**Densidad del Grafo** Grado es el concepto que considera el número de líneas que inciden dentro de un nodo dentro del grafo, también se pueden considerar la proporción de líneas como un todo, realizando el cociente entre el degree correspondiente al nodo entre el número de posibles pares que pueden conformarse puede obtenerse la densidad del grafo como:

$$\Delta = \frac{L}{\frac{g(g-1)}{2}} = \frac{2L}{g(g-1)}$$

La densidad de un grafo puede ir desde  $[0, g(g-1)/2]$   $L$  representa en este caso el número total de líneas sobre la proporción de líneas totales posibles. Es importante decir que cuando un grafo está completo  $\Delta = 1$  esto sucede porque:

$L = \frac{g(g-1)}{2}$  sustituyendo en  $\Delta$  se obtiene la expresión  $\Delta = \frac{2 * \frac{g(g-1)}{2}}{g(g-1)}$  que da como resultado  $\Delta = 1$ , y sucede cuando todos los nodos tienen un grado de  $g - 1$

## distancia Geodésica

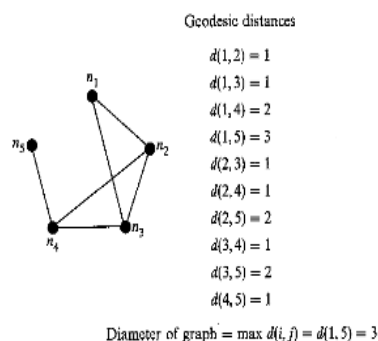


Fig. 4.9 Graph showing geodesics and diameter

Para Definirla se deben tener claros los siguientes conceptos:

- **Caminata (*Walk*):** una caminata es una secuencia de nodos y líneas que empieza y termina con nodos, en donde cada nodo es incidido por la línea que le precede y le sigue en la secuencia. Dentro de los tipos de caminata nos conviene definir dos tipos de caminata:
  - a) **Sendero (*Trail*):** un sendero es una caminata en donde todas las líneas son diferentes incluso si, los nodos están incluidos más de una vez.
  - b) **Camino (*Path*):** un camino es una caminata en donde todas las líneas y los nodos son diferentes.
- **Geodesia (*Geodesic*):** una geodesia se define como el camino más corto entre dos nodos. Si existe más de un camino corto (*shortest path*) entre dos nodos existen dos o más geodesias entre el par de nodos.

La distancia geodesica se define sencillamente como la longitud de las geodesias entre dos nodos dados. Se denotará como distancia Geodésica dados  $(\eta_i, \eta_j)$  como  $d(\eta_i, \eta_j)$ , si no hay un camino posible entre un nodo y otro la distancia será indeterminada o infinita, si el grafo no está conectado (concepto que se definirá con mayor rigor más adelante) la distancia geodesica entre dos nodos no conectados será infinito. Es importante aclarar que la distancia geodesica tiene características de simetría como la gran mayoría de distancias que se mueven en un espacio métrico es decir

$$d(\eta_i, \eta_j) = d(\eta_j, \eta_i)$$

**Excentricidad de un nodo** Fácilmente se puede definir el nodo más excéntrico como el que cuenta con la mayor distancia Geodésica entre él y los demás.

**Diámetro de un grafo** Considere la más vasta distancia geodésica entre cualquier par de nodos de un grafo, esto es la más grande excentricidad entre nodos, el diámetro de un grafo conectado es la longitud de la más larga geodesia entre cualquier par de nodos.

**Conectividad** La conectividad de un grafo consiste simplemente en una función que mide cómo el grafo permanece conectado con la eliminación de líneas o nodos y permanece como un todo.

- **Puntos de corte (*cutpoints*):** un nodo  $\eta_i$  es un punto de corte si el número de componentes en el grafo que contienen  $\eta_i$  es menor que el número de componentes que resultan después de la eliminación de  $\eta_i$  en un subgrafo esto es:

$$\begin{aligned}\eta_i &\subset \eta \subset G \text{ si} \\ \eta - \eta_i &\subset G_s\end{aligned}$$

siendo  $G$  el grafo y  $G_s$  el subgrafo resultado de la eliminación del nodo  $\eta_i$  este nodo se define como punto de corte si todos los componentes que integran el grafo  $G$  se encuentran en menor medida en el grafo  $G_s$ . Si el punto de corte es crítico entonces lo que puede suceder es que por ejemplo si existen dos conjuntos de actores interactuando ninguno de los dos conjuntos podrá comunicarse con los otros. El concepto halla su fundamento en los nodos que se considerarán fundamentales para que la red mantenga su conectividad, este concepto recibe el nombre de k-cutpoints.

- puente (*Bridge*): Si bien un cutpoint se refiere a un nodo que afecta directamente a la conectividad de un grafo existe también el concepto de puente que hace referencia a una relación que afecta directamente la conectividad de un grafo es decir su remoción hace que un importante número de actores en el grafo queden desconectado matemáticamente esto es:

$$\begin{aligned}\iota_i &\subset \iota \subset G \text{ si} \\ \iota - \iota_i &\subset G_s\end{aligned}$$

Se pueden definir aparte conceptos adicionales a la conectividad del grafo se espera en general que grafos con un muy alto degree en sus nodos y con distancias geodesicas en general cortas no tengan presencia de cutpoints, sin embargo existen una serie de términos por definir para aún dejar más clara la noción de conectividad.

- (Point conectivity and Node conectivity): Se define  $k(G)$  como el mínimo número de nodos que deben ser removidos para dejar al grafo desconectado o con una estructura trivial. Es importante resaltar que el  $k$  hace referencia a los  $k$  nodos que pueden ser removidos antes de generar la caída de conectividad del grafo, si  $k > 1$  se dice que el grafo es  $k$  nodos conectado.
- (Line Conectivity or Edge conectivity): Hace referencia al mínimo número de líneas que se pueden remover antes de colapsar la conectividad del grafo o dejarlo conectado de modo trivial. Del mismo modo que el point conectivity si  $l > 1$  se dice que el grafo es  $l$  líneas conectado.

#### 4.1.7. Grafo Direcccionado

El grafo direcccionado se diferencia del grafo no direcccionado debido a que las relaciones cuentan con un sentido y tienen ausencia de reciprocidad, una relación direccional puede ser representada por un grafo directo *digraph*. En un grafo directo  $Gd(\eta, \iota)$  Existen dos conjuntos de información un conjunto de nodos  $\eta$  y un conjunto de arcos o líneas  $\iota$  cada arco representa un par ordenado de nodos  $(\eta_i, \eta_j)$  que es dirigido desde el origen o el generador del mensaje *sender*  $\eta_i$  al receptor  $\eta_j$  *receiver*. Sea  $i$  el número posible de arcos en  $\iota$  como en cada arco existen  $g(g-1)$  posibles arcos en  $\iota$ . en un grafo de relaciones indirectas se suele usar el término adyacente para definir la relación entre dos nodos es decir:

$$\eta_i \text{ es adyacente a } \eta_j$$

en relaciones directas el término correcto sería

$$\eta_i \text{ es adyacente de } \eta_j$$

Para definir las relaciones entre parejas de actores existen tres tipos de isomorfismo a tratar:

- el duo nulo: un duo nulo carece de arco y de dirección en alguno de los nodos

- asimetría: existe una única relación entre un nodo y otro pero no puede darse la situación de que existan relaciones reciprocas.
- Mutualidad: Dado que no existe la simetría lo que podría darse es que existe una relación que define

$$\begin{aligned}\eta^i &\rightarrow \eta^j \\ \eta^j &\rightarrow \eta^i\end{aligned}$$

En este sentido se presentan 3 tipos de relaciones

- Nula: no hay comunicación
- Asimétrica: hay comunicación pero no reciprocidad
- Mutua: Existe reciprocidad

#### 4.1.8. Indegree and Outdegree

En los grafos direccionados no sólo conviene entender el degree , ya que nuevas relaciones surgen que vale la pena analizar. En un grafo no direccionado conviene analizar únicamente la adyacencia de un nodo a otro. Sin embargo, en un grafo dirigido esta situación cambia ya que cobra vital importancia la adyacencia a y la adyacencia de. Explicándolo más abiertamente un individuo cuantifica la tendencia de los actores a hacer elecciones y a recibir elecciones. Definiéndolo formalmente:

Se define el indegree  $d_i(\eta_i)$  de  $\eta_i$ : como el número de arcos de la forma  $\iota(\eta_j, \eta_i)$  para todo  $\iota_i \in \iota$  ,  $\eta_j \in \eta$

Se define el outdegree  $d_o(\eta_i)$  de  $\eta_i$ : como el número de arcos de la forma  $\iota(\eta_i, \eta_j)$  para todo  $\iota_i \in \iota$  ,  $\eta_j \in \eta$

Por medio de el outdegree y el indegree se pueden clasificar los nodos de la siguiente manera:

- Aislado  $d_i(\eta_i) = 0; d_o(\eta_i) = 0$
- Transmisor  $d_i(\eta_i); d_o(\eta_i) > 0$
- Receptor  $d_i(\eta_i) > 0; d_o(\eta_i) = 0$
- Portador (común)  $d_i(\eta_i) > 0; d_o(\eta_i) > 0$

Con la construcción de estas importantes medidas surgen los siguientes estadísticos descriptivos

$$\begin{aligned}\bar{d}_i &= \frac{\sum_{i=1}^g d_i(\eta_i)}{g} = \frac{L}{g} \\ \bar{d}_o &= \frac{\sum_{i=1}^g d_o(\eta_i)}{g} = \frac{L}{g}\end{aligned}$$

adicionalmente puede ser de interés calcular la variabilidad de los degrees ya sea indegree y outdegree

$$\begin{aligned}S_{d_i}^2 &= \frac{\sum_{i=1}^g (d_i(\eta_i) - \bar{d}_i)^2}{g} \\ S_{d_o}^2 &= \frac{\sum_{i=1}^g (d_o(\eta_i) - \bar{d}_o)^2}{g}\end{aligned}$$

la variabilidad mide que tan inequitativamente están cada uno de los actores en sus relaciones de divulgación(outdegree) con respecto a la acogida(indegree).

La densidad de un grafo es igual a la proporción de arcos presentes en el grafo dividido el posible número de arcos es decir

$$\Delta = \frac{L}{g(g-1)}$$

### 4.1.9. Centralidad y prestigio

Una de los primeros usos consiste en la identificación del actor más importante dentro de la red social en este capítulo se intentará definir el concepto de prominencia en una red social, como medidas de localización enfocadas hacia los actores, en este capítulo se mostrarán los conceptos basados en degree (grado), cercanía (closeness), information (información) y rango de un actor. En el caso que nos interesa, por ejemplo se puede definir el actor más importante con el mayor grado de indegree es decir con el mayor número de elecciones proporcionadas por otros actores.

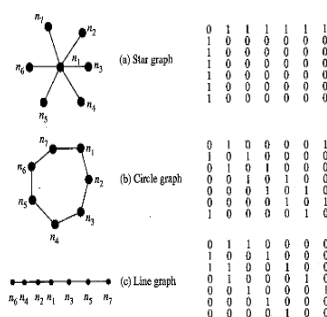


Fig. 5.1. Three illustrative networks for the study of centrality and prestige

En la figura mostrada, se puede ver la importancia de cada uno de los actores o nodos de una manera más gráfica, en general se espera que los actores más importantes dentro de una red tengan una estructura de estrella, esto se conoce como *outranking*, aunque también podría suceder que se de una estructura de anillo este tipo de estructura hace que cada uno de los individuos sean igualmente importantes dentro de la red, debido a que cada uno posee el mismo grado de centralidad son igualmente intercambiables, mientras que en la estructura de línea presentan una relevancia decreciente, es decir el grado de centralidad del primer nodo decrece respecto a los siguientes. En el Grafo existen los centroides y los centros por lo que la idea de este capítulo es de ser posible identificar que hace que un actor sea importante dentro de una red por lo que se definen los siguientes conceptos.

- **Prominencia** Se considerará que un actor es prominente si se destaca de manera visible en la red respecto de los demás actores, para determinar que tan prominente es un actor no se debe sencillamente analizar todos los lazos que lo unen sino que también es necesario observar los caminos, relaciones indirectas. Para grafos no direccionados la  $i$ -ésima fila de la matriz sociométrica correspondiente al actor  $i$  es idéntica a la  $j$ -ésima columna con el mismo conjunto la importancia está en  $g - 1$  posibles relaciones para el actor  $i$  en la matriz. Si la relación es direccional, y la  $i$ -ésima fila que representa al actor, difiere de la  $i$ -ésima columna de tal modo que la prominencia se puede establecer como las  $2(g - 1)$  posibles entradas que tiene la sobrematriz sobre el actor  $i$ , aunque el concepto todavía es vago dentro de la literatura.
- **Centralidad** Aunque la prominencia se relaciona muchísimo con el número de conexiones relacionadas con el actor, se define la centralidad de un actor con el hecho de que el simplemente participe, es decir no existe distinción entre enviar y recibir, sólo se mide que el actor participe en muchas interacciones.
- **Prestigio** El Prestigio surge de la necesidad en los grafos direccionales de distinguir entre los actores que reciben muchas interacciones respecto de los que las emiten, siendo estos primeros los de mayor importancia, es decir se dice que un actor prestigioso es aquel que recibe muchas interacciones.
- **Centralización de grupo o prestigio de grupo** La medida de centralización de grupo está basada en las diferentes medidas de centralización de cada uno de los actores, contando con la

propiedad que entre más grande valor tiene es más probable que el grafo asuma una estructura de estrella es decir un actor tienda a tener un mayor grado de centralidad que el resto de actores, es decir mide que tan variables o heterogeneos los grados de centralización de los diferentes actores son. sea:

$$C_A(n^*) = \max C_A(n_i)$$

el índice de centralidad del actor más alto dentro de la red

$$C_A(n_i)$$

el índice de centralidad de un actor cualquiera Para estas cantidades :

$$\sum_{i=1}^g (C_A(n^*) - C_A(n_i))$$

Es la diferencia entre el índice de más grande valor y los índices observados mientras:

$$\max \sum_{i=1}^g (C_A(n^*) - C_A(n_i))$$

Corresponde a la máxima posible suma teórica de las diferencias de la centralidad del actor. Se define el índice de Freeman como:

$$C_A \frac{\sum_{i=1}^g (C_A(n^*) - C_A(n_i))}{\max \sum_{i=1}^g (C_A(n^*) - C_A(n_i))}$$

Es importante destacar que  $0 < C_A < 1$  y que  $C_A(n^*) = 1$  cuando el actor domina completamente a los demás actores.

A continuación se presentan las medidas descriptivas, de modo más detallado

#### 4.1.10. Medidas descriptivas

##### I Degree

- *Actor Degree Centrality* La idea principal constituye que entre más alto sea el valor del degree más importante será el actor dentro del grafo, esto es

$$C_D = d(\eta_i)$$

una estandarización propuesta de la medida ya muy comentada es

$$C'_D = \frac{C_D}{g-1}$$

- *Group Degree Centralization* El objetivo de una medida de centralización consiste en calcular el rango o la variabilidad de los índices de actores individuales, la siguiente medida indica que tan centralizado el grado(degree) del conjunto de actores es. Fue propuesto por Freeman(1979)

$$C_D = \frac{\sum_{i=1}^g C_D(n^*) - C_D(n_i)}{(g-1)(g-2)}$$

$C_D = 1$  toma su máximo valor cuando un actor elige a todos los demás actores y ellos lo eligen a él, entre más alto sea el valor más centralizado es el grafo.

Otra medida de centralidad corresponde a la varianza de grados

$$S^2 = \sum_{i=1}^g (c_D(n_i) - \bar{c}_D)^2$$

donde  $\bar{C}_D$  es la media del índice de cada actor, la varianza en este sentido puede ser sinónimo de la dispersión o heterogeneidad de los índices de los actores. Este índice toma el valor mínimo es cero cuando todos los grados de cada uno de los actores son iguales, o la estructura del grafo es regular. Para calcular un parámetro de que tan centralizado es un grafo se puede calcular primero su grado promedio

$$\bar{C}_D = \frac{\sum C_D(n_i)}{g} = \sum x_{i+}$$

Se puede notar que esta cantidad varía entre 0 y  $g - 1$  así que para estandarizarla se debe dividir entre  $g - 1$ , al hacer esto da como resultado el cálculo de la densidad del grafo que es equivalente al grado promedio de centralidad de los grados estandarizados.

$$\frac{\sum C_D(n_i)}{g(g-1)} = \frac{\sum C'_D(n_i)}{g} = \Delta$$

La densidad es también en teoría de redes una medida estandarizada del grado. La densidad del grafo es también recomendada como una medida de cohesión de grupo.

- **Centralidad de cercanía** La segunda medida de centralidad de un actor está basada en que tan cercano es un actor a un grupo de actores. La idea principal reside en que si un actor puede rápidamente interactuar con los otros actores, esto se traduce en productividad ya que se ha demostrado que los actores con un alto nivel de cercanía son muy productivos en la transmisión de información a los otros debido a que cuentan con caminatas o paths muy cortos respecto de los otros. En este sentido la cercanía es una medida de centralidad que implica consideraciones económicas muy serias, ya que como es el inverso de la distancia un mayor grado de cercanía constituye un menor costo en la transmisión de la información.

- *Proximidad Central*: La más simple medida fue propuesta por Sabidussi(1966) el cual propuso que la cercanía de un actor debería ser medida en función de sus distancias geodésicas. En el sentido entre más grande sea el valor de sus distancias el valor de proximidad central va a tender a decrecer. Como fue evidenciado por Beauchamp(1965), los actores que ocupan posiciones centrales con respecto a la cercanía pueden ser muy productivos en materia de comunicar información a los otros. Si los actores involucrados se encuentran encargados en la resolución de problemas, las soluciones eficientes suceden cuando existe un mayor grado de cercanía que se traduce en distancias geodésicas cortas.

Este tipo de medida depende no solo de los lazos directos sino de los indirectos es decir cuando dos actores no son adyacentes.

Sea  $d(n_i, n_j)$  el número de líneas en la conexión geodésica entre los actores  $i$  y  $j$ ; como anteriormente se definió la distancia geodésica como una función de distancia. el total de la distancia de  $i$  tomada de los otros actores es  $\sum_{j=1}^g d(n_i, n_j)$ , donde la suma es tomada sobre todos los  $j \neq i$  de acuerdo con la definición de Sabidussi(1966) la cercanía de índice de un actor es:

$$C_c(n_i) = (\sum_{j=1}^g d(n_i, n_j))^{-1}$$

cuando el actor tiene el máximo grado de proximidad el índice se iguala a  $(g-1)^{-1}$ , esto sucede cuando el actor es adyacente a todos los actores, el valor toma valor de cero cuando el actor está definitivamente aislado y no puede ser alcanzable por otro nodo. Con la observación de la medida anterior se puede observar que el valor depende de  $g$ . Beauchamp(1966) propuso estandarizar los índices de tal modo que el máximo grado de proximidad central fuera igual a 1 para realizar esto multiplico  $C_c(n_i)$  por  $g-1$  es decir:

$$C'_c = \frac{g-1}{\sum_{j=1}^g d(n_i, n_j)}$$

Esta versión estandarizada del índice hace que el rango se encuentre entre cero y uno, y puede ser visto como el inverso del promedio entre el actor  $i$  y todos los otros actores, esto es cuando el actor está lo más cerca posible a los otros actores.

- *Proximidad Central de grupo* Ahora se considerará como medir la centralización de grupo usando la proximidad central de los actores. El índice de proximidad central de grupo de Freeman(1979) está basado en la estandarización de las proximidades centrales entre actores y se define como:

$$C_c = \frac{\sum_{i=1}^g (C'_c(n^*) - C'_c(n_i))}{(g-2)(g-1)/(2g-3)}$$

Donde  $C'_c(n^*)$  es el actor con mayor valor de cercanía estandarizada en un grupo de actores. Freeman muestra que el mayor valor de proximidad de grupo es  $\frac{(g-2)(g-1)}{(2g-3)}$  este índice alcanza el valor de uno cuando un actor elige a los  $g-1$  actores es decir el actor tiene una distancia geodésica de 1 con los otros actores y el otro actor tiene distancia geodésica de 2 con los remanentes  $g-2$  actores, esta es exactamente la

situación que ejemplifica un grafo estrella. Por otra parte toma valor de cero cuando las distancias geodésicas son iguales a cero.

