Instituto Tecnológico de Costa Rica

Escuela de Ingeniería en Computación

Programa de Maestría en Computación

Visualización de conocimiento en temas específicos mediante el uso del correo electrónico corporativo

Tesis para optar al grado de Magíster Scientiae en Computación

Berny Alvarado Brenes

Ph. D. Franklin Hernández-Castro

Noviembre, 2014

Resumen

Conocer, en una corporación, quiénes tienen conocimiento sobre un tema, es la clave para una efectiva toma de decisiones que fortalezca la colaboración y un eficiente uso de los recursos. Basados en estudios de análisis de redes sociales, se evalúan técnicas de visualización, se selecciona el paradigma de visualización más apto y se desarrollan mejoras sobre el mismo. Todo esto con el objetivo de probar si es posible determinar la localización del conocimiento dentro de una corporación mediante el uso del correo electrónico corporativo. Al final se muestra un paradigma de visualización que permite analizar claramente las comunicaciones dentro de la corporación y a partir de ahí inferir las experticias y localizaciones del personal en temas específicos.

Abstract

Understanding who within a corporation has knowledge on a given topic, is key for effective decision making that enables the best use of the resources and strengthens collaboration. Based on studies of social network analysis, information visualization techniques are evaluated, the most adequate visualization is selected, and improvements developed over it. All this is done to proof if it's possible to determine the knowledge localization within a corporation through its corporate email. Lastly, a visualization paradigm is shown. This visualization paradigm facilitates a clear analysis of the communications within a corporation and from the ones it's possible to understand the expertise and personnel localization in specific topics.

Programa de Maestría
en Computación
Teléfono: 2550-2402

APROBACIÓN DE LA TESIS

"Visualización de conocimiento en tópicos específicos mediante el uso del correo electrónico corporativo"

TRIBUNAL EXAMINADOR

Ph.D. Franklin Hernández Castro
Profesor Asesor

Dr. Erick Mata Montero
Profesor Lector

Ph.D. Tomás de Camino Beck
Profesional Externo

Dr. Roberto Cortés Morales
Coordinador del Programa
de Maestría en Computación

Noviembre, 2014

Tabla de Contenidos

Dedicatoria

A mi Dios, por ser la fuente de mi vida y mi inspiración.

A mi amiga, esposa y compañera del alma, Irma Guisela Montero Vargas, por su aliento, cariño, amor y apoyo incondicional en todo momento.

A mi hija Amanda Alvarado Montero, por ser inspiración de ternura y esperanza.

A mis padres, Fernando Alvarado Zumbado y Gladys Brenes Rojas, quienes siempre me motivaron y ayudaron en mi formación académica.

Agradecimientos

A mi Dios que me ha dado los insumos y ha propiciado las condiciones necesarias para llevar a feliz término esta meta personal.

Mi agradecimiento especial para mi director de tesis, Ph. D. Franklin Hernández, por la dirección de esta investigación, sus valiosos comentarios y disponibilidad.

A los lectores Ph. D. Tomás De Camino Beck, Dr. Erick Mata Montero y Dr. Roberto Cortés Morales, por su colaboración en la revisión de esta tesis y sus oportunas observaciones de mejora.

Al Ministerio de Ciencia y Tecnología (MICIT), al Consejo Nacional de Ciencia y Tecnología (CONICIT) y a INTEL por su contribución para financiar los costos de la maestría.

Al Instituto Tecnológico por propiciar esta maestría y facilitar su implementación.

1. Planteamiento del Problema

En pro de facilitar la colaboración y hacer un uso efectivo de los recursos en cualquier organización, resulta indispensable comprender cuáles son aquellas personas con conocimiento sobre un tema específico que les permita colaborar para lograr un objetivo en común.

La comprensión de quiénes son las personas relevantes, con conocimiento experto en un tema, no es una tarea sencilla. Es de vital importancia alguna forma de visualización que nos permita abstraer, a partir de un conglomerado de comunicaciones electrónicas, aquellas personas cercanas que pudieran caracterizarse con conocimiento relevante en un tema.

2. Objetivos

2.1 Objetivos Generales

- Implementar una visualización que nos permita descubrir la localización dentro de la corporación del conocimiento de temas específicos a través del correo corporativo.

2.2 Objetivos Específicos

- Analizar diversas técnicas de visualización y evaluar su efectividad para evidenciar la localización del conocimiento en temas específicos.
- Definir el mejor método de visualización para la localización del conocimiento en temas específicos.
- Implementar y evaluar un prototipo que permita la visualización propuesta.

3. Hipótesis

Es posible identificar la localización del conocimiento dentro de la corporación en temas específicos a través de la visualización del correo electrónico corporativo.

4. Justificación

Tanto en la industria como en la academia existe una gran necesidad de promover un ambiente de producción más ágil y efectivo que aproveche al máximo los recursos con que se cuentan. Desafortunadamente, es común encontrar organizaciones y personas que trabajan de forma aislada, desaprovechando el gran valor que la sinergia y la colaboración pueden brindar.

La carencia de algún método que evidencie la localización del conocimiento, priva a las organizaciones de información, con la cual pudiesen establecer planes y estrategias de colaboración. Resulta común descubrir entre grupos de trabajo y organizaciones, “islas” de trabajo, de personas cercanas, trabajando en temas relacionados pero que sin embargo no interactúan entre sí.

Nuestro objetivo de investigación es probar la hipótesis de que un método de visualización puede ayudar a localizar el conocimiento de un tema mediante el análisis de correos electrónicos.

5. Innovación

El presente trabajo busca aportar la aplicación de un paradigma de visualización que visualice la localización del conocimiento y su distribución en una corporación mediante el uso del correo electrónico corporativo. Lo anterior se hará experimentando con diferentes técnicas de visualización, con el objetivo de definir la visualización que resuelva el problema aquí planteado.

6. Antecedentes

Este trabajo está basado en tres grandes vectores de investigación. El primero es el análisis de las redes sociales o bien SNA, por sus siglas en inglés. El segundo es las estrategias de visualización y por último el minado del correo electrónico. De forma adicional, se han analizado trabajos previos relacionados a la colaboración, el manejo de conocimiento y los flujos de este, dentro y fuera de una organización.

6.1 Análisis de Redes Sociales (SNAs)

Dentro de la rama de la sociología se encuentra el análisis de las redes sociales. En 1979, Tichy, Tushman y Fombrun [41] publican un estudio sobre las redes sociales dentro de una organización. En dicho trabajo, estos investigadores hacen un resumen de las propiedades de una red social y proponen su aplicación para el análisis de una organización. Clasifican la red mediante conjuntos de agrupamiento (clusters), ya sea por la jerarquía de los individuos dentro de una organización o bien mediante la interacción entre los individuos (ver figura 6.1).

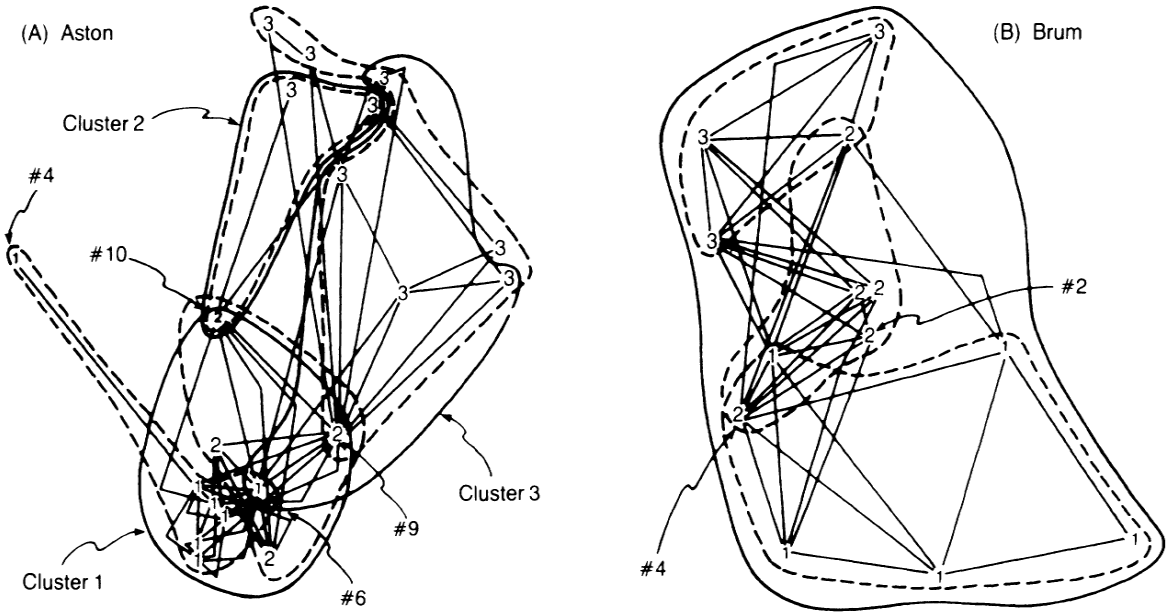


Figure 1
Interaction Networks

------ Prescribed clusters

——— Emergent clusters

Note: 1. Lines represent direct nominations between end points.
2. Clusters were determined from COMPLT and superimposed on SOCK scaling output.
3. Numbers represent formal status of individuals in organization. 1 = top, 2 = middle, 3 = supervisory.

Figura 6.1 Trabajo Tichy, Tushman y Fombrun, en el análisis de redes sociales en organizaciones [41].