Dentro de la cercanía también existe la varianza de los índices de cercanía entre actores representado como

$$S_c^2 = \frac{\sum_{i=1}^g (C'_c(n_i) - \bar{C}_c)^2}{g}$$

que resume la heterogeneidad entre  $C'_c(n_i)$  donde  $\bar{C}_c = \frac{C'_c(n_i)}{g}$  constituye al promedio de cercanía y la varianza toma valor mínimo de 0 en una red para cuando existen iguales distancias en todos los nodos, esto sucede cuando es homogéneo y lo menos centralizado posible (anillo). Por lo que cuando esta varianza toma valores más grandes el grafo toma una estructura más heterogénea y más centralizada.

- *Intermediación Betweenness* Las interacciones entre dos conjuntos no adyacentes, especialmente los actores que se encuentran en el camino entre ellos dos. Estos otros actores potencialmente pueden tener control sobre actores no adyacentes. por ejemplo si la distancia geodésica entre los actores  $n_2$  y  $n_3$  es  $n_2 n_1 n_4 n_3$  esto quiere decir que el camino más corto entre dos actores tiene que ir a través de  $n_1 n_4$  por lo que se podría decir que estos actores tienen el control de las interacciones entre  $n_2 n_3$ . la Intermediación hace referencia a que el actor se encuentra en la mitad o intermediando entre los actores que se encuentran al final del camino.

- *Centralidad de Intermediación entre actores* Considere la probabilidad de una comunicación o simplemente un camino de un actor j al actor k toma una ruta particular. Asumiendo que todas las rutas tienen el mismo peso y que las comunicaciones tenderán a viajar por la ruta más corta, se asume que la comunicación sigue una de las geodésias. Cuando existe más de un camino geodésico entre j y k todas las geodésias son igualmente accequibles. Sea  $g_{jk}$  el número de geodésias que une dos actores. Entonces, si todas las geodésias son igualmente elegibles para un mismo camino la probabilidad de comunicación usando cualquiera de ellas es  $\frac{1}{g_{jk}}$ . Adicionalmente considere, la probabilidad de que el actor i en la comunicación entre los dos actores, sea  $g_{jk}(n_i)$  el número de geodésias que vinculan a los dos actores que contienen al actor i. Freeman(1979) estimó esta probabilidad como  $\frac{g_{jk}(n_i)}{g_{jk}}$ . El índice de intermediación para  $n_i$  es simplemente la suma de estas probabilidades estimadas sobre los pares de actores que no incluyen al i-ésimo actor.

$$C_B(n_i) = \frac{\sum_{j < k} g_{jk}(n_i)}{g_{jk}}$$

para i distinto de j, k de tal modo que el índice cuenta cuantos actores hay entre ellos, es una suma de probabilidades. Si tienen un mínimo de cero quiere decir que no hay geodesias. Su máximo es claramente  $(g-1)(g-2)/2$  el cual corresponde al número de pares de actores que no incluyen a  $n_i$ . El índice alcanza su máximo cuando el i-ésimo actor interactúa en todas las geodesias. Como el índice depende de g se puede estandarizar del siguiente modo

$$C'_B(n_i) = \frac{C_B(n_i)}{(g-1)(g-2)/2}$$

Estandarizado de este modo toma valores entre 0 y 1 y puede ser comparado con otros índices de actores, a diferencia de los índices de cercanía los índices de intermediación estandarizados pueden ser computados incluso si el grafo no esta conectado, por lo que ciertamente es una ventaja.

- *grupo de centralización de intermediación* Los índices de centralización de grupo basados en intermediación permiten al investigador comparar las diferentes redes con respecto a la heterogeneidad de la intermediación de los miembros de las redes. el índice de centralización grupal de freeman index tiene numerador

$$\sum_{i=1}^g C_B(n^* - C_B(n_i))$$

donde  $C_B(n^*)$  es el actor con el índice de mayor valor de intermediación en el conjunto de actores. Freeman(1977) mostró que el valor máximo posible para esta suma es  $(g-1)^2(g-2)/2$  así que el índice de intermediación de grupo es

$$C_B = \frac{\sum_{i=1}^g C_B(n^*) - C_B(n_i)}{(g-1)^2(g-2)}$$

Freeman(1979) simplificó este índice

$$C_B = \frac{\sum_{i=1}^g C'_B(n^*) - C'_B(n_i)}{(g-1)}$$

El cálculo se realiza con los índices estandarizados. Freeman(1977) también demostró que el índice alcanza el valor de uno cuando el grafo posee forma de estrella y cero cuando posee forma de línea o tienen el mismo índice de intermediación.

- *Centralidad de la información* De todos los índices la medida de centralidad de Freeman es una medida basada en la intermediación de los actores y su mayor uso está en las geodésias, por su generalidad. Pero este índice asume que todas las geodesias son iguales cuando se estima una probabilidad crítica de que un actor intervenga en una geodesia particular. Sin embargo la medida de Freeman ignora el hecho de que algunos actores tienen en sus geodesias un mayor grado(degree), y que las geodesias que contienen estos actores son más afines a tener una distancia geodésica más corta y tienden a ser el camino más corto que otros. Es decir si por ejemplo si un actor tiene un grado de diez y este actor está en una geodesia, este actor es más probable que se encuentre en más geodesias que aquellos actores que tienen un grado menor. El principio de Freeman tiene sentido únicamente si todos los actores tienen el mismo grado(degree). Adicionalmente Freeman en sus conteos de intermediación se basa únicamente en el conteo de las geodesias. Lo que quiere decir que caminos con distancias más grandes que la longitud mínima de camino alcanzado por las geodesias son ignorados lo cual no es realista. Por ejemplo si considera las relaciones comunicativas existen buenas razones para que los actores elijan caminos para sus comunicaciones más largos que el camino con menor distancia geodésica. Así que para darle sentido a la noción de proximidad de intermediación para que todos los pesos sean considerados cuando se calculan conteos de intermediación.

- Centralidad de información por actor: Esta versión de centralidad se centra en la información contenida en todos los caminos originados por un actor en específico. La información de un actor promedia la información de estos caminos que en este caso están inversamente relacionados con la varianza de la transmisión de la señal de un actor a otro. Para calcular los índices de centralidad Stephenson y Zalen(1989) reconocieron la información contenida en la matriz de incidencia<sup>16</sup> donde los nodos y las conexiones que hay entre ellos son mostrados. El modelo se centra en las señales fluyendo entre todos los pares de nodos o pares de filas en la matriz de incidencia. Si  $V_{jk}$  es la varianza para estimar la señal del camino que une dos nodos  $n_j$  y  $n_k$  entonces la información asociada con este camino es simplemente  $\frac{1}{V_{jk}}$ . La información de un actor es una función de todos los caminos de información fluyendo hacia el otro actor. Para aplicar esta idea a los grafos la información del actor son funciones de una simple matriz de dimension  $g$  por  $g$ . Un componente fundamental en la fórmula es la suma de las intensidades o los valores para los que la línea es incidente con cada nodo. Esta suma es simplemente un total fila o un total columna de la sociomatriz. Esto se resume en la suma del grado de un nodo cuando la relación medida es dicotómica o la suma de de las intensidades de todas las conexiones a un nodo cuando la relación es valuada.

El cálculo empieza así, primero se crea la matriz  $A$  de dimensión  $g \times g$  cuyos elementos de la diagonal

$$a_{ii} = 1 + \text{Es suma de los valores de todas las líneas que inciden sobre el nodo } n_i$$

$$a_{ij} = 1 \text{ si } n_i \text{ y } n_j \text{ no son adyacentes } a_{ij} = 1 - x_{ij} \text{ si son adyacentes}$$

<sup>16</sup>matriz de incidencia: matriz en donde están contenidas las relaciones entre un conjunto de actores y relaciones como otro conjunto es diferente de la matriz de adyacencia que relaciona conjuntos de actores

$x_{ij}$  es el valor de la conexión entre  $i$  y  $j$  así que los elementos de  $A$  son fácilmente calculados en la sociomatriz. Posteriormente se calcula  $C = A^{-1}$  cuyos elementos son  $c_{ij}$ . Es importante destacar que la matriz  $A$  no siempre es invertible, de hecho si parte de una matriz con dos diferentes conjuntos de actores no lo es, o cuando existen muchos actores aislados por lo que se recomienda eliminarlos antes de calcular  $A$  y posteriormente  $C$ . Para obtener los índices de información se necesitan dos cantidades intermedias que corresponden a sumas de los elementos de  $c$ :

$$T = \sum_{i=1}^g c_{ii}$$

$$R = \sum_{i=1}^g c_{ij}$$

donde  $T$  es simplemente la traza de la suma de las entradas diagonales de la matriz mientras que  $R$  es cualquier suma de filas con estas cantidades se calcula la centralidad de información de  $C$  como

$$C_I(n_i) = \frac{1}{c_{ii} + (T - 2R)/g}$$

Esta medida mide que tanta de la información es contenida en los caminos que se originan y terminan en un actor en específico, el índice tiene un valor mínimo de cero pero no un valor máximo Stephenson and Zelen(1989) recomiendan que se use la información relativa de los índices obtenida por dividir cada índice de información  $C_I(n_i)$  en el total de índices

$$C'_I(n_i) = \frac{C_I(n_i)}{\sum_i C_I(n_i)}$$

este índice está restringido de cero a uno y representan una proporción del total de información que fluye en un grafo controlado por un actor individual.

- Centralización de la información de grupo El resumen de de el índice de información proporcionado por Stephenson y Zelen(1989) es simplemente el promedio de la información entre actores

$$\bar{C}_I = \sum_i (n_i)$$

este índice tiene sus limitantes ya que depende de  $g$ , desafortunadamente es difícil de comparar a través de las redes ya que los promedios no corresponden a índices de centralización una información del índice de nivel de grupo real lo corresponde la varianza de los índices de información del actor

$$S_I^2 = \frac{\sum_{i=1}^g (C'_I(n_i) - \bar{C}_I)^2}{g}$$

Desafortunadamente de acuerdo con esta metodología el índice de centralidad de información es el único índice que puede ser aplicado a las relaciones valuadas.

## 4.2. Teoría no clásica Parte 1: Redes Ponderadas

Las teorías clásicas, por ejemplo la de Wasserman(1994) fueron enriquecidas por autores como Tore Opsahl(2009)<sup>17</sup> en tanto, introdujeron, nuevas herramientas para la medición y el análisis. Por ejemplo a menudo, los datos de redes sociales consisten en relaciones valuadas, en donde la fortaleza o intensidad de las relaciones es cuantificable. Razón por la cual, en la generalización de estos grafos se suele cuantificar el valor del arco o la relación entre ellos. El dólar por ejemplo y las relaciones de comercio entre naciones, o el grado de amistad entre personas pertenecientes a un grupo.

Estos nuevos aportes, han tomado relevancia con el auge de las redes sociales, donde los criterios clásicos resultan insuficientes. Entre los tipos especiales de grafos valuados se encuentran por ejemplo, los grafos valuados o los grafos valuados con enteros (Roberts 1976), así como las cadenas de Markov. (Wasserman and Faust 1994)

<sup>17</sup>Tore Opsahl. Structure and Evolution of Weighted Networks. University of London (Queen Mary College), London, UK. Thesis 2009, 104-122.

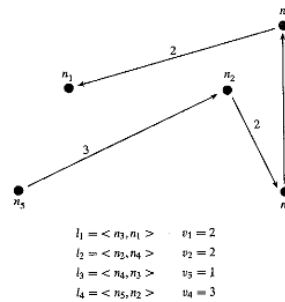
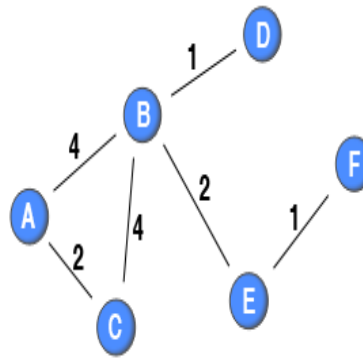


Fig. 4.23. Example of a valued directed graph

#### 4.2.1. Coeficiente de agrupamiento en redes ponderadas

Una medida fundamental que ha recibido atención tanto en la investigación teórica y empírica es el coeficiente de la agrupación. Esta medida evalúa el grado en el cual los nodos tienden a agruparse. La evidencia sugiere que en la mayoría de las redes del mundo real, y en particular las redes sociales, los nodos tienden a crear grupos fuertemente unidos que se caracterizan por una densidad relativamente alta de los lazos (Feld, 1981; Friedkin, 1984; Heider, 1946; Holland y Leinhardt, 1970; Louch, 2000; Snijders, 2001; Snijders et al, 2006; Watts y Strogatz, 1998). En las redes del mundo real, esta probabilidad tiende a ser mayor que la probabilidad media de un lazo establecido al azar entre dos nodos (Holland y Leinhardt, 1971; Wasserman y Faust, 1994; Watts y Strogatz, 1998). Tradicionalmente, las dos versiones del coeficiente de agrupación desarrollados para el ensayo de la tendencia de los nodos a agruparse en grupos muy unidas son el coeficiente de agrupación global (Doreian, 1969) y el coeficiente de agrupación local (Watts y Strogatz, 1998). La versión global fue diseñado para dar una indicación general de la agrupación en la red, mientras que el local, da una indicación de empotramiento de los nodos individuales.

**Coeficiente de agrupamiento global en redes ponderadas** El coeficiente de la agrupación global se basa en tripletas de nodos. Una tripeleta esta compuesta de tres nodos que están conectados por cualquiera de los dos (tripletas abierta) o tres (tripletas) lazos no dirigidos. Un triángulo formado por tres tripletas cerradas, uno centrado en cada uno de los nodos. El coeficiente de la agrupación global.. el número de tripletas cerradas (o 3 x triángulos) sobre el total de tripletas (tanto abiertas como cerradas). El primer intento de medir que fue hecha por Luce y Perry (1949) . Esta medida da una indicación de la agrupación en toda la red (global) , y se puede aplicar a redes no dirigidas y dirigidas (a menudo llamadas medidas de transitividad , ver Wasserman y Faust, 1994 , página 243 ) . Sin embargo, no se puede aplicar a redes ponderadas. Por ejemplo para esta red, el coeficiente de la agrupación global binaria sería 3 sobre 9 o 0.33 . Las tripletas cerradas serian  $B \rightarrow AC$  ;  $A \rightarrow BC$  ;  $A \rightarrow CB$  ; mientras que las tripletas abiertas serian  $A \rightarrow BD$  ;  $A \rightarrow BE$  ;  $C \rightarrow BD$  ;  $C \rightarrow BE$  ;  $D \rightarrow BE$  ;  $B \rightarrow EF$ . Como puede verse a partir de la red del ejemplo , los lazos más fuertes están en el interior del triángulo. Esto no es capturado por el coeficiente binario sino se consideran los pesos.



Una generalización del coeficiente de la agrupación global de redes ponderados, fue propuesto por Opsahl y Panzarasa (2009). Esta generalización requiere un valor tripleta por definir, y luego, divide el valor de las tripletas cerradas en el valor de todas las tripletas. El coeficiente de agrupación generalizada produce el mismo resultado que el coeficiente binario de agrupación global cuando se aplica a una red binaria. Esto es porque todas las tripletas tienen el mismo valor, independientemente del método utilizado para calcular los valores de la tripleta. Además, el coeficiente generalizado tiene las mismas propiedades que C. Por otra parte, si los pesos son asignados al azar en la red, el coeficiente de agrupación ponderada es igual al coeficiente binario.

Cuatro métodos se sugirieron para la definición de un valor de tripleta: la media aritmética, media geométrica, máximo y mínimo de los dos pesos de unión que forman el triplete. La media aritmética es simplemente la media de los pesos de unión. Este método es insensible a las diferencias entre dos lazos ponderados como un valor extremo que pueden tener un gran impacto en el valor de tripleta. La media geométrica de las ponderaciones asignadas a los dos lazos cuenta con el problema que si existe una tripleta formada por un lazo con un valor bajo y un lazo con un alto valor tendrá un valor más bajo que si se utiliza la media aritmética. Adicionalmente, es posible utilizar dos métodos extremos para hacer frente a diferencias en los pesos de los lazos. El método del máximo toma el valor más alto de los dos pesos, y crea una tripleta con un lazo fuerte y un lazo débil igual a una tripleta con dos lazos fuertes. Por el contrario, el método de mínimo tiene el valor más bajo de los dos pesos, y crea tripletas con un lazo fuerte y un débil igual a tripletas con dos lazos débiles. La siguiente tabla muestra las diferencias entre los métodos para definir el valor de la tripleta.

| Method          | Triplet value $\omega$ of |                            |
|-----------------|---------------------------|----------------------------|
|                 |                           |                            |
| Arithmetic mean | $(2+2)/2 = 2$             | $(1+3)/2 = 2$              |
| Geometric mean  | $\sqrt{2 \times 2} = 2$   | $\sqrt{1 \times 3} = 1.73$ |
| Maximum         | $\max(2, 2) = 2$          | $\max(1, 3) = 3$           |
| Minimum         | $\min(2, 2) = 2$          | $\min(1, 3) = 1$           |

Es imprescindible la utilización de un método apropiado para definir el valor de una tripleta y establecer como esto repercute en el resultado del coeficiente. El método debe ser elegido sobre la base de la pregunta de investigación, así como la forma en que la fuerza de los lazos se hacen operativos en pesos. Por ejemplo, en una red donde los pesos corresponden al nivel de flujo, y un lazo débil actuaría como un cuello de botella, el método mínimo podría

ser más apropiado utilizar. Para resaltar el uso potencial del coeficiente de agrupación global binario y generalizado, es posible ver si las tripletas más fuertes son más propensas a ser cerradas que las más débiles. Esto se puede hacer mediante la comparación de la versión ponderada ( $C_w$ ) para la versión binaria, en donde todos los lazos con un peso mayor que 0 se establece en la actualidad ( $C_{GT0}$ ). De este modo, los dos coeficientes se calculan sobre la misma estructura de la red. La única diferencia es que la antigua lleva el peso de los lazos en consideración. Si la razón entre el coeficiente de la agrupación ponderada y el coeficiente binario es mayor que 1, se puede argumentar que las tripletas compuestas por lazos fuertes son más propensas a ser cerradas que para tripletas compuestas por lazos débiles ( $\frac{C_w}{C_{GT0}} > 1 \Rightarrow P(x) > P(Y)$ ). Por el contrario, se puede argumentar que las tripletas compuestas por lazos débiles son más propensas a ser cerradas que las hechas por lazos fuertes si la relación es menor que 1. Para investigar empíricamente si los lazos fuertes son más propensos a ser parte de las tripletas cerradas que la de los lazos débiles, Opsahl utilizó las redes TNET. Por simplicidad la media aritmética método se utiliza para definir los valores del triplete.

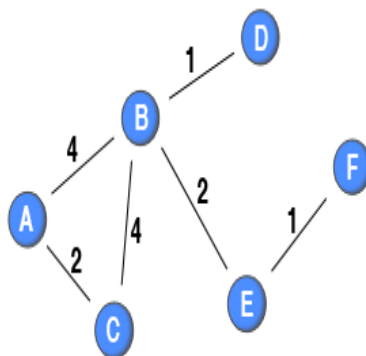
| Network                                      | Nodes | Ties  | $C_{GT0}$ | $C_w$  | $\frac{C_w}{C_{GT0}}$ |
|----------------------------------------------|-------|-------|-----------|--------|-----------------------|
| 1 Freeman's EIES network (time 1)            | 48    | 695   | 0.7627    | 0.7702 | 1.0099                |
| 2 Freeman's EIES network (time 2)            | 48    | 830   | 0.8131    | 0.8214 | 1.0102                |
| 3 Freeman's EIES network (messages)          | 32    | 460   | 0.6386    | 0.7378 | 1.1555                |
| 4 Consulting (advice)                        | 46    | 879   | 0.6932    | 0.7130 | 1.0285                |
| 5 Consulting (value)                         | 46    | 858   | 0.6764    | 0.6852 | 1.0131                |
| 6 Research team (advice)                     | 77    | 2228  | 0.6848    | 0.7127 | 1.0408                |
| 7 Research team (awareness)                  | 77    | 2326  | 0.6785    | 0.6957 | 1.0253                |
| 8 C.elegans' neural network                  | 306   | 2345  | 0.1818    | 0.2364 | 1.3009                |
| 9 US airport network                         | 500   | 2980  | 0.3514    | 0.4765 | 1.3562                |
| 10 Newman's scientific collaboration network | 16730 | 47594 | 0.3596    | 0.3178 | 0.8838                |

Como se puede ver en la tabla anterior, 9 las 10 redes tienen un coeficiente de agrupamiento con un peso que es mayor que el coeficiente de la agrupación tradicional. Esto sugiere que los triángulos en general son, de hecho, como lo sugirió Simmel (1923 [ 1950 ] ), compuestos por lazos fuertes. La décima red, muestra el resultado opuesto, es diferente de las otras nueve redes. Es una proyección de one-mode de una red de two mode, mientras que los primeros nueve redes son las redes de un modo nativo. Esta característica podría explicar el resultado observado como la proyección de un modo está construido de una manera que penaliza el peso de las relaciones dentro de triángulos, o más específicamente, los grandes grupos de nodos totalmente conectados ( Newman, 2001).

**Coeficiente de la agrupación local** El coeficiente de la agrupación local se basa en la red ego de densidad local (Scott, 2000; Uzzi y Spiro, 2005; Watts y Strogatz, 1998). Para un nodo, esta es la fracción del número de lazos presentes sobre el número total de posibles relaciones entre vecinos del nodo. Por lo tanto, el resultado oscila entre 0 y 1 : 0 si no existen vínculos entre los vecinos, y 1 si existen todos los lazos posibles.

El coeficiente de la agrupación local se basa en la red ego de densidad local (Scott, 2000; Uzzi y Spiro, 2005; Watts y Strogatz, 1998). Para un nodo, esta es la fracción del número de lazos presentes sobre el número total de posibles relaciones entre vecinos del nodo. Por lo tanto, el resultado oscila entre 0 y 1 : 0 si no existen vínculos entre los vecinos, y 1 si existen todos los lazos posibles. Para el ejemplo, los nodos A y C pueden conseguir una puntuación de 1, ya que están presentes todos los posibles vínculos entre sus vecinos, mientras que el nodo E pueden consiguen una puntuación de 0, ya que ninguno de los posibles vínculos entre sus vecinos está presentes. Los nodos D y F no consiguen un valor ya que el número de posibles vínculos entre sus vecinos es 0 ( si un nodo tiene menos de 2 vecinos, el coeficiente

es indefinido ). Para el nodo B, uno de cada seis posibles vínculos está presente, por lo que el coeficiente es de  $1/6$  o 0,1667



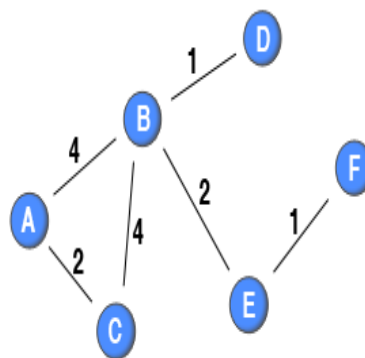
siendo los nodos de una red (por ejemplo Uzzi y Lancaster, 2004). Sin embargo, esta versión de la agrupación coeficiente adolece de tres limitaciones importantes (ver Opsahl y Panzarasa de 2009, para una discusión más larga). En primer lugar, su resultado no tiene en cuenta el peso de los lazos en la red. En segundo lugar, el coeficiente de la agrupación local no se puede calcular en redes dirigidas. En tercer lugar, una correlación negativa con el grado se encuentra a menudo en las redes del mundo real. Esto es debido al hecho de que es "más fácil" para un nodo con dos vecinos para obtener una puntuación de 1 (sólo un lazo es necesario) que para un nodo con 10 vecinos (45 lazos deben estar presentes). Recientemente, ha habido un número de intentos de extender el coeficiente de agrupación local para el caso de redes ponderadas (Barrat et al, 2004;. López-Fernández et al, 2004;. Onnela et al, 2005;. Zhang y Horvath, 2005 ). Me centraré en la medida de Barrat et al.(2004) . Ellos propusieron una generalización del coeficiente de la agrupación local a las redes ponderadas teniendo en cuenta los pesos explícitamente: En primer lugar, se asigna un valor a cada triplete de tripletes en la red basado en la media aritmética (promedio normal). Luego, para cada nodo, que suman el valor de las tripletas cerradas que se centran en el nodo y divide por el valor total de todas las tripletas se centraron en el nodo. Similarmente al coeficiente de la agrupación mundial ponderada, los valores de la triplete no tienen que ser la media aritmética, pero también podran ser la media geométrica, máximo y mínimo de los dos pesos de los lazos. Para la red de ejemplo anterior, todos los nodos, excepto el nodo B, obtendría el mismo resultado que el análisis binario. Los nodos A y C obtienen un valor de 1, porque todos sus vecinos están conectados, mientras que el nodo E obtener un valor de 0, ya que ninguno de sus vecinos están conectados. Los nodos D y E no tienen un valor, ya que tienen menos de 2 vecinos. El Nodo B no obtendra el mismo valor, ya que tiene más de 2 vecinos, y es el centro de los dos tríos abiertos y cerrados. El coeficiente de la agrupación local ponderada es de entre 0,18 y 0,36, mientras que el binario es 0,17. El coeficiente pesado aumento es un reflejo de fuertes lazos de nodo B se dirige hacia los vecinos que están conectados a sí mismos. La siguiente tabla muestra los coeficientes de la agrupación local.

#### 4.2.2. Shortest Paths

Establecer los caminos más cortos o las distancias entre los nodos ha sido durante mucho tiempo un elemento clave de la investigación de redes. Mientras que los caminos más cortos a menudo no son de interés en sí mismos, son el componente clave de una serie de medidas. En primer lugar, una pregunta popular ha sido lo que es la distancia media entre la gente? La pregunta se hizo famosa después de los seis grados de separación experimento de Milgram en 1967 que encontró que las personas en los EE.UU. estuvieron en promedio separados 6 grados de cualquier otra persona (véase la introducción de mi tesis para una historia

más completa y crítica ). En segundo lugar, las distancias más cortas se han utilizado como una métrica subyacente en una serie de medidas, tales como los de centralidad (Freeman, 1978; Opsahl et al., 2010 ). Una de estas medidas es de intermediación, que es aproximadamente igual al número de trayectorias más cortas entre otros, que pasan a través de un nodo. Esta página está destinada a mostrar una generalización del cálculo del camino más corto para las redes de ponderación que tiene el potencial de ser más preciso.

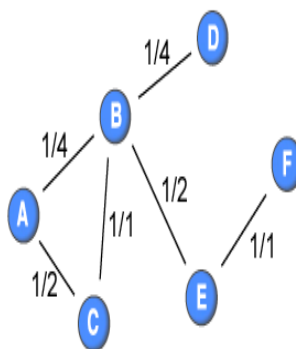
La distancia más corta entre los nodos de una red es bastante fácil de calcular si sólo tiene lazos presentes o ausentes : sólo tiene que contar los vínculos a lo largo de la ruta más corta . Si dos nodos están conectados directamente: La distancia es igual a 1 ; y si no están directamente conectados, pero están conectados a través de intermediarios , entonces es el menor número de nodos intermedios es mayor que 1 . Por ejemplo, si los lazos de esta red no tienen pesos, la distancia entre el nodo A y F sería 3. A continuación se muestra una matriz que muestra las distancias más cortas entre los nodos de la red de ejemplo:



La distancia más corta en esta red es 1,8 (la media de la matriz, excluyendo los elementos de la diagonal).

La dificultad se produce cuando los lazos se diferencian, como lo son en una red ponderada. Por ejemplo, el lazo entre el nodo A y B tiene el doble de la fuerza de la unión entre el nodo A y el nodo C. Esto podría significar que el nodo A tiene un contacto más frecuente con el nodo B que con el nodo C. A su vez, esto podría implicar que el nodo A podría dar al nodo B una pieza de información (o una enfermedad) doble de probabilidades que el nodo C. Si estamos ante la difusión de información o enfermedades en una red, entonces la velocidad que viaja, y las rutas que toma, son claramente afectadas por los pesos.

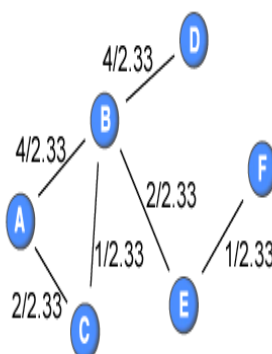
Dijkstra (1959) propuso un algoritmo que suma el costo de las conexiones y encuentra el camino de menor resistencia. Por ejemplo, los dispositivos GPS se utiliza este algoritmo mediante la asignación de un tiempo de costo para cada tramo de la carretera, y luego encontrar la ruta que cuestan menos en términos de tiempo. Este algoritmo también se puede utilizar en análisis de redes sociales. Newman (2001) aplicó a una red de colaboración de científicos al invertir los pesos de unión (1 dividido por el peso). Esto implica que un lazo más fuerte consigue un costo más bajo que un lazo más débil. A continuación se muestra la red desde arriba con los pesos invertidos:



Aplicando el algoritmo de Dijkstra, nos encontramos con que la conexión directa entre el nodo C y el nodo B tiene un costo de 1, mientras que la conexión indirecta a través del nodo A tiene un coste de 0,75 ( $1/2 + 1/4$ ). Por lo tanto, de acuerdo con este algoritmo, la información será viajar más rápido a través de la conexión indirecta. A continuación se muestra una matriz que muestra el coste de las rutas menos costosas entre los nodos de la red de ejemplo

| 1 |   | A    | B    | C    | D    | E    | F    |
|---|---|------|------|------|------|------|------|
| 2 | A | NA   | 0.25 | 0.50 | 0.50 | 0.75 | 1.75 |
| 3 | B | 0.25 | NA   | 0.75 | 0.25 | 0.50 | 1.50 |
| 4 | C | 0.50 | 0.75 | NA   | 1.00 | 1.25 | 2.25 |
| 5 | D | 0.50 | 0.25 | 1.00 | NA   | 0.75 | 1.75 |
| 6 | E | 0.75 | 0.50 | 1.25 | 0.75 | NA   | 1.00 |
| 7 | F | 1.75 | 1.50 | 2.25 | 1.75 | 1.00 | NA   |

El promedio de esta matriz es 0,9833. Esta aplicación del algoritmo de Dijkstra es útil si se quiere identificar los caminos más cortos; Sin embargo, ¿qué significa una distancia media de 0,98? Con este fin, Opsahl sugiere "normalizar" los pesos por el peso medio en la red. Los pesos serían entonces (antes de ser invertida):



Obteniendosé la siguiente matriz

| 1 |   | A      | B      | C      | D      | E      | F      |
|---|---|--------|--------|--------|--------|--------|--------|
| 2 | A | NA     | 0.5825 | 1.1650 | 1.1650 | 1.7475 | 4.0775 |
| 3 | B | 0.5825 | NA     | 1.7475 | 0.5825 | 1.1650 | 3.4950 |
| 4 | C | 1.1650 | 1.7475 | NA     | 2.3300 | 2.9125 | 5.2425 |
| 5 | D | 1.1650 | 0.5825 | 2.3300 | NA     | 1.7475 | 4.0775 |
| 6 | E | 1.7475 | 1.1650 | 2.9125 | 1.7475 | NA     | 2.3300 |
| 7 | F | 4.0775 | 3.4950 | 5.2425 | 4.0775 | 2.3300 | NA     |

El promedio de esta matriz es de 2.29 . esta unidad de distancia a continuación se refiere a un paso con el peso medio de la red. El promedio de la matriz sugiere que en los nodos promedio son 2,29 pasos con un peso promedio de distancia unos de otros(nodos). Esta medida sería comparable a través de redes con diferentes rangos de pesos en los lazos.

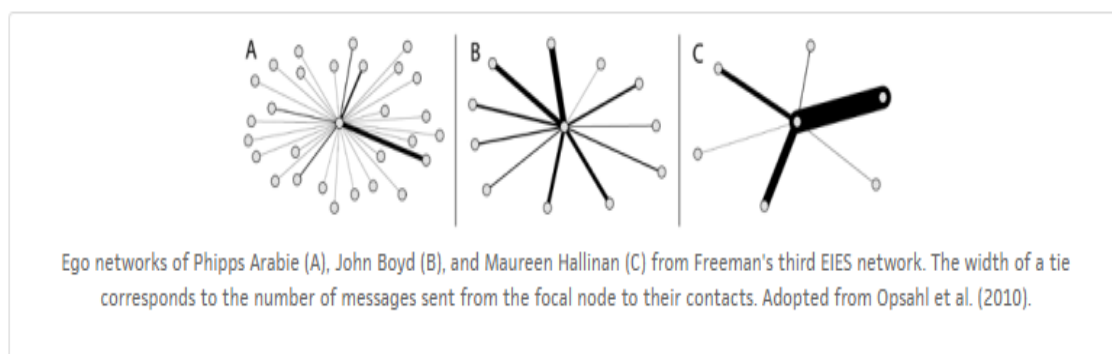
#### 4.2.3. Centralidad de nodo en redes ponderadas

La centralidad de los nodos, o la identificación de qué hay nodos que son más centrales” que otros, ha sido un tema clave en el análisis de redes (Freeman, 1978; Bonacich, 1987; Borgatti, 2005; Borgatti et al., 2006). Freeman (1978) argumentó que los nodos centrales fueron aquellos “en el meollo de las cosas.” puntos focales. Para ejemplificar su idea, él utilizó una red que consta de 5 nodos. El nodo del medio tiene tres ventajas sobre los otros nodos: tiene más lazos, que pueden llegar a todos los demás con mayor rapidez, y que controla el flujo entre los otros. Sobre la base de estas tres características, Freeman (1978) formalizó tres medidas diferentes de centralidad nodo: grado, cercanía, e intermediación. Grado es el número de nodos que un nodo focal está conectado a, y mide la participación del nodo en la red. Su simplicidad es una ventaja: sólo la estructura local alrededor de un nodo debe ser conocida para que pueda ser calculada (por ejemplo, al utilizar los datos de la Encuesta Social General; McPherson et al., 2001). Sin embargo, hay limitaciones: la medida no tiene en cuenta la estructura global de la red. Por ejemplo, aunque un nodo puede estar conectado a muchos otros, podría no estar en condiciones de alcanzar a otros con rapidez para acceder a los recursos, como la información o el conocimiento (Borgatti, 2005; Brass, 1984). Para capturar esta característica, proximidad central se define como la suma inversa de distancias más cortas a todos los demás nodos de un nodo focal. Una limitación principal de cercanía es la falta de aplicabilidad a las redes con componentes desconectados (véase la cercanía centralidad en redes con componentes desconectado). La última de las tres medidas, intermediación, evaluar el grado en el que un nodo se encuentra en el camino más corto entre otros dos nodos, y son capaces de canalizar el flujo en la red. De este modo, un nodo puede ejercer control sobre el flujo. Aunque esta medida tiene la estructura global de la red en consideración y se puede aplicar a redes con componentes desconectados, no es sin limitaciones. Por ejemplo, una gran proporción de los nodos de una red por lo general no está en una trayectoria más corta entre dos nodos cualesquiera otros, y por lo tanto recibe la misma puntuación de 0.

**Degree Ponderado** El Grado es la más simple de las medidas de centralidad mediante el uso de la estructura local en torno a nodos solamente. En una red binaria, el grado es el número de lazos que tiene un nodo. En una red dirigida, un nodo puede tener un número diferente de los lazos entrantes y salientes, y por lo tanto, el grado se divide en a cabo grados y, en grados, respectivamente.

El Grado en general, se ha extendido a la suma de pesos en el análisis de redes ponderadas (Barrat y col., 2004, Newman, 2004;. Opsahl et al, 2008), y marca la fuerza del nodo. Es

igual a la definición tradicional de grado si la red es binario (es decir, cada lazo tiene un peso de 1). Por el contrario, en las redes ponderadas, los resultados de estas dos medidas son diferentes. Dado que la fuerza nodo tiene en cuenta el peso de los lazos, esta ha sido la medida preferida para el análisis de redes ponderadas (por ejemplo, Barrat y col., 2004, Opsahl et al., 2008). Sin embargo, la fuerza de nodo es una medida contundente, ya que sólo tiene en cuenta el nivel total de un nodo de la participación en la red, y no tienen en cuenta la característica principal de las medidas originales formalizados por Freeman (1978): el número de lazos. Esta limitación se destacó por la centralidad de grado por las tres redes de ego tercera red EIES de Freeman. Los tres nodos han enviado más o menos la misma cantidad de mensajes; sin embargo, a un número bastante diferente de los demás. Si se aplica (1978) medida original de Freeman, la puntuación del nodo central en el panel A es casi cinco veces mayor que el nodo en el panel de attains C. Sin embargo, cuando se utiliza la generalización Barrat et al., que reciben más o menos la misma puntuación.