En otra línea de investigación relacionada al análisis de las redes sociales se dan investigaciones sobre la coautoría y la colaboración en investigación entre universidades y países [1] [19] [25]. T Luukkonen, RJW Tijssen, O Persson y G Sivertsen [1] proponen la medición de la colaboración científica mediante medidas absolutas y relativas. Toman como factor de colaboración las relaciones de coautoría y los tamaños de los países para poder determinar la relevancia de la colaboración. Se hace uso de MDS (Multidimensional Scaling) como algoritmo para la generación de las redes. Dicho trabajo se mantiene a nivel de país y no emplea técnicas de visualización tales como focus+context, aun cuando el mismo trabajo sugiere que las variables para la medición de la colaboración pueden ser tantas, que una visualización de 1 plano de dos dimensiones resulta limitada para un análisis detallado de la colaboración (ver figura 6.2).

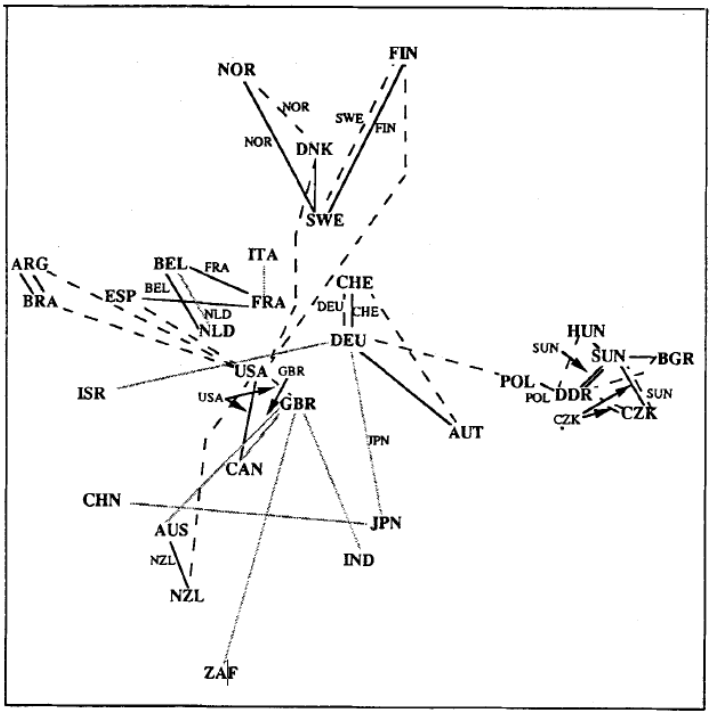


Fig. 4. Network map of international co-authorship relationships amongst 30 countries. Based on Salton's index. Derived with multidimensional scaling. Fit: 68% (RSQ = 0.68).
Connective lines: ——— Strongest link with a country other than the USA
– – – – Second strongest link – second to link with a country other than the USA
····· Second strongest link – second to link with the USA
Missing lines refer to links with the USA. Ambiguties are avoided by labeling selected lines with the respective country involved.

Figura 6.2 Red de relación de coautoría entre países en el trabajo “The measurement of international scientific collaboration” [1].

Nankani, Simoff, et al. [19] sugieren que la organización moderna de una universidad puede ser vista como la de un gran ecosistema. Establece un conjunto de posibles datos con los cuales se pudiera analizar la colaboración, sin embargo, por propósito de practicidad para la investigación, el set de datos es restringido a publicaciones, lista de participación de cada uno de los proyectos de investigación, así como también la caracterización de la personas que supervisan proyectos. El trabajo propone cosas interesantes, tales como la evolución de la red de colaboración (ver figura 6.3), sin embargo, no cubre aspectos tales como la fácil visualización del conocimiento según su localización dentro de la organización. Al mismo tiempo, al igual que otros

trabajos [24] [25] [33] relacionados al análisis de redes sociales, existe un número de conexiones, las cuales no resulta intuitivo visualizar.

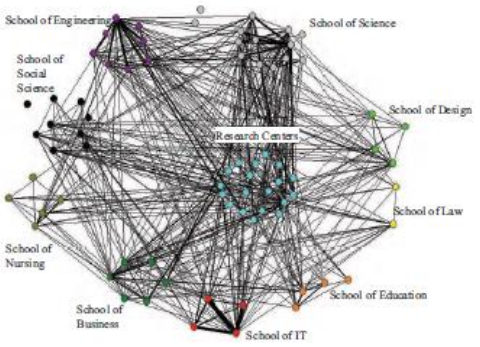


Fig. 3. Network based on departments

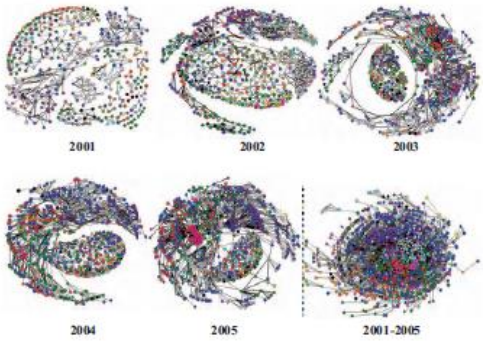


Fig. 5. Evolution of the network in five years

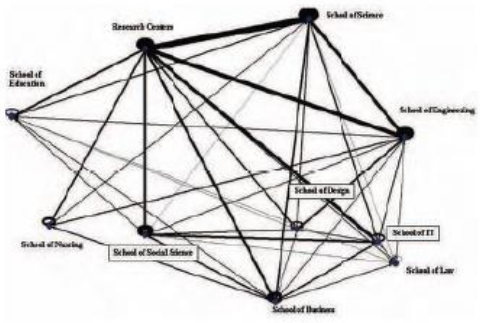


Fig. 4. Network based on schools

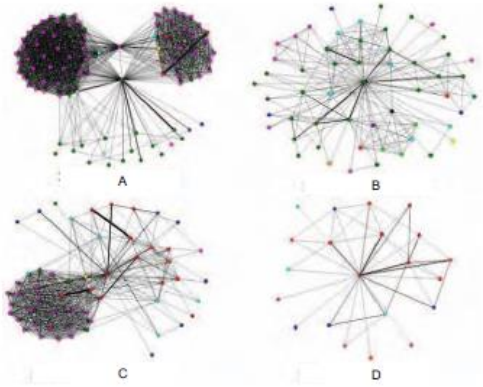


Fig. 6. Egonets of the top four collaborators $\{A, B, C, D\}$

Figura 6.3 Análisis de la colaboración dentro de una universidad [19]

En el 2004, siempre dentro del área de redes sociales, se da una investigación afín a la problemática que nos planteamos. SELaKT [7] busca poder localizar expertos dentro de una red social y poder visualizar la transferencia de conocimiento dentro de esta.

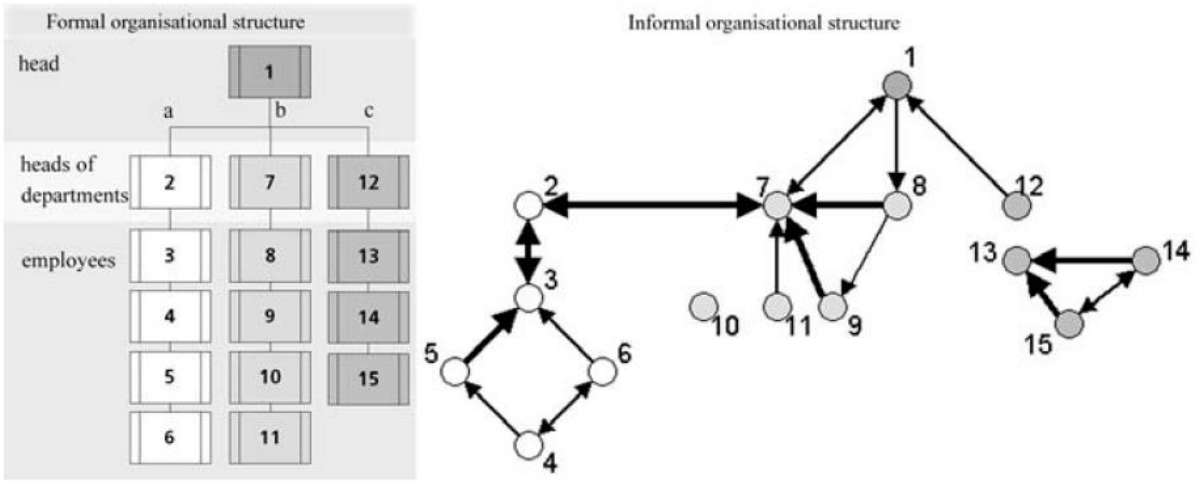


Figure 2: Formal Versus Expert Structure in a Research Organisation

Figura 6.4 El trabajo “SELaKT” [7] busca la localización de conocimiento experto y su localización dentro de una organización.

SELaKT es propuesto más como un método de localización de conocimiento experto en función de la administración del conocimiento dentro de una organización. Sin embargo, su practicidad no es evidenciada en el trabajo y no permite comprender cuáles desafíos el investigador hubiera descubierto en tanto dicho trabajo hubiera sido implementado y probado con datos reales. El método propuesto se resume a estudiar el caso descrito en la figura 2 del trabajo (descrita como “Formal Versus Expert Structure in a Research Organisation”) (ver figura 6.4). A diferencia de nuestra propuesta, este trabajo tampoco utiliza la variable de localización como un factor esencial, de tal forma que el conocimiento en la práctica esté apropiadamente distribuido. Esto es sobre todo cierto para aquellas organizaciones distribuidas en múltiples localizaciones. De igual forma, nuestra propuesta se enfoca comprender y visualizar el conocimiento mediante la minería sobre el correo electrónico. En esta área también existen trabajos previos, los cuales procederemos a describir.

6.2 Minería del correo electrónico

El correo electrónico es similar a un baúl de recuerdos o bien una bitácora de las comunicaciones entre una persona hacia otras y viceversa. El trabajo previo en esta área se encuentra al mismo tiempo relacionado con el análisis de las redes sociales, por una parte para establecer la relación y colaboración entre las personas, y por otra parte, como un histórico de datos útil para contar una historia pasada mediante una ayuda visual que resulta relevante para quien vivió el evento.

Peter A. Gloor, et al. [6] realizan un análisis del correo electrónico sobre un grupo del W3C con el propósito de identificar contribuidores claves dentro de la red, y al mismo tiempo comprender la estructura de la red a partir de las comunicaciones dentro del grupo (ver figura 6.5).

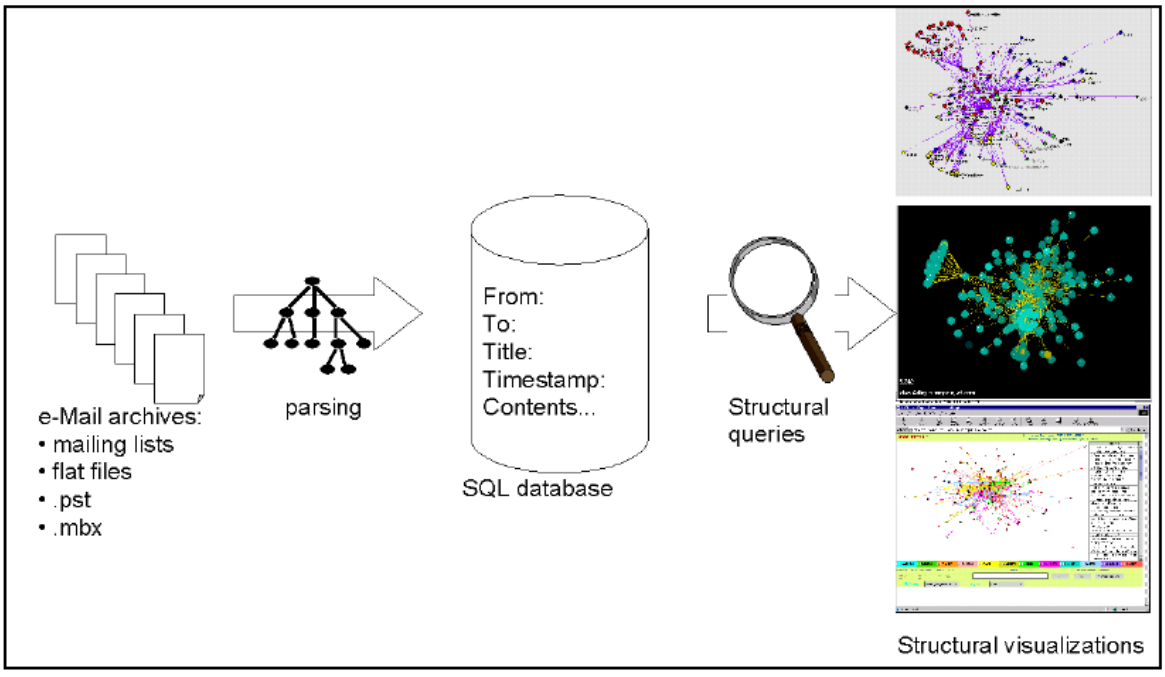


Figure 1. COIN e-mail analysis system architecture

Figura 6.5 Trabajo de A. Gloor, Et al., analiza patrones de comunicación en redes de colaboración mediante el estudio de listas de distribución [6].

A diferencia de nuestra propuesta, tanto la visualización de Gloor así como otros trabajos previos [9] [10], no se pone en el contexto a la persona que está realizando la visualización y cómo la visualización se puede adecuar a la posición de esta dentro de la organización.

FB Viégas, en conjunto con otros investigadores, realiza dos trabajos de visualización sumamente interesantes basados en el correo electrónico. En sus trabajos [39] [40] genera principalmente dos tipos de visualizaciones que permiten recrear una historia a partir de correos electrónicos. El primer tipo de visualización es similar a una nube de palabras y el segundo tipo es una red de comunicaciones entre las personas (ver figura 6.7). En su visualización de red, los contactos se encuentran muy cercanos entre sí y es difícil diferenciar aquellos que cuentan con una baja frecuencia. Por otra parte, cabe destacar que la estrategia de visualización empleada por Viégas y sus compañeros permite la búsqueda por palabras claves.

Figure 1: *Screen shot of Themail showing a user's email exchange with a friend during 18 months.*

Figura 6.6 Visualización del correo electrónico en el tiempo [39].

Fig 1. *PostHistory interface with calendar panel on the left and contacts panel on the right. A contact name has been highlighted and the corresponding emails sent by this person have been highlighted in yellow on the calendar pane*

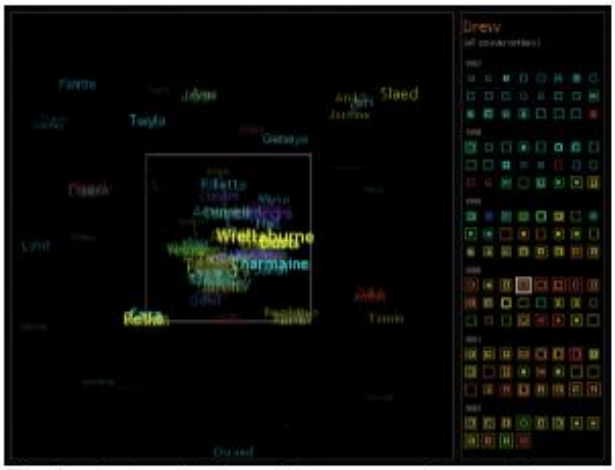


Fig 3. In Social Network Fragments, the social network panel is on the left while the history panel is on the right.

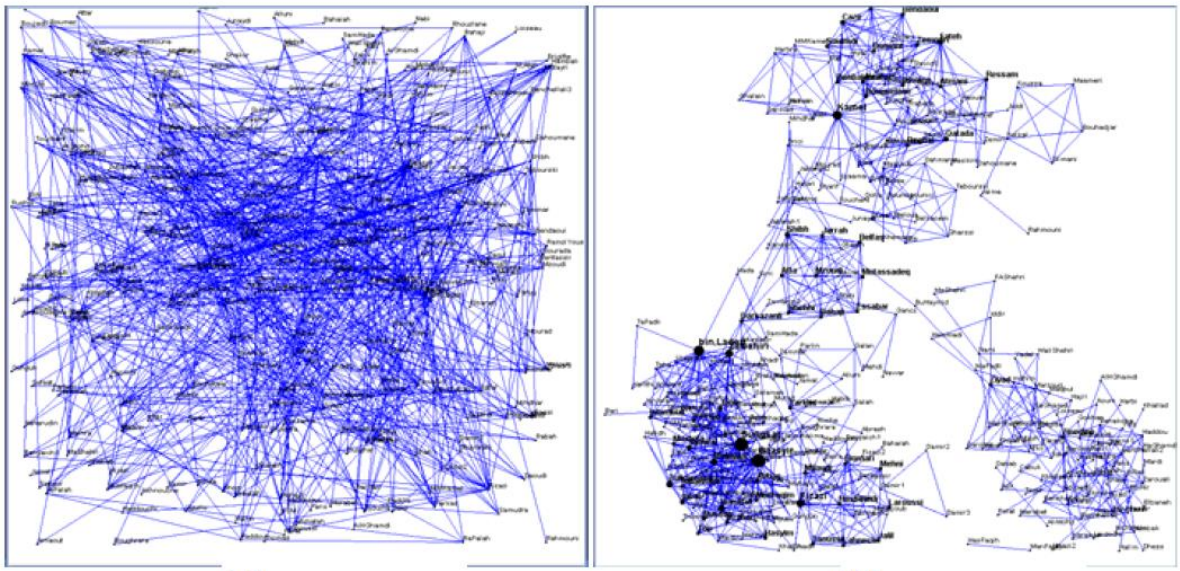
Fig 4. Zooming into the social network panel reveals the structure and the people who operate as bridges.

Figura 6.7 Análisis del correo electrónico para reconstruir una historia [40].