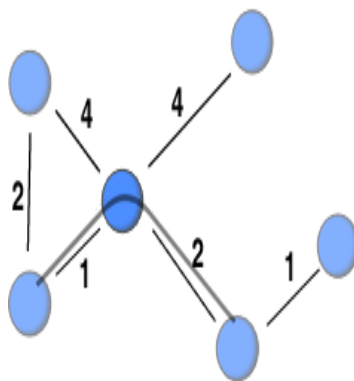
En un intento de combinar tanto grado y la fuerza, Opsahl et al. (2010) utilizaron un parámetro de ajuste para establecer la importancia relativa del número de lazos en comparación con los pesos de unión. En concreto, la medida en grados centralidad propuesta era el producto del número de nodos que un nodo central está conectado a, y el peso medio de estos nodos, ajustada por el parámetro de ajuste. Hay dos valores de referencia para el parámetro de ajuste (0 y 1), y si el parámetro se establece en cualquiera de estos valores, se reproducen las medidas existentes (Barrat y col., 2004, Freeman, 1978). Si el parámetro se establece en el valor de referencia de 0, los resultados de las medidas se basan únicamente en el número de lazos, y son iguales a la encontrada en la aplicación (1978) medida de Freeman a una versión binaria de una red en todo el vínculos con un peso superior a 0 se establecen en el presente. De este modo, los pesos de unión son completamente ignoradas. A la inversa, si el valor del parámetro es 1, la medida se basa en los pesos de unión solamente, y son idénticas a la generalización ya propuesto (Barrat et al., 2004). Esto implica que el número de lazos se tiene en cuenta. La siguiente tabla muestra las diferencias entre los grados-medidas.

| Node                 | Degree measure from |                      |                                 |                                 |
|----------------------|---------------------|----------------------|---------------------------------|---------------------------------|
|                      | Freeman (1978)      | Barrat et al. (2004) | Opsahl et al. (2010; alpha=0.5) | Opsahl et al. (2010; alpha=1.5) |
| Phipps Arabie (A)    | 28                  | 155                  | 66                              | 365                             |
| John Boyd (B)        | 11                  | 188                  | 45                              | 777                             |
| Maureen Hallinan (C) | 6                   | 227                  | 37                              | 1396                            |

**Cercanía(Closeness)** Cercanía (Closeness) La Cercanía se define como la inversa de la lejanía , que a su vez , es la suma de las distancias a todos los otros nodos ( Freeman , 1978 ) . La intención detrás de esta medida fue identificar los nodos que podrían alcanzar otras rápidamente. Una limitación principal de la cercanía es la falta de aplicabilidad a las redes con componentes desconectados: dos nodos que pertenecen a los diferentes componentes no tienen una distancia finita entre ellos. Por lo tanto, la cercanía se restringe generalmente a los nodos dentro de la mayor componente de una red . La cercanía Centralidad entrada de blog en redes con desconectado Componentes sugiere un método para superar esta limitación. La cercanía se ha generalizado a las redes ponderados por Newman (2001) , que utiliza (1959) el algoritmo de Dijkstra (ver rutas más cortas en las Redes tabuladas para más detalles) . Para reiterar rápidamente (1959) Dijkstra y Newman (2001) trabajan aquí:

Dijkstra (1959) propuso un algoritmo para encontrar los caminos más cortos en una red donde los pesos podrían considerarse costos. La ruta menos costosa que conecta dos nodos era el camino más corto entre ellos (por ejemplo, una red de carreteras en las que cada tramo de carretera tiene un tiempo de costo asignar a la misma). Newman (2001) transforma los pesos positivos en una red de colaboración en los costos de inversión de ellos (1 dividiendo por el peso). Sobre la base de los pesos invertidos, Newman (2001) aplica el algoritmo de Dijkstra y encontró los caminos menos costosos entre todos los nodos. El coste total de los caminos desde un nodo a todos los demás era una medida de la lejanía: cuanto mayor sea el número, más cuesta un nodo para llegar a todos los demás nodos. Para crear una medida de la proximidad, Newman (2001), seguido Freeman (1978) y se invierte los números (1 dividido por la lejanía). Por lo tanto, un alto lejanía se transformó en una cercanía baja y una baja lejanía se transformó en una gran cercanía. Similarmente a Barrat y col. (2004) generalización de grado, (2001) algoritmo generalizado de Newman se centra únicamente en la suma de pesos de unión, y no tiene en cuenta el número de lazos en los caminos. Opsahl et al. (2010) generalización de trayectorias más cortas se puede aplicar a la determinación de la longitud de ellos.

**Intermediación(Betweenness)** Significa que tanto un nodo es parte de las transacciones entre los otros nodos. Se puede estudiar usando medida (1978) intermediación de Freeman otros nodos.



Brandes (2001) propuso un nuevo algoritmo de cálculo de intermediación más rápido. Además de reducir el tiempo, este algoritmo también se relajó el supuesto de que los lazos tenían que estar presente o ausente (es decir, una red binaria), y de intermediación permite que se calcula en redes ponderada (observar que esta generalización es independiente de la medida de flujo propuesto por Freeman et al., 1991, que podría ser más apropiado en ciertas configuraciones). Esta generalización tiene en cuenta, que en las redes ponderados, la transacción entre dos nodos podría ser más rápida a lo largo de las trayectorias con más nodos intermedios que están fuertemente conectados a las rutas con menos nodos intermedios conectados débilmente. Esto se debe al hecho de que los nodos intermedios fuertemente conectados tienen, por ejemplo, el contacto más frecuente que los débilmente conectados queridos. Por ejemplo, el lazo entre el nodo de la parte superior izquierda y el nodo central en la red de ejemplo anterior tiene cuatro veces la fuerza de la unión entre el nodo inferior izquierda y el nodo focal. Esto podría significar que el nodo superior izquierdo tiene un contacto más frecuente con el nodo central que el nodo inferior izquierda tiene. A su vez, esto podría implicar que el nodo superior izquierda podría dar el nodo central de una pieza de información (o una enfermedad) cuatro veces más rápido que el nodo inferior izquierda. Si estamos estudiando los nodos que tienen más probabilidades de ser la canalización de información o enfermedades en una red, entonces la velocidad a la que viaja, y las rutas que toma, están claramente afectadas por los pesos. La identificación de los caminos más cortos en las redes tabuladas también se puede utilizar en la identificación de los nodos que canalizan las transacciones entre los demás nodos de redes ponderados. Si asumimos que las transacciones en una red ponderada siguen los caminos más cortos identificados por el algoritmo de Dijkstra en lugar de la que tiene el menor número de nodos intermedios, entonces el número de caminos más cortos que pasan a través de un nodo puede cambiar.

| Node | Betweenness measure from |                |                                     |
|------|--------------------------|----------------|-------------------------------------|
|      | Freeman (1978)           | Brandes (2001) | Opsahl et al. (2010; $\alpha=0.5$ ) |
| 1    | 0                        | 4              | 0                                   |
| 2    | 8                        | 8              | 8                                   |
| 3    | 0                        | 0              | 0                                   |
| 4    | 0                        | 0              | 0                                   |
| 5    | 4                        | 4              | 4                                   |
| 6    | 0                        | 0              | 0                                   |

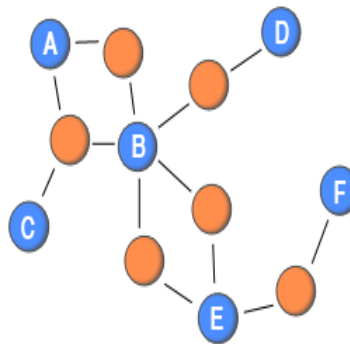
Ahora, el nodo 1 (A) ha conseguido la puntuación de intermediación 4 también. Esto se debe a la trayectoria indirecta desde el nodo B al nodo C por A se utiliza en lugar de la conexión directa. Similarmente al (2001) generalización de Newman de cercanía, de Brandes

(2001) algoritmo generalizado se centra únicamente en la suma de pesos de unión, y no tiene en cuenta el número de lazos en los caminos. Opsahl et al. (2010) generalización de los caminos más cortos también puede aplicarse a la identificación de ellos.

### 4.3. Teoría no clásica Parte 2 :Redes Ponderadas de dos modos

Uno de los grandes aportes de Tore(2009)<sup>18</sup> fue el estudio de las redes ponderadas de dos modos o bipartitas.

Muchos conjuntos de datos de la red pertenecen a las redes de dos modos de definición (también conocidas como redes de afiliación o bipartitas). Estos son un tipo particular de redes con dos conjuntos de nodos y lazos que se establecen entre los nodos que pertenecen a diferentes conjuntos. Este diagrama ilustra una red de dos modos binario en el que el color representa el conjunto de nodos al que pertenece un nodo. Uno de los primeros conjuntos de datos de dos modos para ser analizados fue el conjunto de datos Davis Sur de las mujeres del club (Davis et al., 1941), que registró la asistencia de un grupo de mujeres (conjunto de nodos 1) a una serie de eventos (serie 2). Una mujer estaría vinculada a un evento si asistió a ella. Otro tipo de conjunto de datos de dos modos que se ha popularizado en los últimos años son las redes de colaboración científica. En este tipo de redes, se establece un lazo entre un científico (nodo set 1) y un documento si el científico autor de ese documento (por ejemplo, Newman, 2001) (2 conjuntos de nodos).



Los vínculos en redes de dos modos también pueden ser diferenciados por pesos. En estos casos, la red sería una red de dos modos ponderado. Un ejemplo de este tipo de red son foros en línea donde los usuarios (conjunto de nodos 1) puede crear múltiples mensajes en temas o hilos (conjunto de nodos 2). De manera similar a las redes de un modo ponderado, se les pueden incorporar medidas en los lazos. Esta sección se ocupa principalmente de las redes de dos modos binarios, pero cuando se aplique, la documentación incluirá también información sobre las redes de dos modos ponderados. En la medida que los científicos en general, deciden si les gustaría trabajar en un papel, a menudo se asume que los nodos son primarios. Sin embargo, no siempre es obvio que el conjunto de nodos sea el principal, y en estos casos, la pregunta de investigación guía la elección. Por ejemplo, en el caso de los directorios entrelazados, donde los dos conjuntos de nodos son directores y juntas corporativas, y los lazos representan la afiliación de los directores con tablas, no siempre está claro si los directores o juntas son el conjunto de nodos primarios (por ejemplo, Levine, 1979; Mizruchi, 1996; Seierstad y Opsahl, 2011). Es probable que sea debido a la formación de un proceso mutuo, donde (1) se invitará a los directores para unirse a la junta, y (2) acepten la invitación.

<sup>18</sup>Tore Opsahl. Structure and Evolution of Weighted Networks. University of London (Queen Mary College), London, UK. Thesis 2009, 104-122.

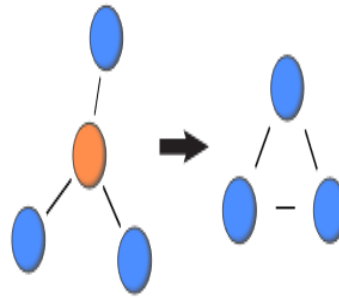
#### 4.3.1. ¿Cómo Analizar las redes de dos modos?

Las redes de dos modos rara vez se analizan sin transformarlas. Esto se debe a que la mayoría de las medidas de red se definen únicamente para las redes de un modo, y sólo unos pocos de ellos se han redefinido para las redes de dos modos (Borgatti y Everett, 1997; Lapaty et al., 2008). La transformación de una red de dos modos para una red de un solo modo se hace a menudo con un método conocido como proyección. Este método funciona mediante la selección de uno de los dos conjuntos de nodos (a menudo el conjunto de nodos primario) y la vinculación de los nodos a uno establecido si estuvieran conectados al menos a un nodo común en el otro conjunto. Aunque la estructura de dos modos se descarta en este proceso, es posible definir los pesos de unión basados en ella. Específicamente, los pesos del nodo se definen a menudo como el número de nodos comunes. Este método se extendió por Newman (2001), quienes argumentaron que los pesos de unión entre los autores en redes de colaboración científica deben ser descontados si los autores colaboraron en papeles con muchos otros. Para obtener más información, consulte la página en la proyección. La proyección de las redes de dos modos crea una serie de problemas. En primer lugar, cada lazo en una red de un solo modo prototípico se supone que es creado por separado; sin embargo, este no es el caso en las redes de dos modos proyectadas. Por ejemplo, mientras que una llamada telefónica estándar crea un lazo de comunicación de una persona a otra, un director forma lazos con todos los otros directores en un tablero cuando él o ella se une a esa junta. Esto tiene implicaciones directas para los marcos que utilizan redes aleatorias para detectar un nivel de referencia (por ejemplo, Opsahl et al., 2008) y al comparar las medidas observadas en una red con las que se encuentran en las redes aleatorias correspondientes. Esto es debido al hecho de que los lazos en las redes aleatorias clásicas se supone que son independientes entre sí (Erdos y Rényi, 1959). Aunque este no es el caso en un modo prototípico ni proyecta redes de dos modos, las redes aleatorias son menos comparables a las redes de dos modos proyectados de las redes de un modo prototípicos como múltiples lazos pueden ser creados debido a un solo evento en estas redes. En segundo lugar, dependiendo del grado de distribución del conjunto no proyectado nodo, una red de dos modos proyectada tiende a tener más y más grandes grupos conectados completamente que las redes de un modo prototípicos (Wasserman y Faust, 1994). Estos se producen cuando tres o más nodos están conectados a un nodo común de la red de dos modos (por ejemplo, todos los directores en una sola tarjeta están conectados y forman un clique totalmente conectado). Esta característica afecta a una serie de medidas de red, especialmente los basados en triángulos incluyendo las medidas estructurales huecos (Burt, 1992, 2005) y los coeficientes de la agrupación (para una revisión, véase Opsahl y Panzarasa, 2009). Para ejemplificar las grupos, y los muchos triángulos, produce cuando se proyecta una red de dos modos, la siguiente figura muestra el componente principal de la red interpersonal entre los directores noruegos.

Las redes ponderadas de dos modos: Los vínculos en redes de dos modos a menudo pueden ser diferenciados sobre la base de calcular la fuerza de interacción. En estos casos, la red sería una red de dos modos ponderado. Al igual que con las redes de un modo ponderado, las ponderaciones de acoplamiento deben estar basadas en una escala de razón (es decir, 0 es igual a nada, y un valor de 4 es el doble de 2). Para obtener más información, consulte la página de Definición de Redes ponderado. Un ejemplo de este tipo de red es foros en línea donde los usuarios (conjunto de nodos: 1) puede crear múltiples mensajes en temas o hilos (conjunto de nodos 2). Entre otras medidas, hay dos maneras de hacer operativos los pesos de unión en estas redes. En primer lugar, que podría ser el número de mensajes de un mensaje de usuario a un hilo. De esta manera la herida contar el número de interacción en un hilo, pero sea insensible al tamaño de los mensajes. Como tal, una segunda forma de definir los pesos de unión como la suma del número de caracteres de los mensajes enviados. Mientras se incluye el tamaño de los mensajes, este método no incorpora la parte de atrás adelante y hacia atrás de los contenidos. Sin embargo, puede incluir pesos empate, el análisis de un foro en línea podría ser más preciso que simplemente ignorarlos.

### 4.3.2. Proyección

Las Redes de dos modos rara vez se analizan en su forma original. Aunque esto es preferible, existen algunos métodos para tal fin. Como tal, estas redes a menudo se transforman en las redes de un solo modo (un solo tipo de nodos) para ser analizadas. Este procedimiento se refiere a menudo como la proyección. La proyección se realiza mediante la selección de uno de los conjuntos de nodos y la vinculación de dos nodos de la establecida si estuvieran conectados al mismo nodo (de la otra especie). Este proceso se ilustra para los nodos azules de este diagrama.

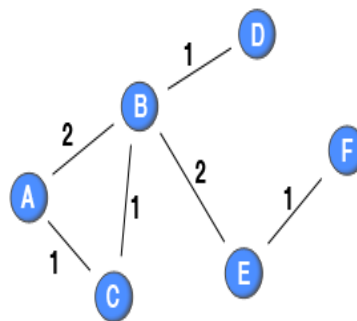


### 4.3.3. Redes binarias de dos modos

Tradicionalmente, los lazos en las redes de un modo proyectados no tienen pesos unidos a ellos. Sin embargo, los estudios empíricos recientes de redes de dos modos, ha creado una red de un modo ponderado mediante la definición de los pesos como el número de co-ocurrencias (por ejemplo, el número de eventos que a lo que han-asistido dos individuos o el número de artículos en el que dos autores han colaborado. Para describir formalmente este método y facilitar la comparación entre los diferentes métodos introducidos, este método se puede formalizar como:

$$w_{ij} = \sum_p 1$$

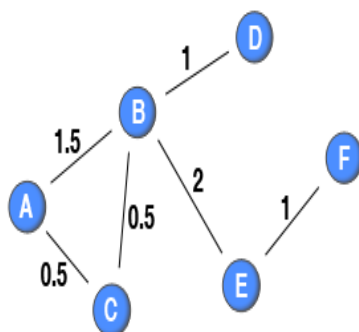
en donde  $w_{ij}$  es el peso entre el nodo  $i$  y el nodo  $j$ , y  $p$  es los nodos de la otra clase que el nodo  $i$  y el nodo  $j$  están conectados a, sus co-ocurrencias (por ejemplo, los nodos de color rojo en el diagrama anterior). En la red de la muestra, el vínculo entre el nodo A y el nodo B tiene un peso de 2, ya que estos dos nodos tienen conexiones a dos nodos rojos comunes, mientras que la unión entre el nodo A y el nodo C tiene solamente un peso de 1, ya que estos nodos tienen conexiones a sólo un nodo rojo común.



Newman (2001) extendió este procedimiento mientras trabajaba con redes de colaboración científica. Sostuvo que los lazos sociales entre los científicos que colaboran con algunos otros en un papel eran más fuertes que los lazos entre los científicos que colaboran con muchos en un papel. El propuso descontar para el tamaño de la colaboración mediante la definición de los pesos entre los nodos utilizando la siguiente fórmula:

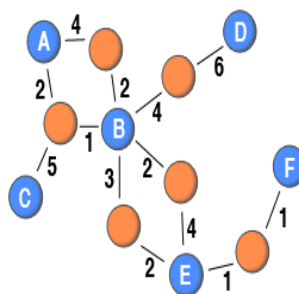
$$w_{ij} = \sum_p \frac{1}{N_p - 1}$$

donde  $N_p$  es el número de autores en el papel  $p$  (por ejemplo, el número de nodos conectados azul al nodo rojo  $p$ ). En el contexto de redes de colaboración científica, esto implica que si dos científicos que escriben solamente un solo papel junto con ningún otro co-autor (es) obtienen un peso de 1 (por ejemplo, el nodo B y el nodo D). Por otra parte, si dos científicos han escrito dos artículos juntos sin ningún co-autor, el peso de su eliminatória sería 2 (por ejemplo, el nodo B y E). Sin embargo, si los dos científicos tienen un co-autor, el peso sobre el vínculo entre ellos es de 0,5 (por ejemplo, el nodo A y el nodo C). Un ejemplo más complicado es el lazo entre el nodo A y el nodo B en el diagrama. Ellos han escrito dos artículos juntos: uno sin ningún otro co-autor y uno con el nodo C como un co-autor. El primer documento daría su eliminatória un peso de 1, y el segundo lazo añadiría 0,5 al peso de este lazo. Por lo tanto, el peso atado a su lazo es de 1.5. Descontando para el número de nodos azules conectados al mismo nodo de red, estos métodos crea una proyección de un modo en el que la fuerza de un nodo es igual al número de los lazos que procedan de dicho nodo de la red de dos modos (por ejemplo la suma de pesos unido a lazos originarios desde el nodo a en la proyección de un modo es 2, y el nodo a está conectado a dos nodos rojos)



#### 4.3.4. Redes Ponderadas de dos modos

Hasta ahora, los métodos sólo se han ocupado de las redes de dos modos binarios. Sin embargo, también existen redes de dos modos ponderados. Por ejemplo, en los foros en línea, donde los conjuntos de nodos son los usuarios y los hilos (o temas), y un empate entre dos nodos se establece si un usuario publica un mensaje a un hilo, es posible diferenciar los lazos de dos modos basados en el número de mensajes o caracteres enviados por usuarios a un hilo específico. Otro ejemplo, es la compra de un comportamiento en el que el conjunto de nodos son personas y mercancías, y el peso se puede definir en términos de número de compras. Similarmente a las redes de un modo ponderado, los datos utilizados para el análisis son más ricos si los pesos de los nodos se incluyen en la red de dos modos que si se descartan. Una red de dos modos ponderado se ilustra en el diagrama mostrado abajo.



En un espíritu similar se puede recurrir simplemente al número de co-ocurrencia de una red de dos modos binario, la proyección de un solo modo de una red de dos modos ponderada podría estar basado en los pesos de los dos nodos se han dirigido hacia los nodos comunes (de tipo diferente). Además, con un método de este tipo es posible diferenciar cómo los dos nodos interactúan con el nodo común, y para proyectarla de una manera dirigida sobre una red de un solo modo ponderado. En este tipo de red proyectada, el peso de un lazo de un nodo a otro no es necesariamente el mismo que el peso unido a la atadura del último nodo al antiguo nodo. No es una verdadera red dirigida como existen dos lazos dirigidos entre cualquier par de nodos conectados. Más específicamente, todas las díadas se componen de dos lazos dirigidos (de inversión) o sin lazos dirigidos (null), y no hay díadas están hechos de un solo lazo dirigido (asimétrica). Por ejemplo, si el diagrama anterior se refirió a un foro en línea, el nodo B ha publicado 4 mensajes a un hilo que el nodo D participar en, dando así el lazo dirigido entre el nodo B y el nodo D en la proyección de un modo un peso de 4 . Por el contrario, el nodo D publicado 6 mensajes a ese hilo, y por lo tanto, el peso unido al empate desde el nodo D al nodo B es igual a 6. Este método se puede formalizar la siguiente manera:

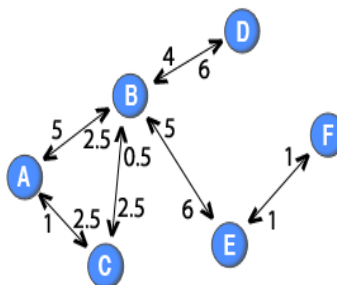
$$w_{ij} = \sum_p w_{i,p}$$

En un espíritu similar al método utilizado por Newman (2001), también es posible descontar el número de nodos cuando se proyectan las redes de dos modos ponderados. Por ejemplo, se podría argumentar que, si muchos usuarios en línea alrededor de un tema, sus lazos deben ser más débiles que si hay pocas personas que publican en la red.

Una generalización recta mas simplificada es la siguiente función.

$$w_{ij} = \sum_p \frac{w_{i,p}}{N_p - 1}$$

Esta fórmula crea una red de un solo modo dirigido en la que el outdegree de un nodo es igual a la suma de los pesos unidos a los lazos en la red de dos modos que se originó a partir de dicho nodo. Por ejemplo, el nodo C tiene un lazo con un peso de 5 en la red de dos modos y un fuera de la fuerza de 5 en la proyección de un modo



#### 4.3.5. Caminos más cortos

Los caminos más cortos entre los nodos ha sido durante mucho tiempo un elemento clave de la investigación de redes. Mientras que los caminos más cortos a menudo no son de interés en sí mismos, son el componente clave de una serie de medidas. En primer lugar, una pregunta popular ha sido cuál es la distancia media entre la gente? La pregunta se hizo famoso después de los seis grados de separación experimento de Milgram en 1967 que encontró que las personas en los EE.UU. fueron en promedio de 6 pasos de cualquier otra persona (véase la introducción de mi tesis para una historia más completa y crítica) . En segundo lugar, las rutas más cortas se han utilizado como una métrica subyacente en una serie de medidas, tales como los de centralidad ( Freeman , 1978 ) . Una de estas medidas es de intermediación, que es aproximadamente igual al número de trayectorias más cortas entre otros, que pasan a través de un nodo

La distancia más corta entre los nodos de las redes de un modo binario es una recta hacia adelante para calcular: usted sólo tiene que contar los vínculos a lo largo de la ruta más corta . Si dos nodos están conectados directamente: distancia = 1 ; y si no están directamente conectados, pero están conectados a través de intermediarios , entonces es el menor número de nodos intermedios +1 . Para obtener más información, así como la manera de identificar los caminos más cortos en las redes ponderados y calcular su longitud, ver rutas más cortas en las Redes tabuladas. Para ejemplificar cómo se calculan las distancias más cortas, la distancia entre el nodo A y F en la red de la muestra a la derecha sería 3 ya que este es el número mínimo de lazos entre ellos. Las distancias entre todos los pares de nodos se muestran en la siguiente matriz . La distancia más corta media (o distancias de la red) se calcula tomando la media de la matriz de exclusión de la diagonal. En estos casos, es 1,8

|   |     |     |     |     |     |     |     |
|---|-----|-----|-----|-----|-----|-----|-----|
| 1 |     | [A] | [B] | [C] | [D] | [E] | [F] |
| 2 | [A] | NA  | 1   | 1   | 2   | 2   | 3   |
| 3 | [B] | 1   | NA  | 1   | 1   | 1   | 2   |
| 4 | [C] | 1   | 1   | NA  | 2   | 2   | 3   |
| 5 | [D] | 2   | 1   | 2   | NA  | 2   | 3   |
| 6 | [E] | 2   | 1   | 2   | 2   | NA  | 1   |
| 7 | [F] | 3   | 2   | 3   | 3   | 1   | NA  |

En las redes de dos modos, las rutas más cortas se han identificado en un número de maneras ( Borgatti y Everett, 1997 ; Latapy et al, 2008 , Newman, 2001 ) . Mientras que los métodos varían, por lo general ellos conducen sustancialmente al mismo resultado. La distancia entre dos nodos es igual al número de vínculos entre ellos. En primer lugar, Borgatti y Everett (1997) sugirieron para representar las redes de dos modos de  $(n + p) \times (n + p)$  -matrices donde n y p es el número de nodos primario y secundario , respectivamente, en lugar de dos en modo rectangular matrices ( NXP ) . Este tipo de matriz es extremadamente ineficiente, ya que es al menos cuatro veces más grande que el rectangular sin mantener ninguna información adicional. Sin embargo, tiene una característica fundamental: es cuadrada. Como tal, las métricas de una red de un modo, como los de dos modos se les puede aplicar directamente la misma. Este método ha sido el abogado general para los usuarios de UCINET. La red de ejemplo de dos modos está representada en una matriz tal a continuación

|    | [A] | [B] | [C] | [D] | [E] | [F] | [1] | [2] | [3] | [4] | [5] | [6] |
|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 1  | [A] | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 1   | 0   | 0   | 0   |
| 2  | [B] | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 1   | 1   | 1   | 0   |
| 3  | [C] | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   |
| 4  | [D] | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0   |
| 5  | [E] | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 1   |
| 6  | [F] | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   |
| 7  | [1] | 1   | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   |
| 8  | [2] | 1   | 1   | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   |
| 9  | [3] | 0   | 1   | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 0   |
| 10 | [4] | 0   | 1   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 0   |
| 11 | [5] | 0   | 1   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 0   |
| 12 | [6] | 0   | 0   | 0   | 0   | 1   | 1   | 0   | 0   | 0   | 0   | 0   |

Un segundo método consiste en proyectar las redes de dos modos a las redes de un modo ( $n \times n$  - matriz), y luego aplicando el modo más corto algoritmo de ruta estándar. La proyección de trabajos de conexión de nodos primarios una vez conectados a los nodos comunes en la estructura de dos modos permite ejemplificar esta estructura. La matriz siguiente es la proyección binario de la red de dos modos:

|   | [A] | [B] | [C] | [D] | [E] | [F] |
|---|-----|-----|-----|-----|-----|-----|
| 1 | [A] | 0   | 1   | 1   | 0   | 0   |
| 2 | [B] | 1   | 0   | 1   | 1   | 1   |
| 3 | [C] | 1   | 1   | 0   | 0   | 0   |
| 4 | [D] | 0   | 1   | 0   | 0   | 0   |
| 5 | [E] | 0   | 1   | 0   | 0   | 0   |
| 6 | [F] | 0   | 1   | 0   | 0   | 1   |
| 7 | [F] | 0   | 0   | 0   | 0   | 1   |

El uso de los dos procedimientos anteriores produce un resultado similar. La única diferencia es que todas las distancias que se encuentran con el primer método son el doble de las distancias encontradas usando el segundo método. Por ejemplo, las distancias que se encuentran desde el nodo B al nodo F son 4 y 2 mediante los dos métodos, respectivamente. Hay una serie de problemas con estos dos métodos:

- Dos nodos conectados a través de dos nodos comunes (por ejemplo, B y E) se supone que tienen una conexión idéntica a dos conectados a través de un único nodo común (por ejemplo, B y D).
- Todos los nodos que no son proyectados se consideran iguales. Esto implica que los nodos de conexión muchos nodos son iguales a los que enlace solamente dos nodos. Newman (2001) argumentó que los nodos secundarios que conectan unos nodos primarios dieron esos nodos de un nivel más alto de interacción.
- No hay manera establecida de cálculo de rutas más cortas en la red de dos modos ponderado.

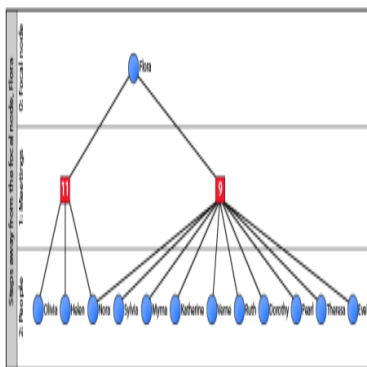
Basándose en el segundo método, la proyección, es posible hacer frente a algunas de estas limitaciones. Mediante la proyección de las redes de dos modos (tanto binarios y ponderados) en las redes de un modo ponderado, y luego aplicar el algoritmo de ruta más corta ponderada, este procedimiento tiene la posibilidad de capturar más información que simplemente decir que los pesos no importan. Es importante utilizar un método apropiado que se proyecte ya que influirá en el resultado de la medida. Para ejemplificar este método, las distancias en la red del ejemplo de dos modos utilizando el método de la suma proyección serían:

|   |     |      |      |      |      |      |      |
|---|-----|------|------|------|------|------|------|
| 1 |     | [A]  | [B]  | [C]  | [D]  | [E]  | [F]  |
| 2 | [A] | NA   | 0.67 | 1.33 | 2.00 | 1.33 | 2.67 |
| 3 | [B] | 0.67 | NA   | 1.33 | 1.33 | 0.67 | 2.00 |
| 4 | [C] | 1.33 | 1.33 | NA   | 2.67 | 2.00 | 3.33 |
| 5 | [D] | 2.00 | 1.33 | 2.67 | NA   | 2.00 | 3.33 |
| 6 | [E] | 1.33 | 0.67 | 2.00 | 2.00 | NA   | 1.33 |
| 7 | [F] | 2.67 | 2.00 | 3.33 | 3.33 | 1.33 | NA   |

Como se puede ver en esta matriz, la distancia entre B y E ( 0.67 ) es la mitad de la distancia de B a D ( 1,67 ) . Esto se debe a que tiene dos nodos comunes en lugar de sólo uno, y por lo tanto tener una conexión más fuerte.

#### 4.3.6. Centralidad de nodo en redes de dos modos

La centralidad de los nodos, o la identificación de qué nodos son más centrales”que otros, ha sido un tema clave en el análisis de redes (Freeman, 1978). Freeman (1978) argumentó que los nodos centrales son aquellos que están .en el meollo de las cosas.º puntos focales. Sobre la base de este concepto, se formaliza tres medidas: grado, cercanía, e intermediación. Para un análisis más a fondo sobre estas medidas, consulte Nodo Centralidad en Redes ponderado. Grado es el número de lazos que un nodo tiene o el número de otros nodos con el que un nodo está conectado. En las redes de dos modos, este concepto se puede aplicar directamente. Sin embargo, hay ligeras complicaciones. En las redes de dos modos, .el número de otros nodos con los que un nodo está conectado a.es ambigua. Bien podría ser el número de nodos secundarios de un nodo primario a los que está conectado (y viceversa), o el número de nodos primario a los que un nodo primario está conectado. Para aclarar la diferencia entre estos dos números, me he referido a ellos como el grado de dos modos y un modo de nodos ', respectivamente. Para ejemplificar la diferencia, la imagen siguiente se muestra la red local que rodea la flora en el Davis (1940) Sur de la Mujer conjunto de datos (adaptado de Opsahl, 2011).



Como puede verse a partir de este diagrama, el grado de dos modos de Flora es 2 y el grado de un modo es 12. Si la red se proyecta a una red de un modo, la medida en grados estándar sería 12. También es posible llegar al grado de dos nodos modo 'después de una red ha sido proyectada utilizando el método de Newman (2001). Este método de proyección fue desarrollado para redes de coautoría científicos, y establece el peso empate entre dos autores igual a la suma a través de los papeles co-autor de más de 1 número de autores en que el papel menos 1. En otras palabras, para cada co-autor papel, un nodo divide 1 por los otros autores.

Como tal, el peso total del lazo es igual al número de trabajos en coautoría. La única diferencia entre este método y el grado de dos modos es trabajos de un solo autor. Estos están excluidos en la primera y se incluyen en el segundo método.

#### 4.3.7. Cercanía e intermediación

La parte principal de las medidas de cercanía e intermediación es el camino más corto y su longitud. La proximidad es la suma inversa de longitudes de las rutas más cortas, e intermediación es el número de caminos más cortos que pasan a través de un nodo. Al tomar ventaja del algoritmo del camino más corto de dos modos, es posible extender fácilmente estas dos medidas a las redes de dos modos. Para recapitular rápidamente este algoritmo.

Utilizando un método de proyección adecuada para la identificación de las rutas más cortas, y calcular su longitud, en las redes de un modo ponderado (Brandes, 2001; Dijkstra, 1959; Newman, 2001) permite encontrar la longitud de los caminos más cortos, la medida de cercanía sería simplemente la suma inversa de ellos. Del mismo modo, intermediación sería fácil de calcular si se mira a los nodos intermedios en las rutas más cortas, y contar, para cada nodo, el número de veces que el nodo es un intermediario. Nota: si hay varios caminos más cortos, es importante dividir por el número de ellos para asegurarse de que cada camino sólo cuenta una vez. Ejemplo Para ilustrar las cuatro medidas, cuento de Davis (1940) Sur de la Mujer conjunto de datos. La asistencia a las reuniones de 18 mujeres en 14 sesiones se registra en este conjunto de datos. La tabla siguiente muestra el resultado de las cuatro medidas (Newman, 2001, método de proyección se utiliza para la cercanía e intermediación como el nivel de interacción entre los participantes en los eventos de menor tamaño tienden a ser más altos).

| Nodo.     | two-mode degree | one-mode degree | closeness | betweenness |
|-----------|-----------------|-----------------|-----------|-------------|
| EVELYN    | 8               | 17              | 0.053     | 5           |
| LAURA     | 7               | 15              | 0.051     | 1           |
| THERESA   | 8               | 17              | 0.060     | 48          |
| BRENDA    | 7               | 15              | 0.050     | 1           |
| CHARLOTTE | 4               | 11              | 0.044     | 0           |
| FRANCES   | 4               | 15              | 0.041     | 0           |
| ELEANOR   | 4               | 15              | 0.043     | 0           |
| PEARL     | 3               | 16              | 0.037     | 0           |
| RUTH      | 4               | 17              | 0.043     | 0           |
| VERNE     | 4               | 17              | 0.045     | 0           |
| MYRNA     | 4               | 16              | 0.042     | 0           |
| KATHERINE | 6               | 16              | 0.052     | 0           |
| SYLVIA    | 7               | 17              | 0.054     | 11          |
| NORA      | 8               | 17              | 0.059     | 60          |
| HELEN     | 5               | 17              | 0.046     | 0           |
| DOROTHY   | 2               | 16              | 0.027     | 0           |
| OLIVIA    | 2               | 12              | 0.035     | 0           |
| FLORA     | 2               | 12              | 0.035     | 0           |

Una limitación clave de la medida de intermediación se puede ver en esta tabla: la mayoría de las personas a alcanzar una puntuación de 0 (es decir, la medida es cero inflado). Las parejas de correlaciones entre las medidas se indican a continuación. Mientras que todas las medidas tienen altas correlaciones, es interesante observar que el grado de dos modos de medir tiene una mayor correlación con la cercanía e intermediación que el grado medida de un modo. Esto podría sugiere que el computacionalmente barato de dos modos grado - medida es más capaz de replicar los computacionalmente costosas medidas de cercanía e intermediación.

|                    | 1    | 2    | 3    | 4    |
|--------------------|------|------|------|------|
| 1: two-mode degree | 1.00 |      |      |      |
| 2: one-mode degree | 0.51 | 1.00 |      |      |
| 3: closeness       | 0.95 | 0.44 | 1.00 |      |
| 4: betweenness     | 0.59 | 0.34 | 0.64 | 1.00 |

### 4.3.8. Clustering

Un tema que ha recibido mucho la atención tanto en la investigación teórica y empírica es la tendencia de los nodos a agruparse. La evidencia sugiere que en la mayoría de las redes del mundo real, y sobre todo en las redes sociales, los nodos del clúster en grupos esta densamente conectados (Holland y Leinhardt, 1970; Opsahl y Panzarasa, 2009). Una serie de medidas se han desarrollado para probar esta tendencia. En concreto, el coeficiente de la agrupación global que evalúa el nivel general de la agrupación en una red (Luce y Perry, 1949), y el coeficiente de la agrupación local evalúa la agrupación en la red inmediata de un solo nodo (es decir, el nodo y sus contactos; Watts y Strogatz, 1998). Formalmente, el coeficiente global es la fracción de trios semiabiertos o de dos caminos (es decir, tres nodos conectados por dos lazos) que están cerradas por la presencia de un empate entre el primer y el tercer nodo, y el coeficiente de la agrupación local es la fracción de trios de caminos centrados en el nodo focal que se cierra por la presencia de un empate entre el primero y el tercer nodo. Para una revisión, véase Opsahl y Panzarasa (2009) o la agrupación en redes ponderado páginas. La aplicación de los dos coeficientes de la agrupación directamente a una red de dos modos es sin sentido como las redes de dos modos son bipartitas y, por lo tanto, los contactos de un nodo no pueden ser conectados entre sí por la construcción y no hay triángulos pueden existir (Borgatti y Everett, 1997). Cuando las medidas de un modo no se pueden aplicar directamente a las redes de dos modos, hay dos opciones: la primera es, transformar las redes de dos modos en las redes de un modo, hasta donde la medida de un solo modo se puede utilizar, y la segunda, el uso de la especial medidas diseñadas para redes de dos modos. ¿Por qué no transformar las redes de dos modos en las redes de un modo en la medición de la agrupación? Una serie de cuestiones están presentes en la medición de la agrupación en la red de dos modos transformada. La transformación de una red de dos modos para una red de una sola modo se hace a menudo con el método conocido como proyección.

*Es importante resaltar que este metodo ya se encuentra explicado y no se aplicará ya que en los datos no existe esta estructura*

## 5. Marco Metodológico

Para lograr describir y crear criterios de interpretación de las interrelaciones entre los usuarios y las bases de datos, a través de técnicas de Social Networking Analysis(SNA),que permitan elevar el nivel de seguridad de la información de las operaciones que realizan los usuarios sobre las bases de datos del DEPÓSITO, se definieron cuatro etapas a saber:

- i) **Etapas de recolección de los datos:** Para el desarrollo de esta etapa, el DEPÓSITO, particularmente en el área de seguridad de la información, cuenta con el software DAM, el cual muestra las diferentes operaciones realizadas por los usuarios con las bases de

datos registradas por segundo. para este trabajo se estableció como fecha de corte el mes de enero del presente año.

La estructura de los datos recolectados, está definida por las siguientes variables: tiempo en segundos en los que se realizó la operación, usuario, tabla(base de datos), operación. Este tipo de estructura recibe el nombre de edgelist as cases ya que cada fila representa una operación de un usuario con una base de datos y por ende un caso.

Posteriormente, con SPSS se realizó la transformación correspondiente a la estructura de los datos descategorizandola y convirtiendola en un conjunto de matrices de adyacencia ponderadas multirelacionales, vale la pena resltar como ya esta explicado en el marco teórico que las matrices multirelaconales se generán para una única operación es decir una para Select, Delete, Insert y Update, donde cada fila representa un usuario y cada columna una tabla y los valores contenidos en esta matriz de adyacencia representan el número de interacciones que se realizan entre usuarios y bases de datos.

Se transformaron las matrices de adyacencia ponderada multirelacionales en directed weighted edge lists<sup>19</sup>, es decir listas donde se evidencia la interacción del usuario con las tablas y el número de veces que las realizó, en tres columnas distintas y durante el mismo periodo. El conjunto de relaciones directas (directed ties) que se dan entre las bases de datos (cabeza) y los usuarios (cola), categorizandolas por el tipo de operacion (Relacion multirelacional) que se daban entre elementos del conjunto. Esta accion se realizó mediante sentencias definidas en R.

Para mayor claridad, se deja a modo de ejemplo de edgelist as cases la siguiente tabla.

- Formato Original de los datos “edgelist as cases”

Tabla 1: Edge list as cases 1.

| Seg       | Usuario | Tabla | Actividad |
|-----------|---------|-------|-----------|
| segundo1  | U1      | t1    | Delete    |
| segundo 2 | U2      | t207  | Insert    |

## ii) Etapa Definición de actores

Una vez recolectados los datos, se procedió a definir el conjunto de actores que representan a los usuarios así como también el conjunto de actores que respresentan las tablas(bases de datos) posteriormente, se definió el conjunto de relaciones correspondientes.