6.3 Estrategias de visualización

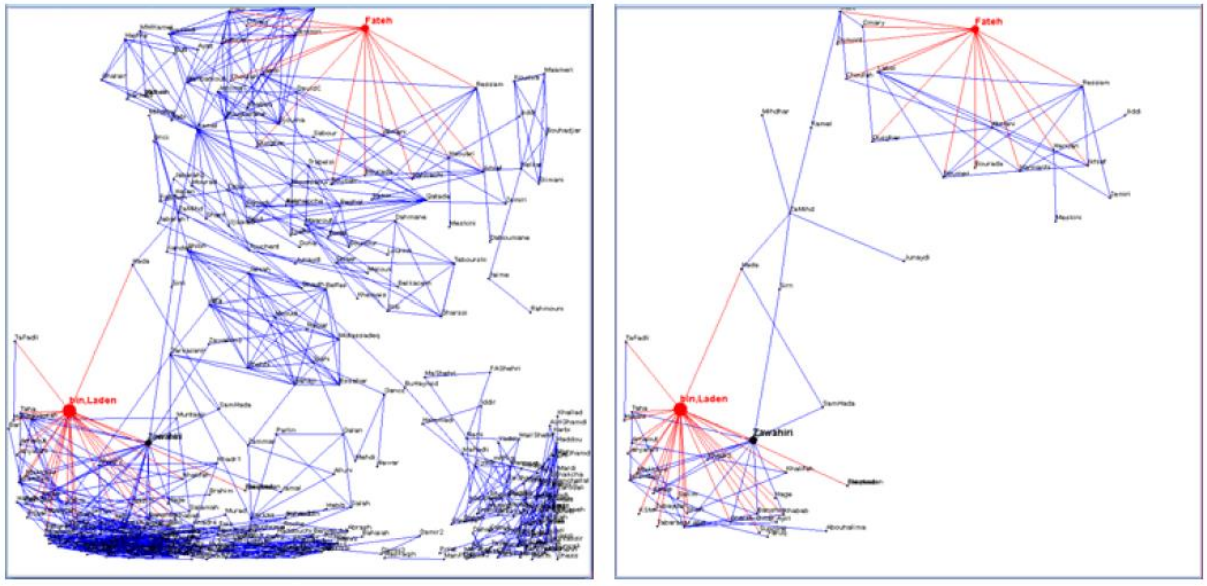
La última área de investigación previa en el que se fundamenta nuestro trabajo es la de las estrategias de visualización y las técnicas detrás de estas.

Si bien el trabajo previo de las estrategias de visualización es amplio [3] [5] [8] [15] [21] [35], nos basamos en una selección de trabajos afines sobre los cuales estaremos eligiendo el conjunto de técnicas de visualización fuertemente relacionadas con esta investigación. Dentro de esas cabe destacar el trabajo de CC Yang, N Liu y M Sageman, quienes emplean técnicas de F+C tales como el “ojo de pez” para visualizar redes de terrorismo [11].



(a) (b)

Fig. 1. (a) Initial Layout (b) Layout after applying the spring embedder algorithm



(a) (b)

Fig. 4. (a) Fisheye view with both Fateh and Bin Laden as focuses (b) Combined fisheye and fractal view with both Fateh and Bin Laden as focus and a fractal value threshold of 0.6

Figuras 6.8 Uso de análisis de redes sociales y la técnica ojo de pez (“fisheye”) para la identificación de líderes dentro de una red de terroristas [11].

Tal y como se describe en [50], la técnica de visualización de “ojo de pez” permite la ampliación del foco de atención a aquellos puntos que interesan sin mayor pérdida del contexto.

En el trabajo de Yang et al. [11], es evidente por medio de la visualización los puntos líderes de la red (puntos rojos dentro de la figura 4) sobre los cuales el foco de atención se centra (ver figura 6.8), sin embargo, existen métodos [28] que se creen pudieran ser más eficientes para resaltar los puntos de enfoque y permitir una mayor comprensión del objetivo meta de la visualización (ver figura 6.9). Por ejemplo, si se quiere focalizar la atención en los líderes de la red, puede decidirse no desplegar las conexiones de aquellas subredes que no tienen relación con los líderes.

Otra visualización [79] que hace una implementación interesante de la técnica de ojo de pez es este mapa (ver figura 6.10) que muestra las conexiones de las referencias bibliográficas realizadas entre publicaciones científicas, agrupas según su revista de publicación. La cercanía entre los nodos es dada por la cantidad de citas que existen entre las revistas de publicación. Entre mayor sea la cantidad de citas bibliográficas entre las revistas, más próximamente se encontrarán entre sí en la visualización. La implementación de ojo de pez resulta interesante ya que el ojo de pez es claramente visible por una circunferencia el cuál se puede desplazar por el usuario conforme el foco que éste le quiera dar a la visualización. A partir de esta visualización una vez más concluimos que el ojo de pez aporta a concentrar el enfoque del usuario en un punto dado sin la mayor parte del contexto del resto de la información, sin embargo, se

encuentra uno limitado a un punto de enfoque cuando lo que buscamos para nuestro objetivo puede involucrar múltiples puntos de enfoque.

Los estudios sobre F+C ("focus+context") en el área de la fotografía aportan una técnica de enfoque que en su trabajo Kosara define como SDOF (Semantic Depth of Field, por sus siglas en inglés) [28]. En dicha técnica, se difuminan los elementos que forman parte del contexto, y se les da claridad a aquellos elementos que forman parte del enfoque.

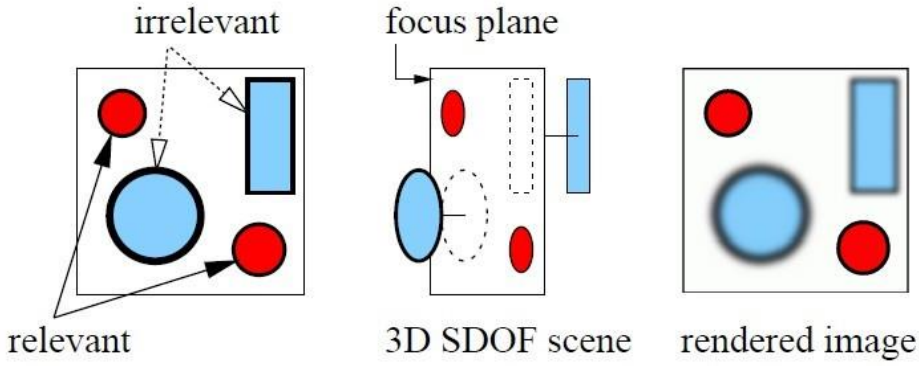


Figure 3: The basic idea of SDOF: applying DOF individually to scene objects depending on their semantics.

Figura 6.9 Técnica de "Semantic Depth of Field" para la aplicación de focus+context [28].

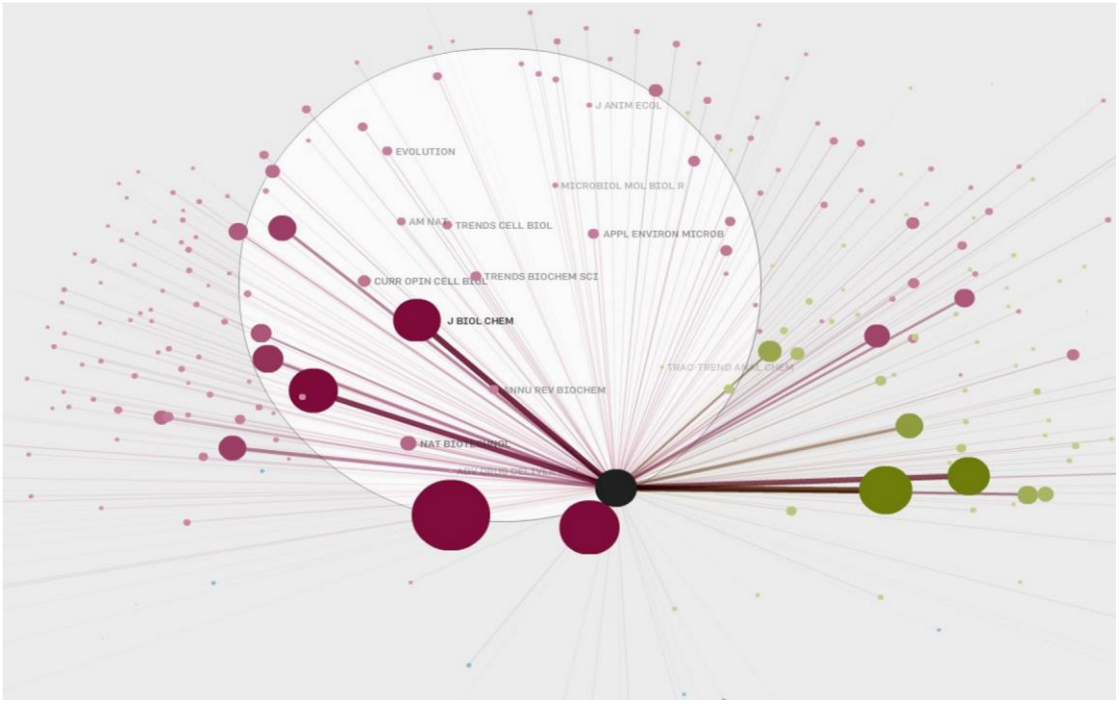


Figura 6.10 Fish-eye aplicado sobre red de referencias bibliográficas en revistas [79].

De aquellos paradigmas de visualización útiles para visualizar flujos y conexiones, se pueden estudiar y citar varios que giran alrededor del concepto de la infoesfera. La infoesfera es un concepto en sí, el cual describe como un todo, la mezcla de comunicaciones y datos provenientes de múltiples fuentes. Estos generan una gran red de conocimiento de todo el ecosistema que se alimenta la infoesfera [62] [71][77] [80].

El mapa de vientos [78] permite observar flujos claramente definidos, así como también concentraciones de estos. En las siguientes figuras (ver figuras 6.11 y 6.12) se pueden observar claramente los flujos de los vientos en dos momentos distintos en el tiempo y se puede observar las áreas sobre las cuales existe una mayor concentración de estos.

El mapa de vientos, así como otras visualizaciones por mencionar en este apartado, cuentan con animación que permite observar el flujo de los datos y la información a través del tiempo.