Una vez definidos estos conjuntos, se procedió a construir el grafo que representa los dos conjuntos de actores y sus interelaciones en este caso directas(dirigidas).

## iii) Etapa Descriptiva

Elaborado el grafo, se procedió a definir cuales herramientas eran más relevantes para el análisis que exigía el presente trabajo. De la teoría clásica se escogieron con criterios de pertinencia y practicidad las siguientes herramientas:

### A' Metodología Clásica

- Outdegree
- Indegree
- Cercanía

### B' Metodología No Clásica Ponderada

- Outdegree
- Indegree
- Cercanía

Después de un análisis sobre la utilidad de las herramientas de intermediación se decidió que estas no son relevantes para el presente trabajo ya que no existen caminos dentro

<sup>19</sup><http://www.shizukalab.com/toolkits/sna/weighted-edgelist>

del grafo que involucren a más de dos actores, lo que impide que se puedan establecer comunicaciones indirectas dando como resultado 0 para la variable intermediación en todos los actores.

Dado que para el análisis actual propuesto, sólo existen tripleteas semicerradas y no existen tripleteas cerradas, tal y como se definen en el marco teórico para definir el coeficiente de agrupamiento global, resulta irrelevante calcular el coeficiente global y local de agrupamiento pues es igual a cero en todos los actores.

- iv) Etapa de formulación de criterios para guiar el Monitoreo y generación de alertas tempranas para situaciones atípicas  
 Dados los resultados de la descripción y la etapa de mediciones realizadas en las etapas anteriores, fue posible discriminar la naturaleza de las relaciones entre los usuarios con las bases de datos y se identificaron las relaciones inadecuadas. Así como la estructura de los datos. Estos resultados se notificaron al jefe inmediato quien definió los criterios de corrección pertinentes para garantizar la integridad, disponibilidad y confidencialidad de las bases de datos.

## 6. Hallazgos en el desarrollo de cada etapa del proyecto

### 6.1. Hallazgos en la etapa de recolección de los datos

Tal como se enuncio en los apartados anteriores, durante esta etapa, se recolectaron los datos del DAM correspondientes a las operaciones realizadas por todos los usuarios del mercado de valores colombiano durante el mes de enero del presente año. Estos registros alcanzaron el número de 1'027.000. Se filtraron los datos usando como criterio aquellos relacionados estrictamente con el DEPÓSITO llegando al número de 460.000. Estos a su vez se filtraron acorde con los criterios definidos por el administrador de seguridad de la información llegando al número de 11.130 relaciones.

En seguida, se encriptó la información de los usuarios y las bases de datos, y se procedió a resumirlas en matrices de adyacencia multirrelacionales guardadas en archivos csv, para después convertirlas en directed weighted edgelist multirrelacionales mediante el software R.

### 6.2. Hallazgos en la etapa de definición de actores

En este apartado, los resultados permitieron construir los grafos que tipifican el conjunto de relaciones entre los usuarios y las tablas (bases de datos). Para esta definición se usaron herramientas de la teoría clásica y no clásica especificadas en el marco teórico y en el marco metodológico del presente trabajo.

**Grafos teoría clásica** A continuación se presentan los diferentes grafos correspondientes a cada una de las relaciones múltiples de acuerdo con la metodología clásica, para identificar a los usuarios se usaron las siguientes convenciones:

Actores: Color azul claro Tablas: Color Naranja

**Grafos teoría no clásica** Para complementar la construcción que ofrecía la teoría clásica del grafo se recurrió a construir los grafos como redes ponderadas, tomando en cuenta los siguientes intervalos, donde  $r$  es la intensidad de la relación:

- si  $r < 2$  se le asignará el color amarillo y se considerará una relación de baja intensidad (situación atípica)
- si  $2 \leq r \leq \bar{r}$  se le asignará el color naranja y se considerará una relación de intensidad media (situación no atípica)

- si  $r > \bar{r}$  se le asignará el color azul y se considerará una relación de alta intensidad (situación atípica)

Adicionalmente, se utilizaron las siguientes convenciones:

Actores: Color rojo actores Tablas: Color verde

#### 6.2.1. Grafo Select

En el siguiente grafo, se evidencia comparativamente, diferencias notables entre la metodología clásica y la metodología ponderada, en la primera se observa las relaciones que bajo concepto de la dirección de seguridad de la información son atípicas. En el grafo ponderado, se obtiene una descripción más detallada de las tablas más seleccionadas por los usuarios que tienen más relaciones atípicas, relaciones consideradas en el grafo como aquellas en las que sobresalen los lazos de color amarillo y azul.

#### 6.2.2. Grafo Delete

En el siguiente grafo, se evidencia comparativamente, diferencias notables entre la metodología clásica y la metodología ponderada, en la primera se observa las relaciones que bajo concepto de la dirección de seguridad de la información son atípicas. En el grafo ponderado, se obtiene una descripción más detallada de las tablas más seleccionadas por los usuarios que tienen más relaciones atípicas, relaciones consideradas en el grafo como aquellas en las que sobresalen los lazos de color amarillo y azul.

#### 6.2.3. Grafo Insert

En el siguiente grafo, se evidencia comparativamente, diferencias notables entre la metodología clásica y la metodología ponderada, en la primera se observa las relaciones que bajo concepto de la dirección de seguridad de la información son atípicas. En el grafo ponderado, se obtiene una descripción más detallada de las tablas más seleccionadas por los usuarios que tienen más relaciones atípicas, relaciones consideradas en el grafo como aquellas en las que sobresalen los lazos de color amarillo y azul.

#### 6.2.4. Grafo Update

En el siguiente grafo, se evidencia comparativamente, diferencias notables entre la metodología clásica y la metodología ponderada, en la primera se observa las relaciones que bajo concepto de la dirección de seguridad de la información son atípicas. En el grafo ponderado, se obtiene una descripción más detallada de las tablas más seleccionadas por los usuarios que tienen más relaciones atípicas, relaciones consideradas en el grafo como aquellas en las que sobresalen los lazos de color amarillo y azul.

Figura 1: Grafo Select Clásico

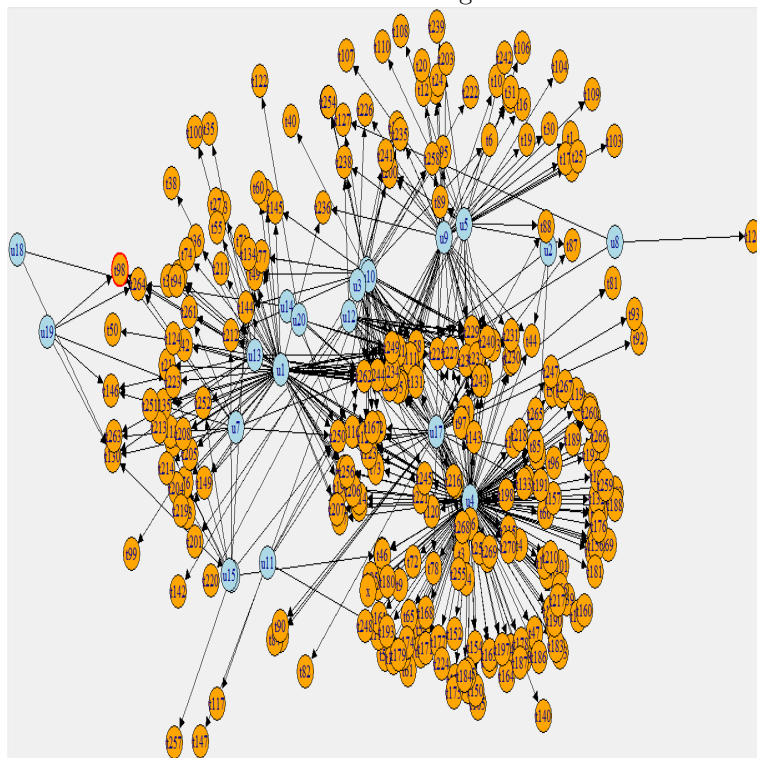


Figura 2: Grafo Select No Clásico

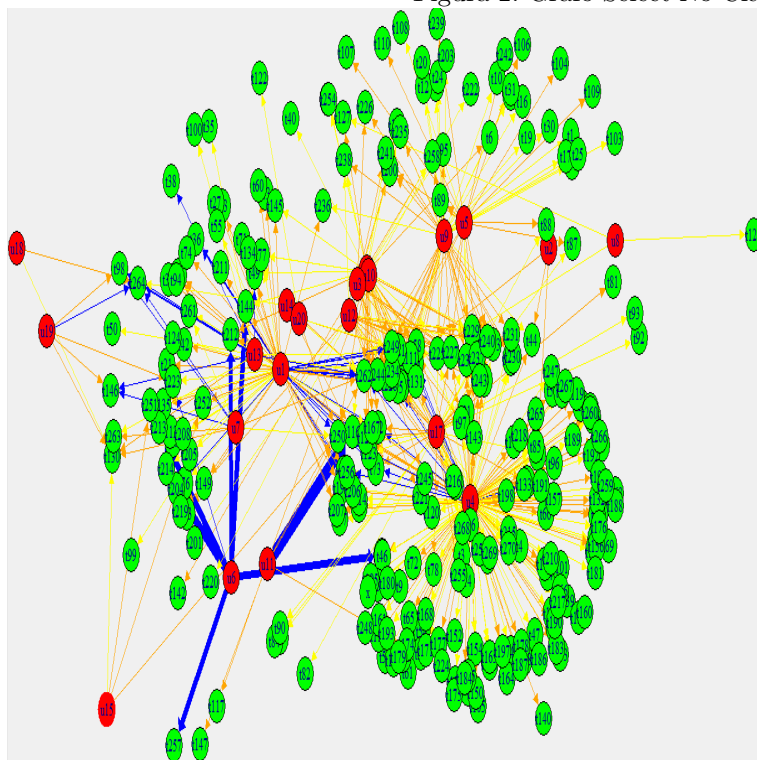


Figura 3: Grafo Delete Clásico

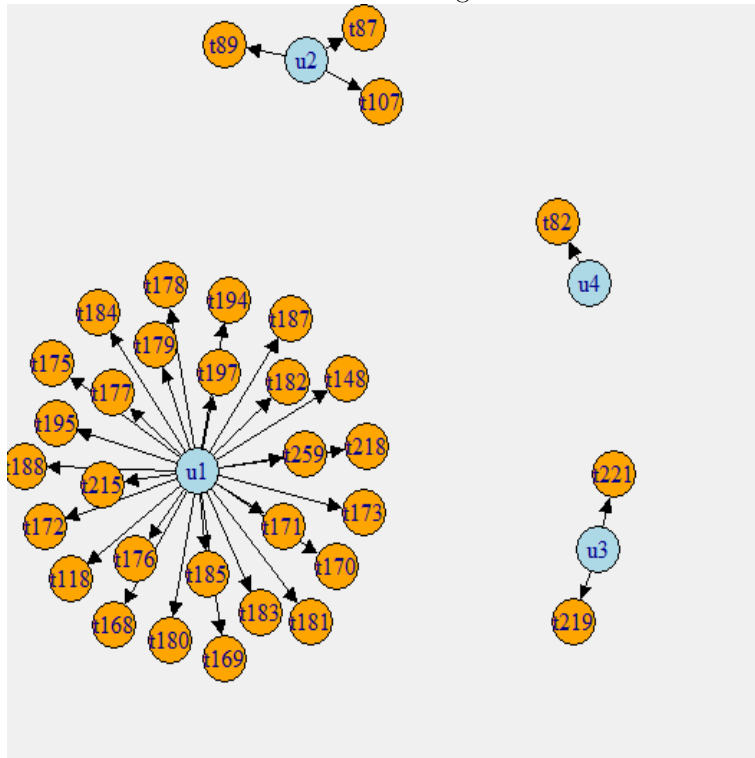


Figura 4: Grafo Delete No Clásico

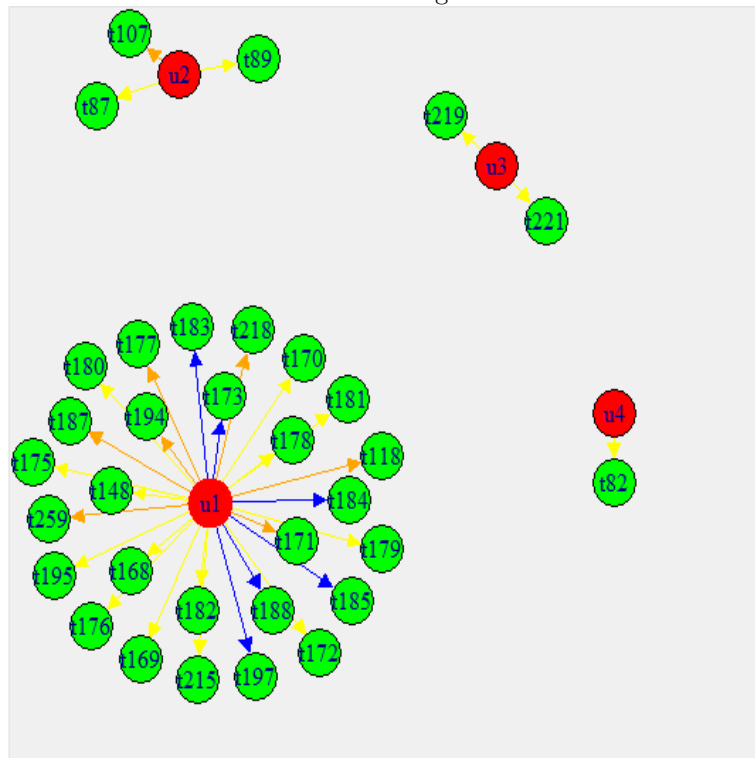


Figura 5: Grafo Insert Clásico

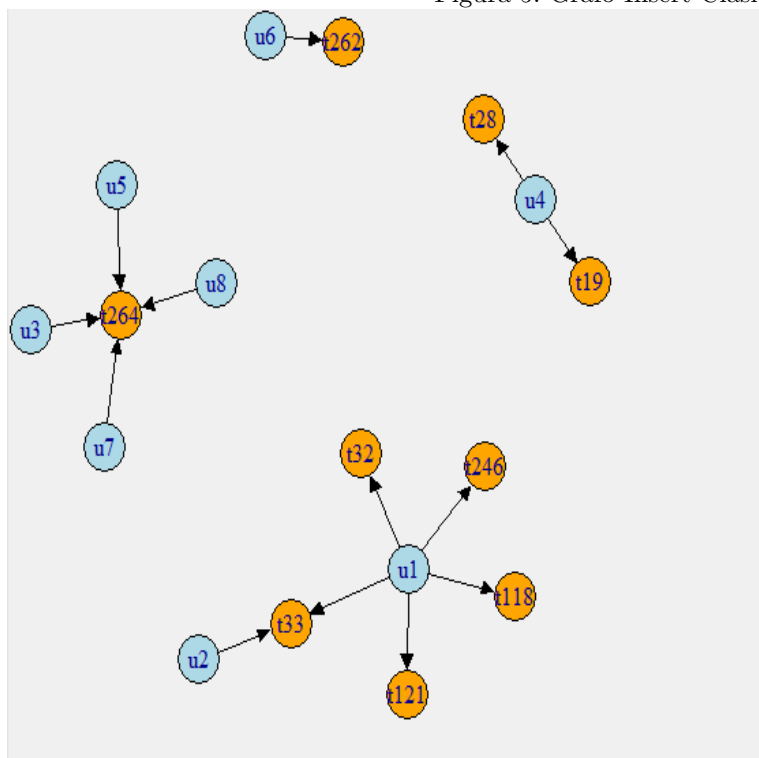


Figura 6: Grafo Insert No Clásico

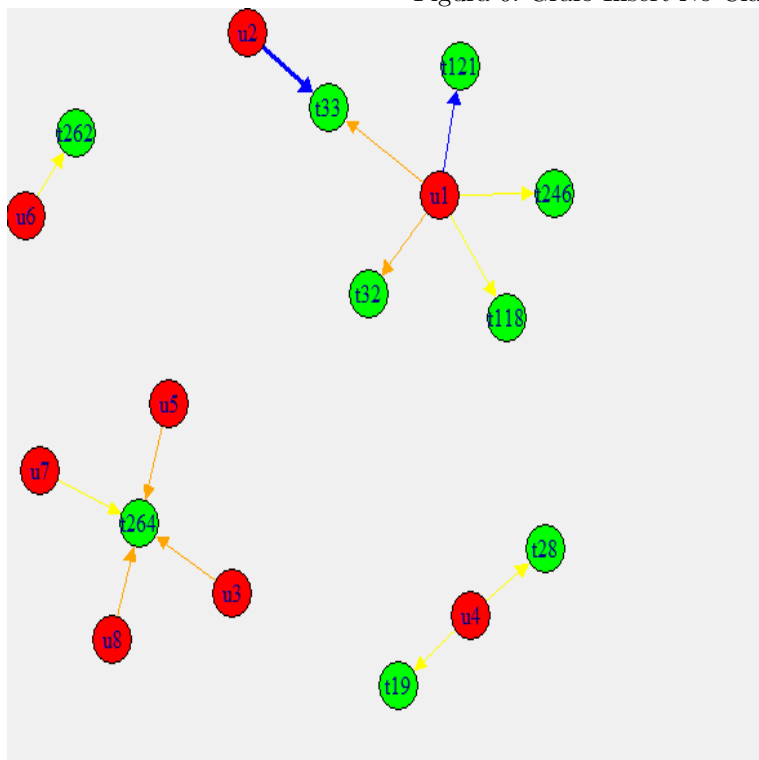


Figura 7: Grafo Update Clásico

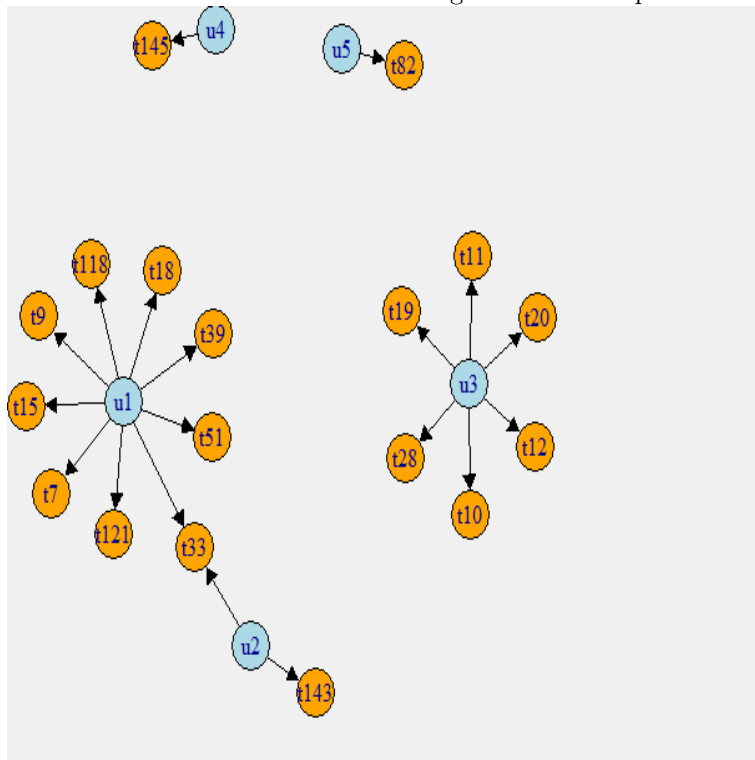
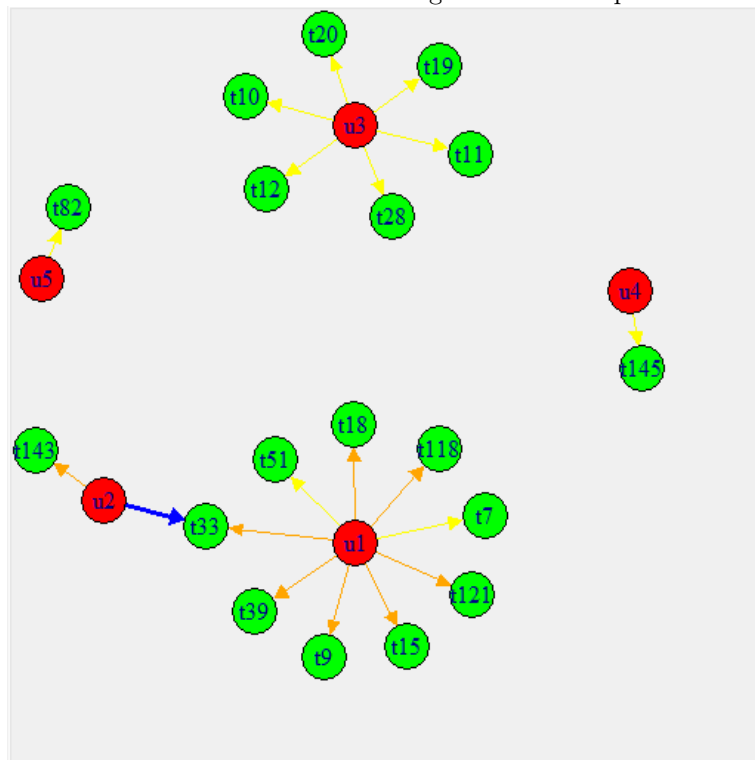


Figura 8: Grafo Update No Clásico



### 6.3. Hallazgos en la etapa Descriptiva

Para medir a través de las relaciones la importancia de los usuarios y las tablas(bases de datos) se usaron diferentes herramientas.