Figura 6.11 Mapa de vientos del 3 de Enero, 2013 según el sitio Wind Map [78]

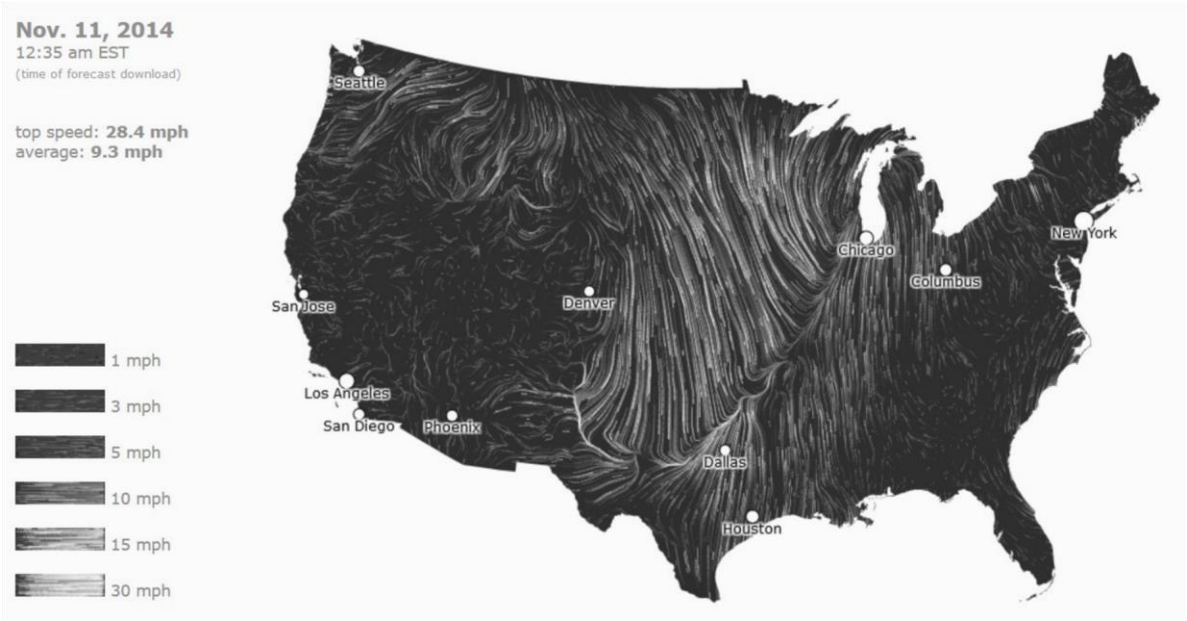


Figura 6.12 Mapa de vientos del 3 de enero, 2013 según el sitio Wind Map [78].

Otra alternativa de visualización de concentración de información se puede encontrar en el trabajo por Michael Kreil [73], en donde se visualiza la localización de teléfonos celulares utilizando técnicas de luminosidad y difuminación sobre un contexto de un mapa geográfico para poder observar la concentración y desplazamiento de

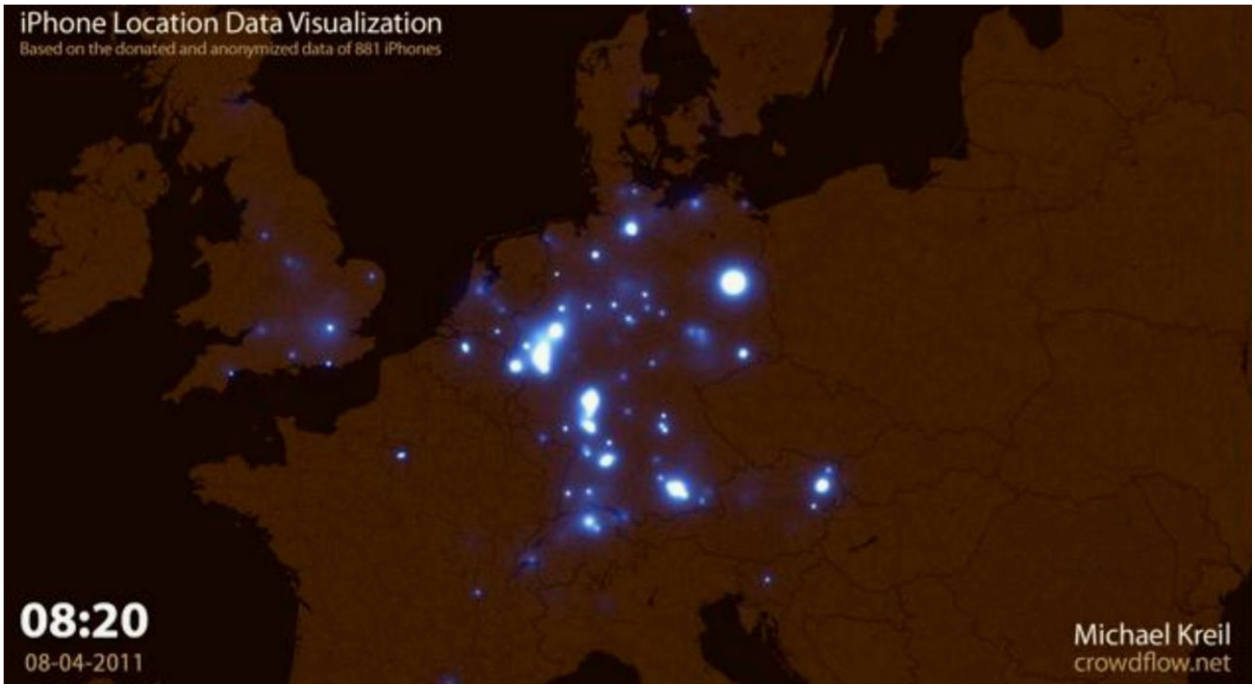


Figura 6.13 Fireflies – Visualización de localización de iPhones [73].

ochocientos ochenta y un celulares en Europa. De forma similar, el trabajo de Jeff Clark, de movimientos sobre Manhattan [53] [54] [55], permite visualizar concentraciones y movimientos de personas a partir de la localización de su celular, con la particularidad de que en este caso ocurre en el contexto de un mapa mucho más detallado. Esto agrega un gran valor para poder llegar a conclusiones concretas sobre la información que se muestra en la visualización (ver figura 6.14). La visualización de dicho trabajo se puede observar en un video en donde llama la atención las concentraciones de las personas conforme las horas de la noche avanzan. Al mirar el contexto, representado por el mapa,

se puede claramente observar que la concentración existe primordialmente en el área donde se encuentra ubicado el metro en Manhattan.

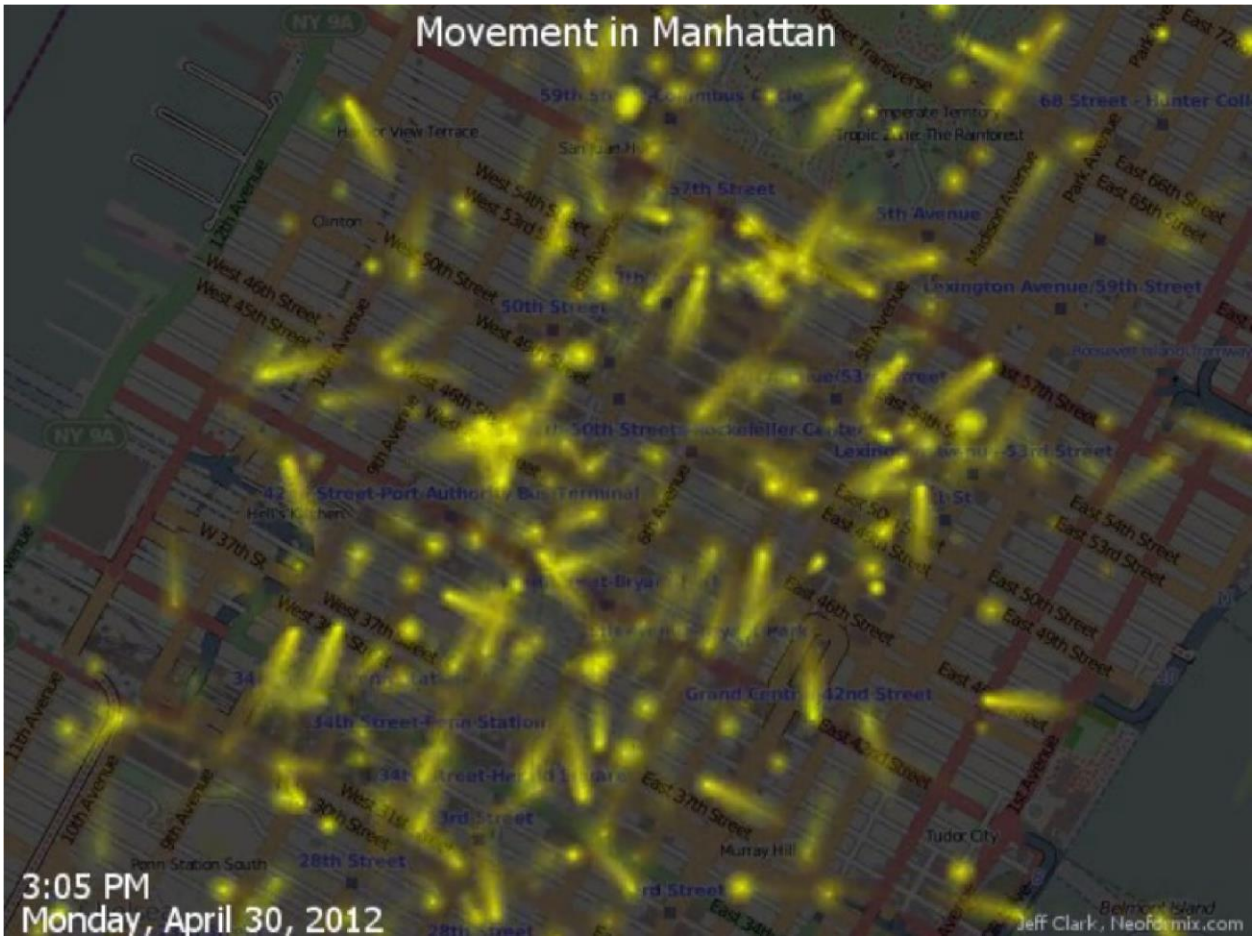


Figura 6.14 Movimientos en Manhattan visualizados a partir de la localización de tweets [54].

Figura 6.15 Colaboración e influencia en proyectos de código libre [83].

El trabajo de visualización de la colaboración y la influencia en la comunidad de código libre [83] permite analizar aspectos de colaboración en una comunidad establecidos por las relaciones en las actualizaciones de código en proyectos de código libre de gran relevancia, tales como rails, git, perl, entre otros (ver figura 6.15). La visualización de la colaboración y la influencia en la comunidad de código libre, permite observar las conexiones entre los distintos contribuidores de la comunidad en el contexto de su ubicación en el mundo. Al mismo tiempo, se pueden ver aquellas ubicaciones que cuentan con una mayor contribución a la comunidad. Un aporte interesante de este trabajo son los diagramas matriciales sobre los cuales se muestran las contribuciones a través del tiempo, agrupando la información por cuatrimestres por cada uno de los principales proyectos de código libre estudiados [83].

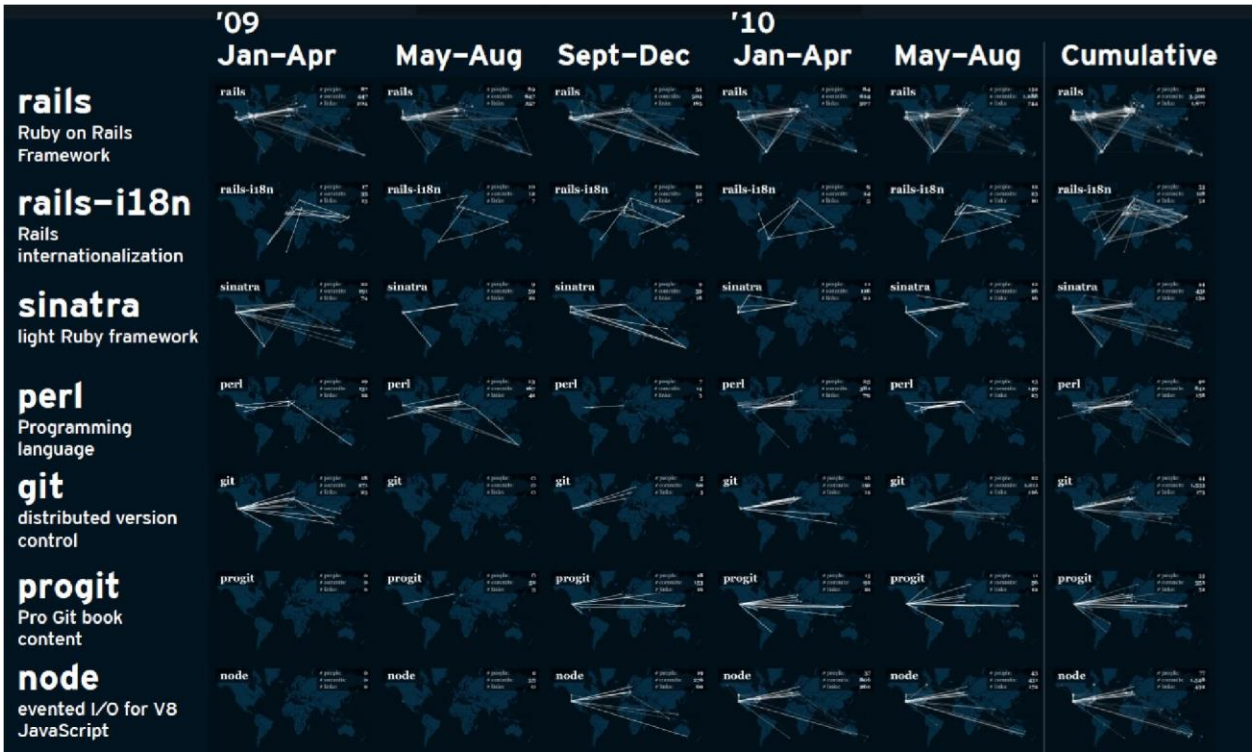


Figura 6.16 Evolución en el tiempo de contribuciones a proyectos de código libre [83].

De igual forma, existen otros paradigmas de visualización que pueden ser relevantes para nuestra investigación. Tal es el caso de los mapas de árboles, mejor conocidos como “treemaps” por su traducción en inglés. Los treemaps son un paradigma de visualización basado en el ordenamiento jerárquico de la información mediante estructurales de árboles visualizados sobre una representación plana. En su artículo Estrategias de Representación de Estructuras Jerárquicas [50], los profesores Franklin Hernández-Castro y Jorge Monge Fallas describen en detalle la composición y variaciones existentes alrededor de los treemaps (ver figura 6.17).

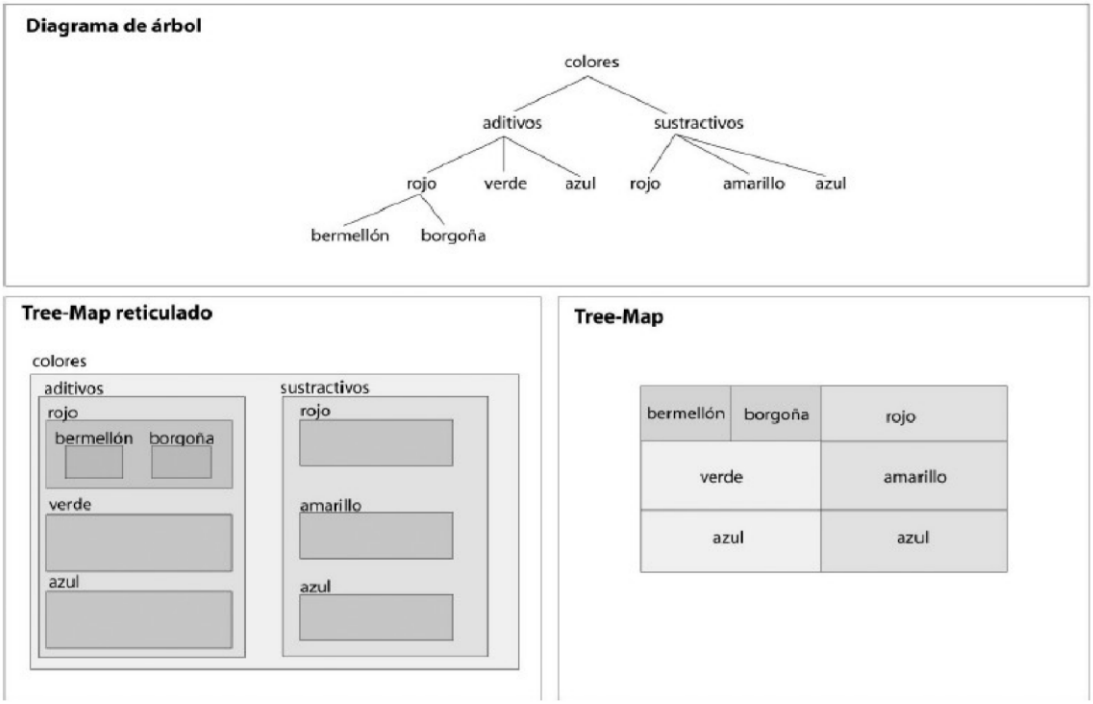


Figura 6.17 Treemaps desarrollados a partir del árbol base [83].

El paradigma de visualización de treemaps es limitado en el sentido de que no cuenta con conexiones. En implementaciones donde sí cuenta con conexiones [68], las relaciones son difíciles de visualizar cuando existe una gran cantidad de datos (ver figura 6.18).

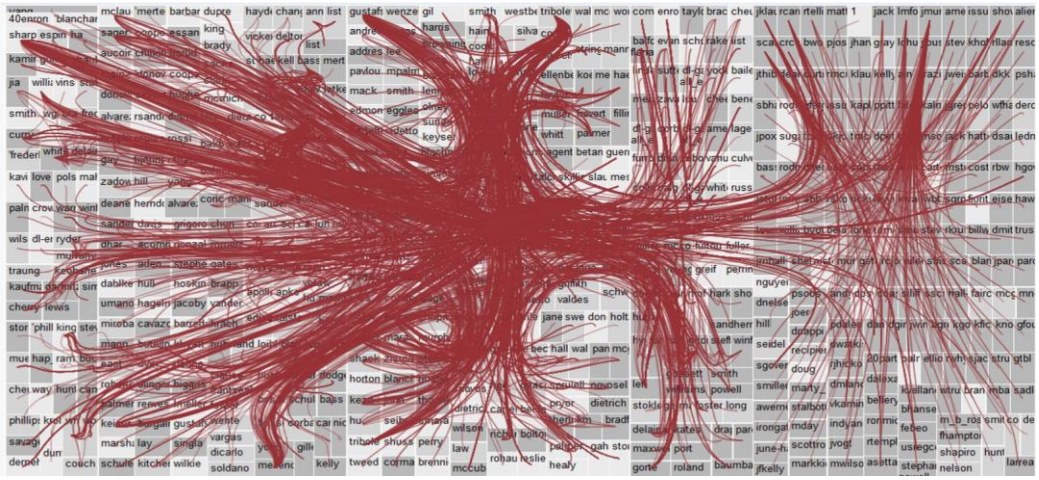


Figura 6.18 Evaluación de concepto de un treemap con conexiones.