#### 6.3.1. Outdegree Clásico

Para determinar el nodo más importante en cuanto a usuarios, se empleo la medida del outdegree como una estimación de un parámetro de centralidad que permite de acuerdo con la metodología clásica determinar el usuario que se interrelaciona con más tablas(bases de datos). Consiste en determinar la cantidad de enlaces salientes<sup>20</sup> que genera un actor.

En la metodología clásica un outdegree alto evidenciará un usuario con gran número de interacciones con diferentes tablas, medida muy útil a la hora de verificar actividades sospechosas por parte de los usuarios, se espera que por lo general dada la naturaleza privilegiada del acceso a las bases de datos los usuarios tengan valores muy concentrados es decir, con poca varianza y con poca densidad.

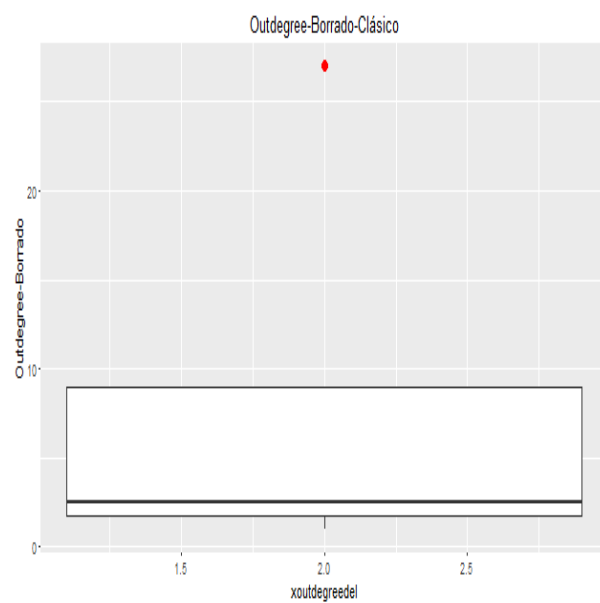
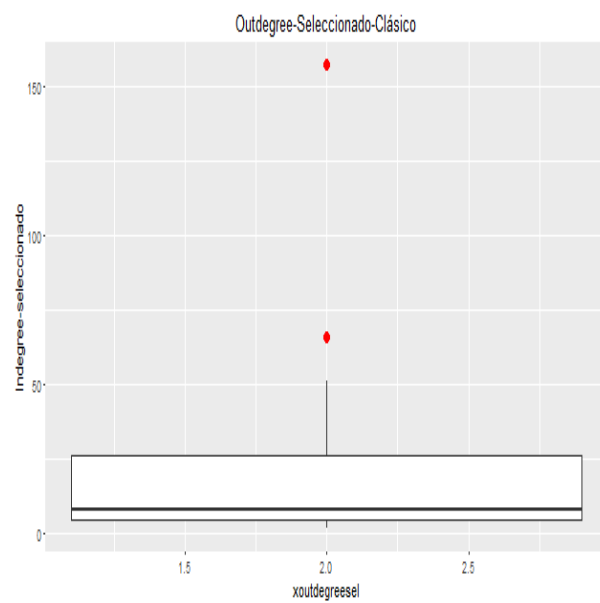
Como se observa en la mayoría de histogramas los valores de outdegree tienden a ser muy concentrados, sin embargo importantes actividades en el grafo y valores con alto valor de outdegree de baja frecuencia son fundamentales para el estudio ya que reflejan la atipicidad de los datos para aquellos actores que cuentan un alto nivel de centralidad

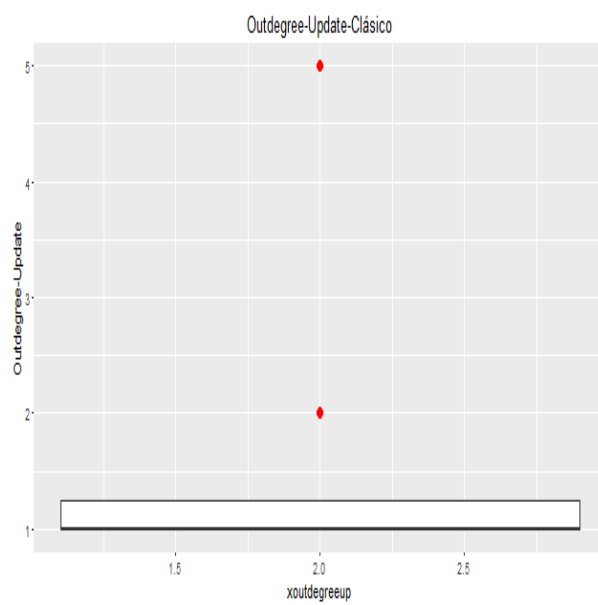
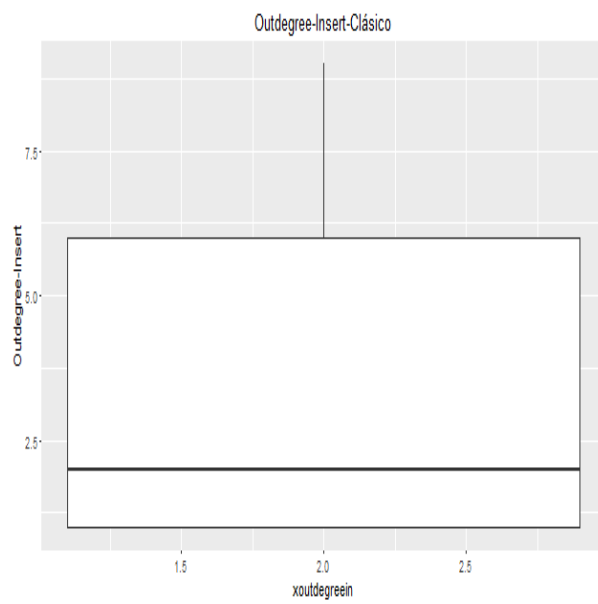
---

<sup>20</sup><http://cs.uns.edu.ar/agm/mineriaweb/downloads/Slides/clase20-slides-marcelo-durante.pdf>

Se presentan los boxplot del outdegree de la metodología clásica, los boxplot tienen la ventaja de ser no paramétricos y los espacios entre las diferentes partes de la caja muestran el grado de dispersión en rango y la asimetría de los datos, también en los boxplot muestran los outliers es decir aquellos puntos que exceden por mucho la dispersión en rango de los datos y que gráficamente pueden ser identificados como evidencia de outdegrees atípicos, o muy por encima de lo esperado.

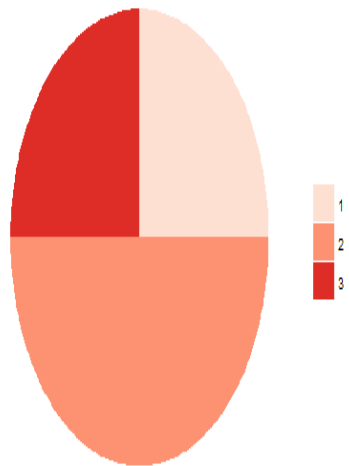
vale la pena resaltar que los boxplot en algunos casos donde los cuartiles son idénticos presentan unos datos en rango muy concentrados, es decir los boxplot no resultan muy informativos, esto sucede por que el outdegree se repite numerosas veces.



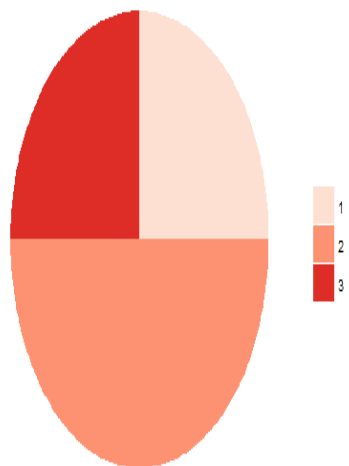


La variable artificial creada, para la elaboración de los boxplot es representada a continuación en pies o diagramas circulares definiendo como los usuarios de interés a aquellos que se encuentran por debajo de la mediana entre 0 y el cuartil que representa el 25 por ciento de la información, así como también aquellos que se encuentran por encima de la mediana como la población de usuarios que son de interés para este estudio.

outdegree-Seleccionado-Clásico



Outdegree-Borrado-Clásico



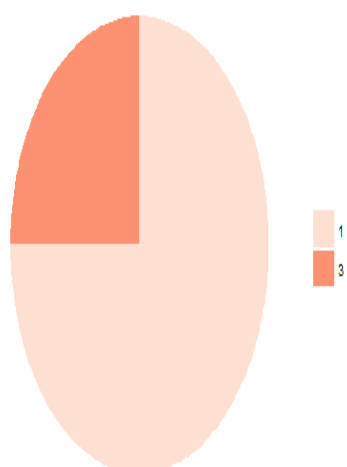
### 6.3.2. Outdegree Ponderado(no clásico)

Una vez evidenciado los outdegrees Clásicos se muestra el estudio del outdegree Ponderado, este surge con el objetivo de calcular no los usuarios que interactúan con más tablas sino más bien calcular aquellos usuarios cuya intensidad de relación medida frecuentemente se destaca sobre el resto, razón por la cual el outdegree clásico constituye una medida complementaria más no sustitutiva del outdegree clásico.

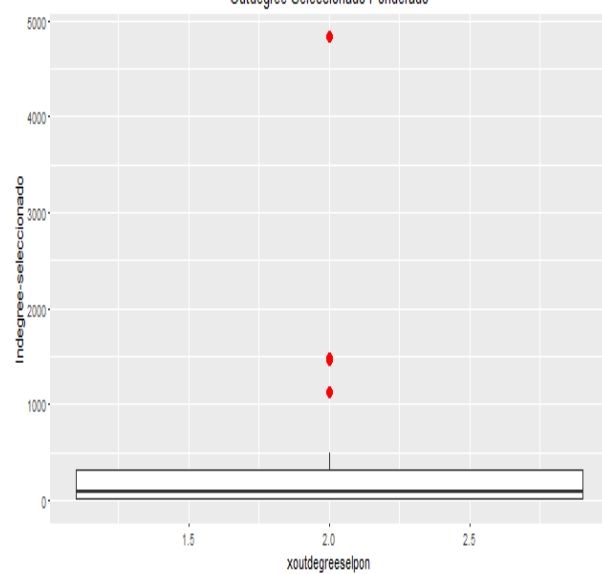
En los histogramas se evidencia un cambio total en la concentración de los datos, se logra evidenciar valores de frecuencia mucho más altos que anteriormente no se habían evidenciado con total claridad. Los usuarios que contaban con un outdegree clásico alto no cuentan con el mismo nivel de relevancia en cuanto a sus relaciones con tablas en el cálculo del outdegree ponderado, detectando de este modo la presencia de más usuarios atípicos.

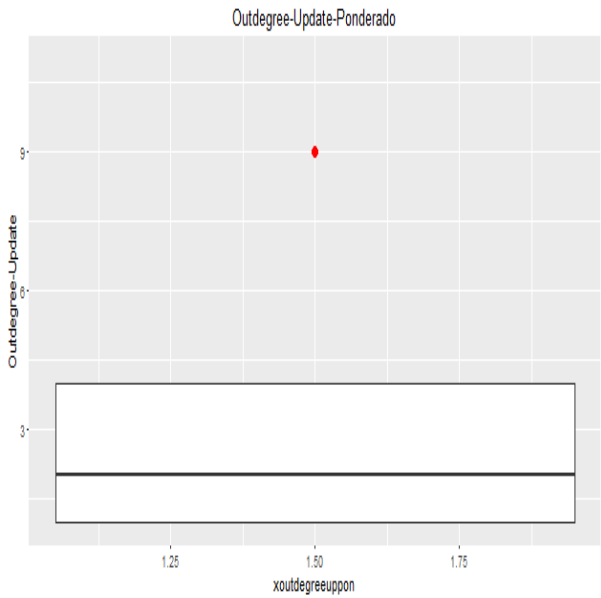
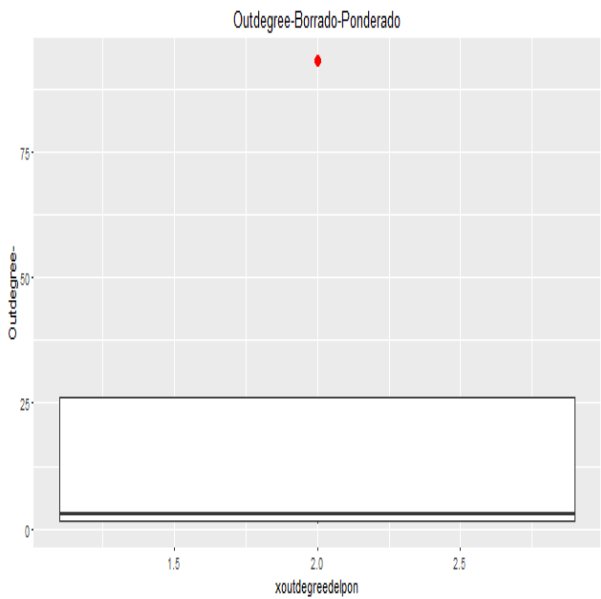
Un cambio relevante, dentro del outdegree ponderado se observa también en los boxplot siendo estos mucho más informativos se analiza y se observa la particularidad de una mayor dispersión de las observaciones con valores de outdegree superiores a la mediana así como la presencia de outliers nuevamente en los datos.

Outdegree-Update-Clásico



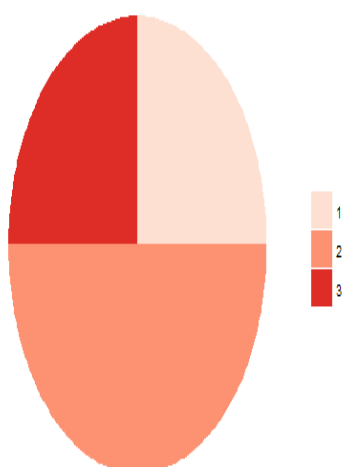
Outdegree-Selecionado-Ponderado



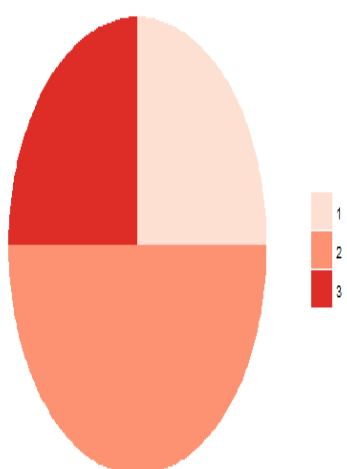


Se muestran los pies analizando nuevamente la concentración de las poblaciones objetivo tomando como población a analizar aquellos individuos que tienen un outdegree por debajo entre 0 y el cuartil que representa el 25 porciento de la información así como aquellos con un outdegree que excede la mediana de los datos.

outdegree-Seleccionado-Ponderado



Outdegree-Borrado-Ponderado



### 6.3.3. Densidades del grafo Ponderado y Clásico por su Outdegree

A continuación se presentan las densidades clásicas de cada uno de los grafos, tomando en cuenta que la densidad constituye el degree promedio de la variable estandarizada, se espera que el valor sea bajo para el outdegree ya que un grafo menos denso se ve representado en una distribución más equitativa de la centralidad de las relaciones de los usuarios con las bases de datos es decir, menor concentración de las operaciones respecto de un único usuario.

Tabla 2: Densidad Outdegree

| Grafo  | Densidad Clasica | Densidad Ponderada maximo Tore Opsahl |
|--------|------------------|---------------------------------------|
| Select | 0,005893254      | 0,1082056                             |
| Delete | 0,024774775      | 1                                     |
| Insert | 0,037549407      | 0,2465278                             |
| Update | 0,047794118      | 1                                     |

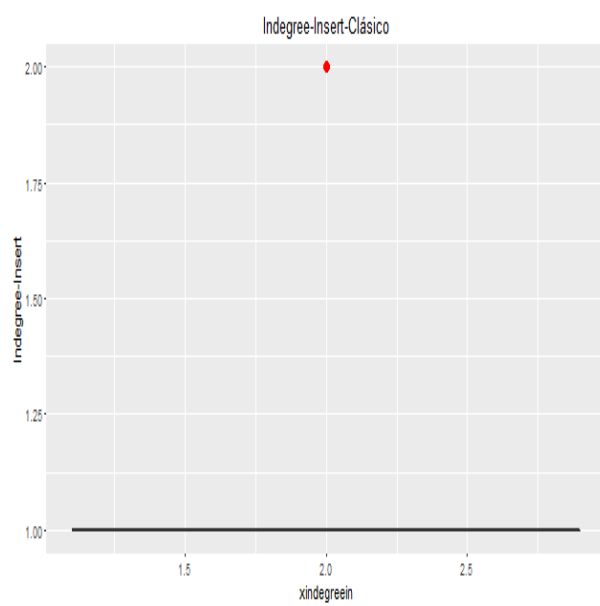
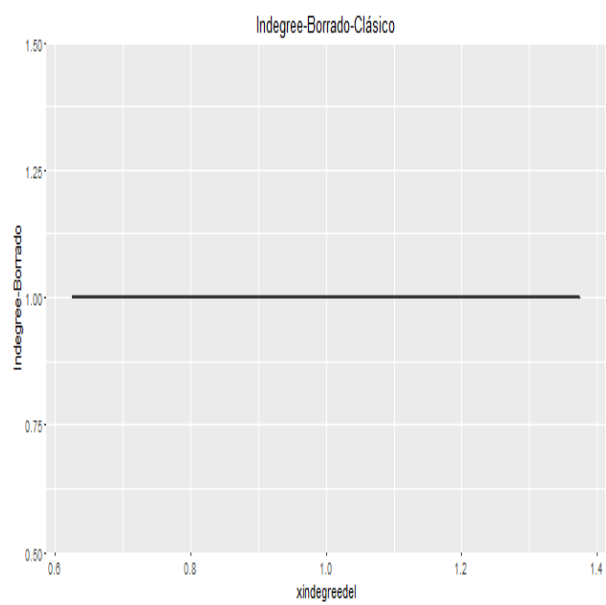
### 6.3.4. Análisis del Indegree Clásico.

En este cálculo se tiene como objetivo calcular las tablas más prestigiosas en la estructura de bases de datos, estas son las que cuentan con mayor número de operaciones hechas por diferentes usuarios privilegiados de las bases de datos o indegree.