7. Metodología

El proceso de investigación para la elaboración de este trabajo cuenta con un enfoque cualitativo y está compuesto de tres grandes etapas sobre las cuales se aplica continuamente un proceso de análisis, experimentación y mejora.

Tal y como se muestra en la figura 7.1, Análisis, Diseño e Experimentación, son las tres grandes etapas que ejecutadas en ese orden cubren aspectos de 3 grandes vectores, sobre los cuales se basa este trabajo:

1. Los datos: La proximidad de los datos con escenarios reales nos ayudarán a probar tanto la hipótesis como también la factibilidad de emplear el mecanismo de extracción de la información propuesto sobre correos corporativos reales.

2. Evaluación de paradigmas de visualización: En este vector se centra gran parte de la investigación de esta tesis. Se tomarán 3 paradigmas de visualización seleccionados por su proximidad en poder resolver el problema de investigación planteado en este trabajo. A cada uno de estos se le aplicará un mismo conjunto de métricas para finalmente poder hacer un contraste entre los diferentes paradigmas de visualización y así poder seleccionar aquel que tenga un mayor alineamiento con los objetivos planteados en este trabajo.

3. Propuesta final: Finalmente, basado en el paradigma de visualización seleccionado en el vector anterior, se analizarán aquellas áreas donde la visualización podría mejorarse y se presentará un prototipo final, el cuál será evaluado a la luz de un conjunto de casos de uso.

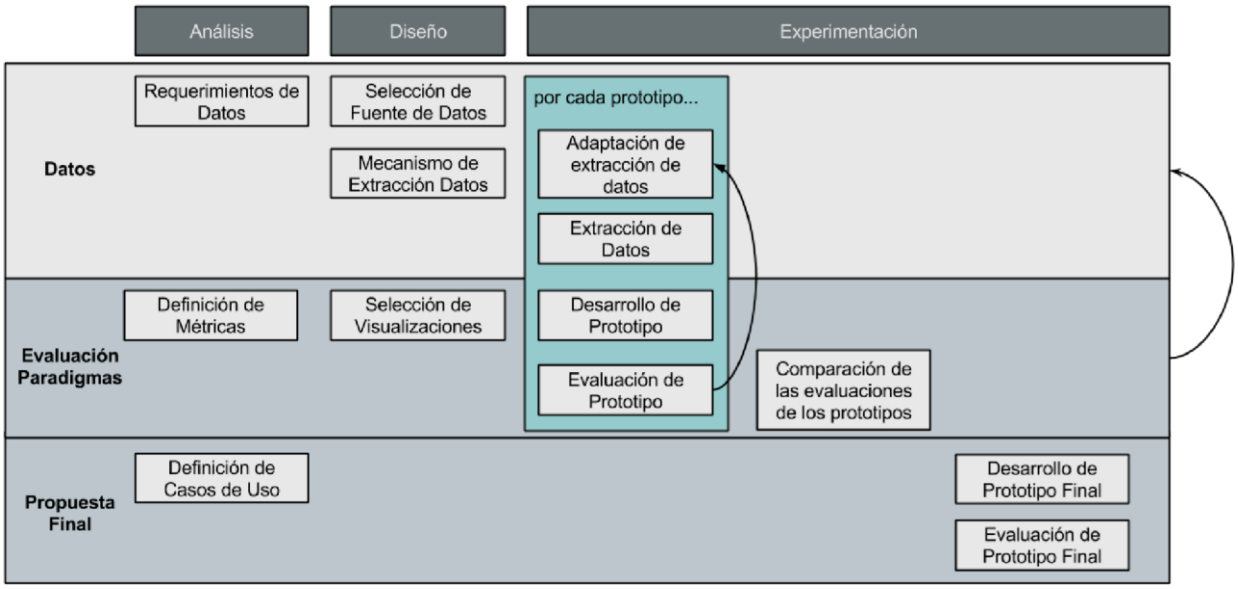


Figura 7.1 Ilustración gráfica de metodología.

A continuación se procede a describir en un mayor nivel detalle la metodología en el orden cronológico:

7.1 Análisis

En esta etapa se definen los requerimientos de la fuente datos con el objetivo de que estos sean lo más próximos a poder representar el correo electrónico de una corporación. Dentro de las características a tener en consideración se encuentran:

- El tamaño del correo corporativo: Este debe ser representativo al tamaño de un correo corporativo y así poder probar la efectividad del mecanismo de extracción de datos propuesto, con respecto a fuentes de datos de gran tamaño

- El formato de almacenamiento de los datos: En lo posible los correos deben existir en un formato estándar sobre el cual la mayoría de servidores de correos puedan exportar sus correos, y con esto permitir una mayor portabilidad de la solución propuesta.

- La privacidad de la información: Dado a que este trabajo es de carácter público, es de vital importancia poder resguardar cualquier aspecto de privacidad o bien de propiedad intelectual que pudiera existir en los datos a visualizar. Esta área fue identificada como un área de alto riesgo para la conclusión de este trabajo, de ahí que en el desarrollo se evaluarán distintas alternativas. Se escogerá aquella opción que mejor cubra las necesidades de privacidad y propiedad intelectual de la información.

En la etapa de análisis, se define un conjunto de métricas sobre las cuales se medirá la efectividad, de los paradigmas de visualización y de la propuesta final, en responder al problema planteado.

En el caso de la evaluación de paradigmas de visualización, se definirán 4 métricas que buscarán medir aspectos relevantes con los cuales los paradigmas de visualización debieran cumplir para poder resolver el objetivo de investigación. En el caso

de la propuesta final, esta será evaluada contra una rigurosa lista de diez casos de uso que ejemplifican escenarios reales a los cuales se espera se pueda dar respuesta y así poder concluir si la hipótesis es cierta.

7.2 Diseño

En la etapa de diseño se hace un análisis de las posibles fuentes de datos y se selecciona aquella fuente que sea más apropiada para la investigación según los requerimientos definidos en la etapa de análisis. Analizaremos la fuente de datos y escogeremos un conjunto de palabras claves que representen un tema, las cuales no predominen como un tema común o ambiguo. Dichas palabras claves representarán nuestro tema de interés de estudio durante toda la investigación.

Una vez definida la fuente de datos, se procede a diseñar y desarrollar una solución, con la cual se extraigan los datos para la evaluación de los distintos paradigmas de visualización. Dicho mecanismo de extracción y transformación de los datos se irá modificando conforme se vayan analizando variaciones sobre cada uno de los paradigmas de visualización durante la etapa de desarrollo.

En la etapa de diseño, se escogen los tres paradigmas de visualización a evaluar. Se hace un análisis de las fortalezas de cada uno, que expliquen el porqué de su selección como un paradigma de visualización que potencialmente podría responder a los problemas planteados en esta investigación.

7.3 Experimentación

Tomando las métricas definidas en la etapa de análisis, se definen y evalúan los distintos paradigmas de visualización. Del análisis de las evaluaciones, se proponen e implementan mejoras sobre la visualización más apta para resolver el problema planteado. La propuesta final será analizada a la luz de los casos de uso definidos en la etapa de análisis.

El proceso de experimentación comienza por comprender el formato de los datos que requiere el paradigma de visualización a evaluar y el modificar el mecanismo de extracción de acuerdo con este. Conforme se vayan extrayendo los datos y se vayan probando los distintos paradigmas de visualización, se irán incorporando diferentes variantes conforme los paradigmas de visualización lo permitan, y es conforme a estas variantes que se evaluará cada paradigma contra las cuatro métricas definidas en la etapa de análisis.