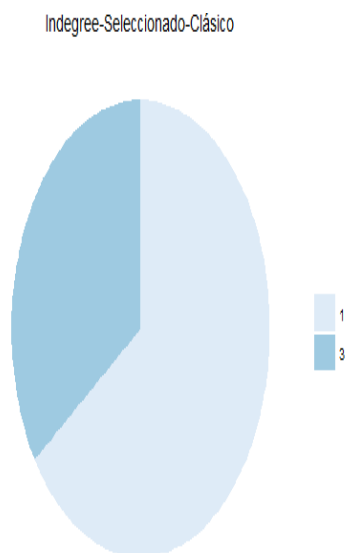
En este contexto, el histograma evidencia que tan concentradas se encuentran las operaciones sobre las bases de datos y llama la atención aquellos valores de alto valor que poseen una baja frecuencia, ya que dentro de la teoría SNA estas cobran gran valor ya que si se analiza con detenimiento el concepto de prominencia —es aquel actor que los actores eligen por encima de los otros — en definitiva su indegree.

El objetivo de este estudio consiste en calcular cuales tablas son pozos o nodos donde las elecciones de los usuarios de bases de datos se concentran con el fin de medir la importancia de la tabla

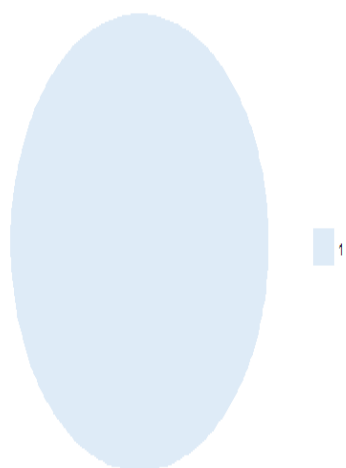
Los boxplot en estos casos representan la concentración de los valores de indegree para cada una de las 271 tablas analizadas, se espera que los datos no sean tan dispersos es decir, que los cuartiles primero tercero y segundo coincidan haciendo que el boxplot sea poco informativo de lo contrario se asumirá que la estructura tiende a centralizarse lo que quiere decir que existe una distribución irregular del prestigio de las tablas



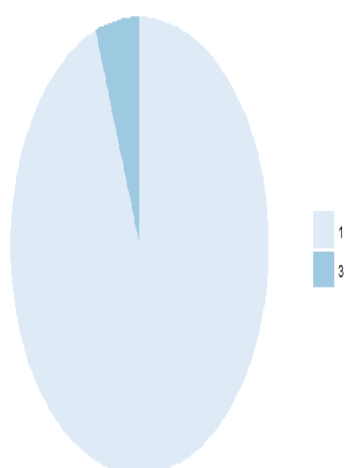
Por ultimo se representan los pies haciendo énfasis en el análisis de aquellas tablas que se encuentran entre 0 y el cuartil 25 con un valor bajo así como aquellas observaciones que se encuentran por encima de la mediana, como tablas cuyo contenido vale la pena analizar.



Indegree-Borrado-Clásico



Indegree-Insert-Clásico

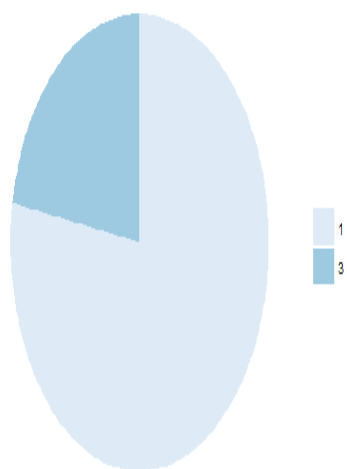


### 6.3.5. Análisis del Indegree Ponderado

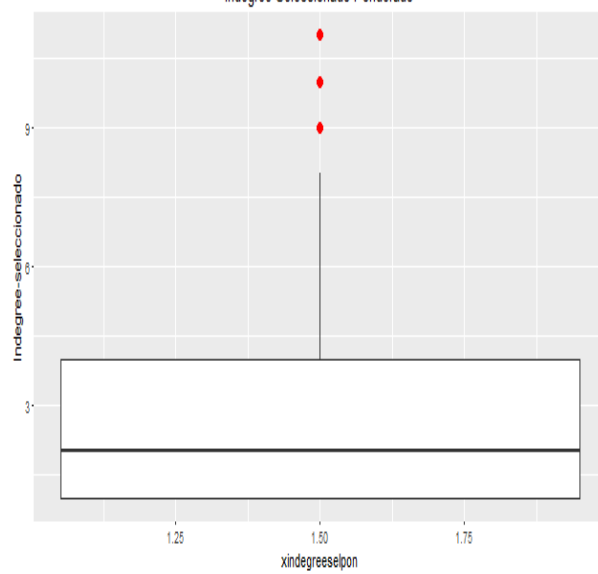
En el indegree ponderado, o cálculo del prestigio de las tablas, los resultados cambian considerablemente mostrando como atípico aquellos valores nuevamente de alto valor de baja frecuencia como aquellas tablas que vale la pena analizar y que no son iguales a las halladas por el prestigio clásico.

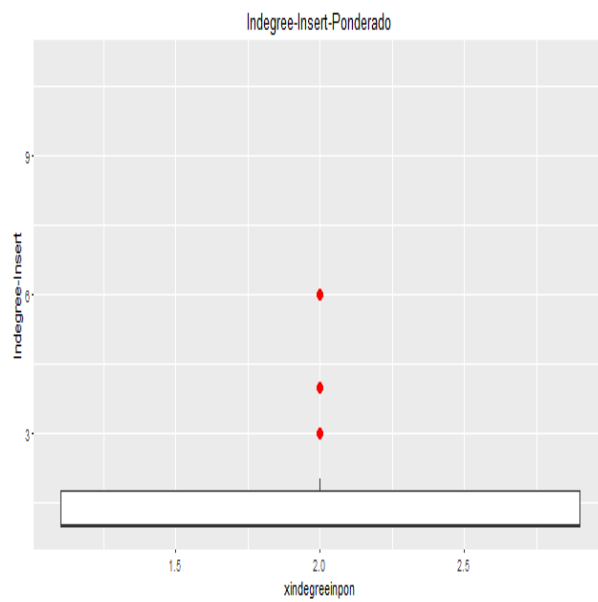
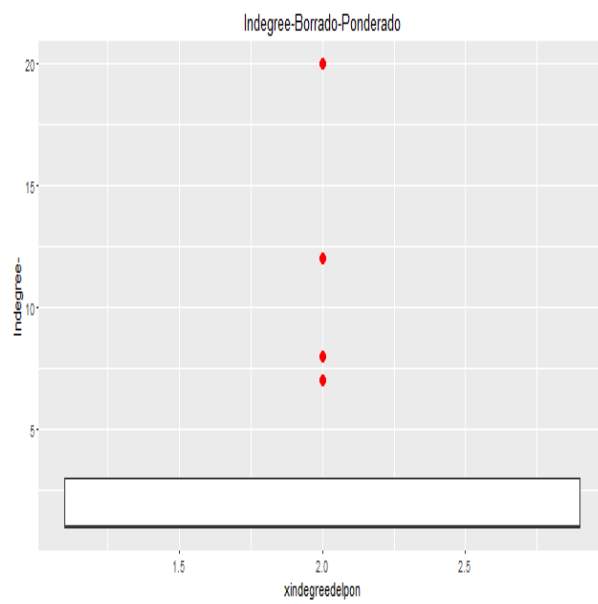
Se muestran los boxplot que evidencian la concentración del prestigio destacando nuevamente aquellos valores con un prestigio mayor a la mediana y la presencia de outliers como valores que vale la pena analizar.

Indegree-Update-Clásico

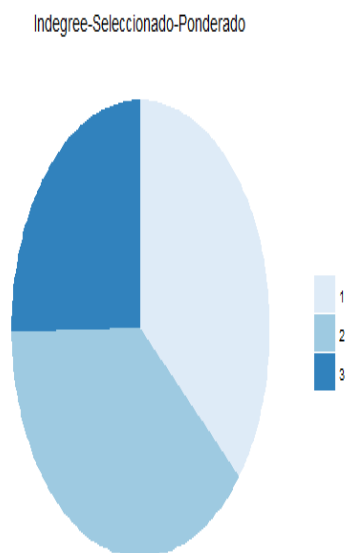
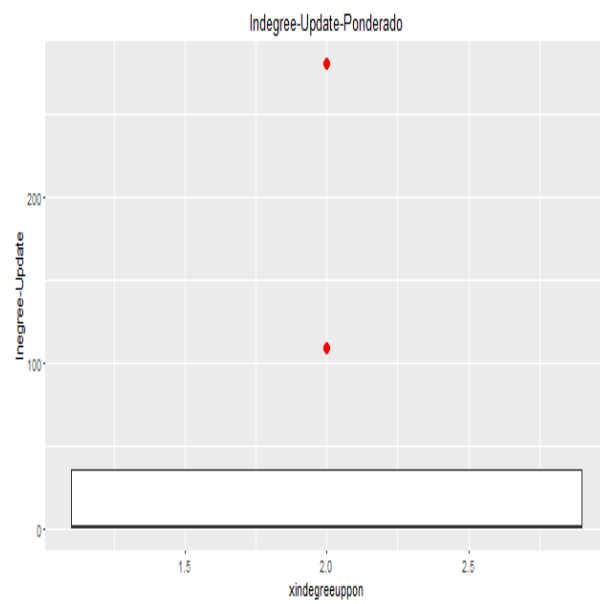


Indegree-Seleccionado-Ponderado

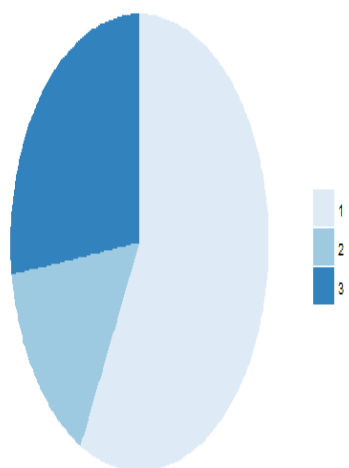




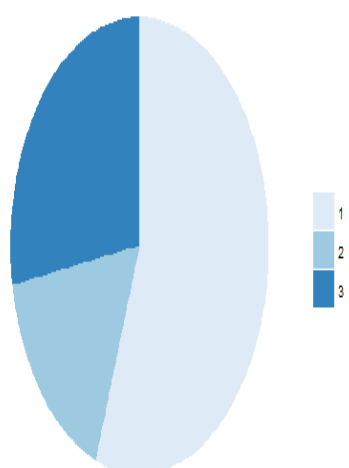
Finalmente se presentan los pies que muestran como se distribuyen las tablas en cada uno de los cuartiles mostrando cuales se deberían analizar nuevamente ya sea por bajo o alto valor en el indegree.



Indegree-Borrado-Ponderado



Indegree-Insert-Ponderado



### 6.3.6. Densidad del Grafo por Indegree

A continuación se muestran las densidades del indegree clásico mostrando la regularidad o centralidad dependiendo del valor de la densidad del grafo, en este grafo en esta tabla en particular se transformó la variable indegree en una variable con distribución uniforme cero uno y se procedió al cálculo de la densidad del grafo mediante el cálculo del promedio de la variable.

Tabla 3: Densidad grafo por Indegree

| Grafo  | Densidad Clasica | Densidad Ponderada | Densidad Ponderada maximo Tore Opsahl |
|--------|------------------|--------------------|---------------------------------------|
| Select | 0,005893254      | 0,01074047         | 0,02724805                            |
| Delete | 0,024774775      | 0,13513514         | 0,10685805                            |
| Insert | 0,037549407      | 0,04281298         | 0,06597222                            |
| Update | 0,047794118      | 0,09054622         | 0,16806053                            |

### 6.3.7. Conectividad

El número de nodos que debe ser eliminado para que el grafo quede desconectado

Tabla 4: Add caption

| Grafo  | conectividad |
|--------|--------------|
| Select | 1            |
| Delete | 1            |
| Insert | 1            |
| Update | 1            |

### 6.3.8. Cercanía

La cercanía representa de acuerdo con Opsahl 2009 representa la suma del inverso de las distancias geodésicas que no contienen la relación directa entre los nodos medidos .

Se esperará en los histogramas que la cercanía tome valores de alta frecuencia de baja cercanía para los Usuarios como evidencia de que existen perfiles de acceso a la información que maximizan la distancia geodésica minimizando la cercanía de estos respecto a las tablas.

Se espera por otra parte que la cercanía de las tablas respecto a los usuarios tome valores de alta frecuencia de baja cercanía como evidencia de la existencia de políticas que restringen el acceso a la información sensible.

Por último se espera que la distribución sea asimétrica y leptocúrtica en cuanto a su cercanía con el fin de evidenciar la ausencia de centralización ya sea en un usuario o tabla.

**kurtosis** es importante aclarar que si en los valores encontrados

- $x = 0$  (distribución mesocúrtica).
- $x > 0$  (distribución leptocúrtica).
- $x < 0$  (distribución platicúrtica).

**Asimetría** es importante aclarar que si en los valores encontrados

- $x = 0$  (distribución simétrica; existe la misma concentración de valores a la derecha y a la izquierda de la media)
- $x > 0$  (distribución asimétrica positiva; existe mayor concentración de valores a la derecha de la media que a su izquierda)

Indegree-Update-Ponderado

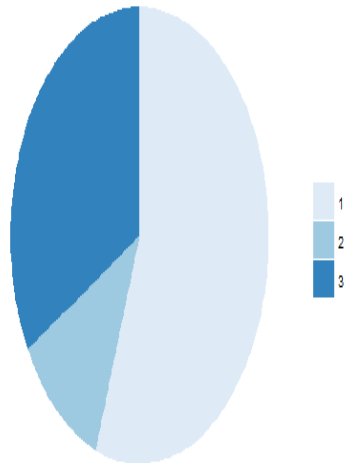


Tabla 5: Kurtosis Cercanía Indegree

| Grafo  | kurtosis clásica indegree | kurtosis Ponderada indegree |
|--------|---------------------------|-----------------------------|
| Select | 6,0712039                 | 6,07364                     |
| Delete | 4,3712121                 | 2,82076                     |
| Insert | 5,3898586                 | 4,807424                    |
| Update | 0,7143423                 | 15,531397                   |

Tabla 6: kurtosis cercanía outdegree

| Grafo   | kurtosis clasica outdegree | kurtosis Ponderada outdegree |
|---------|----------------------------|------------------------------|
| Select  | 215,187148                 | 214,836354                   |
| Delete  | 31,919245                  | 31,85525                     |
| outsert | 7,893222                   | 8,24604                      |
| Update  | 9,512957                   | 4,345566                     |

- $x < 0$  (distribución asimétrica negativa; existe mayor concentración de valores a la izquierda de la media que a su derecha)

Tabla 7: Asimetría cercanía Indegree

| Grafo  | skewness clasica | skewness Ponderada |
|--------|------------------|--------------------|
| Select | 2,3336058        | 2,203332           |
| Delete | -2,524126        | -2,090243          |
| Insert | 2,3072521        | -2,330724          |
| Update | -0,2774817       | -4,039769          |

Tabla 8: Asimetría cercanía outdegree

| Grafo   | skewness clásica | skewness Ponderada |
|---------|------------------|--------------------|
| Select  | 14,095084        | 14,071527          |
| Delete  | 5,819215         | 5,8109857          |
| outsert | 2,878869         | -2,8814123         |
| Update  | 3,240586         | 0,9604051          |

#### 6.4. Hallazgos en la etapa de formulacion de criterios para guiar el Monitoreo y generación de alertas tempranas para situaciones atípicas

Se enunciarán a continuación objetivo por objetivo como se cumplieron las expectativas

Se Describieron y crearon criterios de interpretación, a través de técnicas de Social Networking Analysis, que permitieron elevar el nivel de seguridad información de las operaciones que realizan los usuarios sobre las bases de datos del DEPÓSITO y que se gestionan en SQL server e identifico sus atipicidades. Al detectar mediante un esquema de intervalos las relaciones atípicas encontrando evidenciando e identificando lo siguiente:

- Existencia de usuarios privilegiados con acceso distanciado de las tablas manejadas por el común de los usuarios del DEPÓSITO.
- Existencia de Tablas temporales con las cuales los usuarios realizan las diferentes actividades de actualizado e inserción.
- Ausencia de estructura jerárquica que impide a los usuarios tener el mismo grado de cercanía a las tablas.
- Existencia de un portal, según el cual la mayoría de usuarios de deceval interactúan directamente con las bases de datos lo que genera un aumento de la vulnerabilidad ya que los grafos son un nodo vulnerables.
- Se identificaron bots que sobrecargaban al registrar su trazabilidad la velocidad del sistema de gestión de bases de datos.

Se Recolectaron los datos a través del software diseñado por IBM conocido como DAM , el cual mostró las diferentes operaciones realizadas por los usuarios con las bases de datos del DEPÓSITO registradas por milisegundo. cumpliendose el primer objetivo específico.

Se Describió a través de un grafo el conjunto de relaciones directas (ties) que se dan entre los usuarios y las bases de datos del DEPÓSITO. Tanto en el esquema Clásico como en el ponderado.

Se Medieron a través de relaciones, la importancia de los usuarios y las bases de datos Por medio de sus medidas de centralidad.

Se Identificaron y analizaron relaciones recurrentes para descubrir situaciones atípicas.

## 6.5. Conclusiones

Por medio del desarrollo anteriormente plasmado se debe destacar que en este trabajo, se cumplió con el objetivo propuesto de describir y crear criterios que permitieron elevar el nivel de seguridad de la información de las operaciones que realizaron los usuarios sobre las bases de datos del DEPÓSITO ya que se contó por primera vez una herramienta que permite la visualización de los datos de un modo más práctico y eficiente según el cual se pudieron identificar relaciones atípicas de los usuarios con las bases de datos dentro del DEPÓSITO. Por primera vez gracias a este trabajo, se generó un esquema de alertas tempranas y una metodología a seguir para los datos arrojados por la analítica en el área de Seguridad de la información dentro de la organización. El trabajo desarrollado permitió que área de Riesgos y Cumplimiento desarrollara un interés por la metodología propuesta y propusiera llevarla a un esquema mucho más elevado en el área de operaciones dentro del DEPÓSITO. Se propone la aplicabilidad de esta metodología para futuros monitoreos tales como:

- Monitoreo de mensajería corporativa a cuentas personales
- Monitoreo de actividad en firewalls
- Monitoreo de entrega de información
- Monitoreo copia de información a usb

## Referencias

- [1] IBM. *DAM*. DAM main page privative use. Tore Opsahl.
- [2] *Structure and Evolution of Weighted Networks*. [University of London (Queen Mary College), London, UK]. Thesis 2009, 104-122.
- [3] *Social Network Analysis. Methods and Applications* 1994.
- [4] School of Neurological science University of Nebraska Lincoln ,  
<http://www.shizukalab.com/toolkits/sna/weighted-edges>
- [5] Seguridad de la información,  
[https://es.wikipedia.org/wiki/Seguridad\\_de\\_la\\_informacion](https://es.wikipedia.org/wiki/Seguridad_de_la_informacion)
- [6] <http://igraph.org/redirect.html>
- [7] [https://en.wikipedia.org/wiki/Social\\_network\\_analysis](https://en.wikipedia.org/wiki/Social_network_analysis)
- [8] <http://kateto.net/networks-r-igraph>
- [9] <https://assemblingnetwork.wordpress.com/2013/06/10/network-basics-with-r-and-igraph-part-i/>

## A. Códigos

los codigos se presentan a continuación<sup>2122</sup>

---

<sup>21</sup><http://kateto.net/networks-r-igraph>

<sup>22</sup><https://assemblingnetwork.wordpress.com/2013/06/10/network-basics-with-r-and-igraph-part-ii-of-iii>