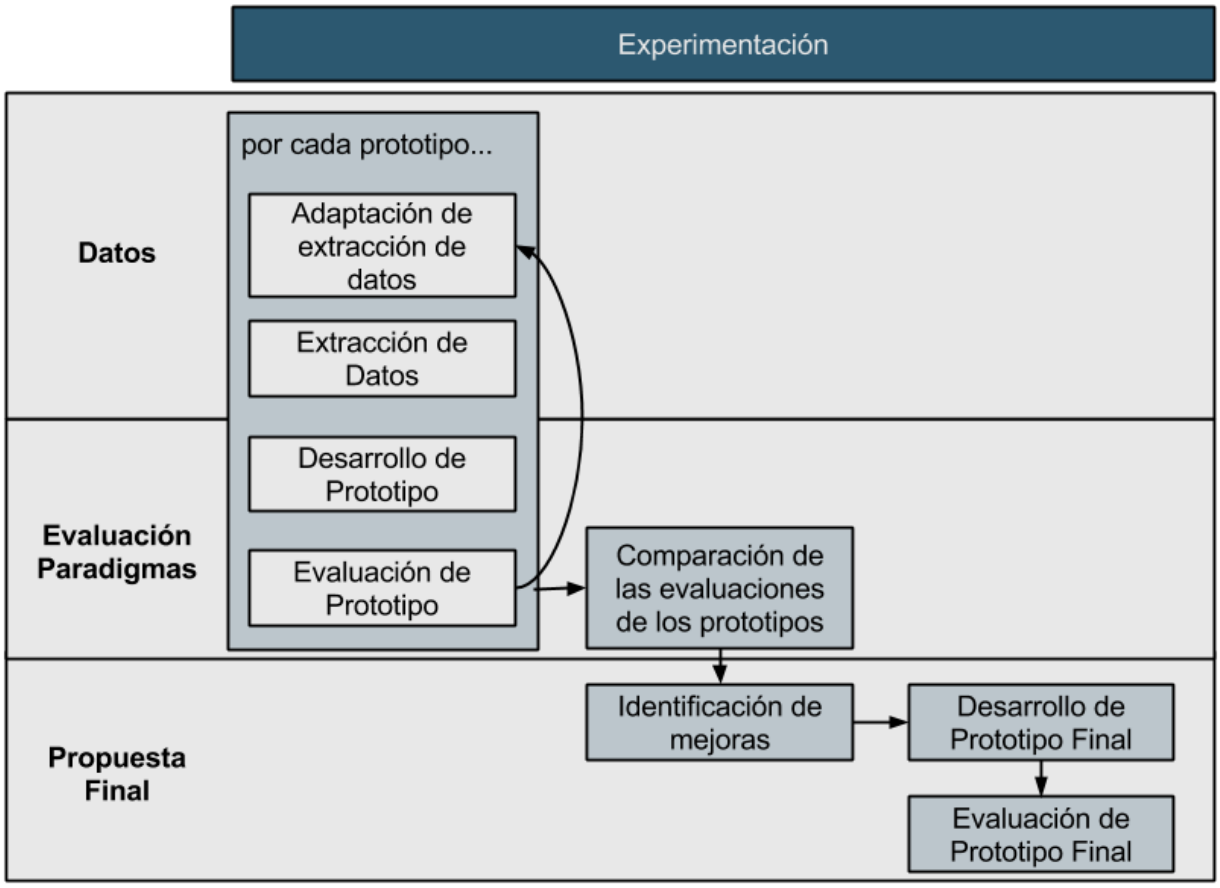


Figura 7.2 Proceso de experimentación.

De los paradigmas de visualización evaluados, se hará un contraste de las evaluaciones hechas sobre cada uno y se escogerá el paradigma de visualización más apto según el análisis sobre las métricas. Una vez escogido aquel paradigma que mejor cumple las métricas, se hará un análisis de las áreas a mejorar, en pro de lograr un mayor alineamiento con los objetivos de investigación inicialmente planteados. Dicho paradigma de visualización será finalmente analizado a la luz de un conjunto de diez casos de uso que permitirán concluir el resultado de la hipótesis.

8. Desarrollo

8.1 Requerimientos de Datos

Para la selección de la fuente de datos se han definido un conjunto de requerimientos mediante los cuales se analizará cuál de las distintas opciones es más conveniente para desarrollar esta investigación. A continuación se detallan dichos requerimientos:

1) El volumen de datos: Se busca un set de datos que comprenda por lo menos 1 GB de datos, de tal forma que se compruebe que el mecanismo de extracción y análisis de los datos es capaz de procesar grandes cantidades de información. Esta característica permitirá probar atributos del método de extracción de datos a desarrollar, los cuales solo pueden ser evaluados contra grandes cantidades de datos.

2) El formato de almacenamiento de los datos: Al inicio de este trabajo se evaluó hacer uso del formato .pst (Personal Storage Table, por sus siglas en inglés) [76]. PST era el formato del correo electrónico corporativo sobre el cual se comenzó a elaborar la idea de esta investigación. Las pruebas iniciales para tomar los datos básicos del .pst usando librerías de Outlook sobre una implementación sencilla Microsoft .Net rápidamente mostraron un desempeño muy deficiente de tanto la librería como también la aplicación de Outlook, que rápidamente iba incrementando su memoria RAM reservada conforme la cantidad de registros que eran leídos. Por su mal desempeño y

debido a que el formato .pst no es abiertamente adoptado por la mayoría de servidores de correos, decidimos evaluar un formato estándar sobre el cual se pudieran exportar de forma sencilla la mayor cantidad de servidores. Es por esto que se decidió la selección del formato MIME [75], el cual está desarrollado sobre el SMTP [51]. Mediante el uso del formato MIME, se puede asegurar una gran portabilidad de la solución sobre los diferentes tipos de servidores de correos electrónicos en la industria, los cuales en su gran mayoría hacen uso de SMTP/MIME.

3) La privacidad de los datos: Lo ideal para este trabajo es que el set de datos no sea un riesgo de privacidad de las personas involucradas en los correos. Entre las alternativas como fuente de datos se encuentran:

- Datos simulados: Una de las opciones que se analizó fue el generar datos simulados para hacer la investigación. La idea sería generar un conjunto de datos que representaran comunicaciones aleatorias de un grupo de personas fuente a un conjunto de personas destino, con palabras que formasen títulos aleatorios, a partir de un conjunto de palabras previamente definidas. Dicha opción no se desestimó por completo, sin embargo, se dejó como la última opción en caso de que no se lograra acceso a un correo corporativo.

- Datos abiertos: La búsqueda por una fuente de datos abierta comenzó a partir de fuentes de datos utilizadas en los trabajos descritos en el estado del arte [83], las cuales hacen referencia primordialmente al estudio de listas de distribución públicas tales como aquellas empleadas por grupos de código abierto o grupos

de investigación que comparten un interés común. Dicho tipo de datos no era el ideal para este trabajo de investigación, debido a que las comunicaciones eran muy relacionadas entre sí. Las listas de correo públicas pierden la naturaleza y la variedad de comunicaciones que se pueden encontrar en un correo corporativo. En la búsqueda de una lista de distribución pública con una gran variedad de temas de discusión, se descubrió la base de datos Enron [60]. Dicha base de datos pertenece a un gran escándalo financiero que salió a la luz en octubre del 2001 [59]. Como parte de las investigaciones se generó una base de datos la cual luego fue comprada para ser hecha pública con fines de investigación [87]. Para esta investigación se usará el set de datos de Enron publicado en el 2011 [60]. Dicha versión ha contado con varios procesos de depuración, para eliminar así aquellos correos de personas quienes han solicitado que sus correos no sean parte de dicha base de datos.

8.2 Selección de Fuente de Datos

Tal y como se mencionó en el enunciado previo, la base de datos más apropiada y por ende seleccionada para realizar esta investigación es la base de datos Enron. Dicha base de datos está representada por archivos planos en formato MIME organizados de la forma en que se ilustra en la Figura 8.1. Los archivos suman en conjunto 1.31 GB y se encuentran compuestos de 512 334 correos.

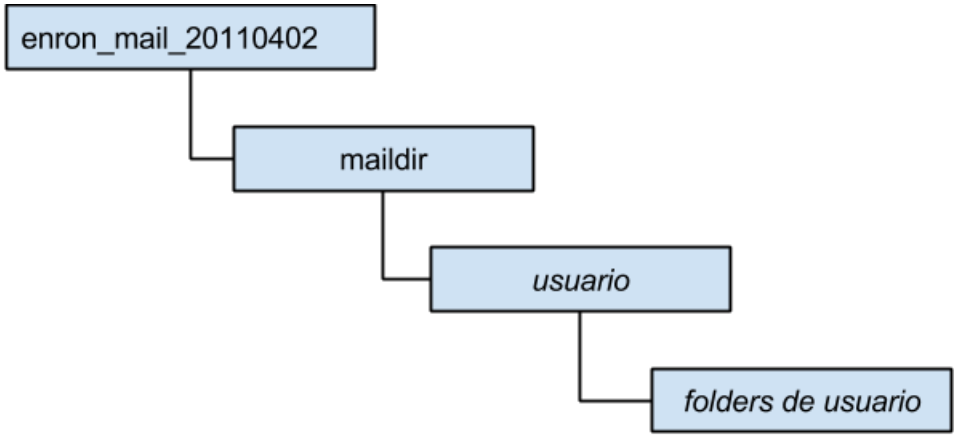


Figura 8.1 Organización del correo Enron en formato MIME.

Al ser la base de datos Enron ampliamente usada en investigaciones de análisis de redes sociales, se entiende que no existe un riesgo de privacidad para las personas involucradas en los correos. Dichos datos son públicos y han contado con varios procesos de depuración para eliminar así aquellos correos de personas quienes han solicitado que sus correos no sean parte de esta base de datos.

De esta forma, se cumplen con los requisitos previamente definidos de tamaño de la base de datos, formato y las consideraciones de privacidad, con los cuales se esperaba cumplir.

8.3 Mecanismo de extracción de datos

Se implementó por medio del lenguaje de programación Java un mecanismo de extracción de la base de datos Enron. Dicho mecanismo asume que los correos existen en un conjunto de directorios de forma jerárquica y que cada correo existe en un formato tipo MIME que contiene palabras claves estándar con las cuales se recorre el archivo

buscando el remitente, el destinatario o los destinatarios y el título del mensaje. Es importante aclarar que, para efectos de esta investigación académica, únicamente se hace uso del título de mensaje del correo electrónico ya que con eso se obtiene la información requerida para hacer la investigación y poder así probar la hipótesis.

La esencia del diseño del extractor de datos es permitir la lectura del programa de forma secuencial, tomando un archivo a la vez, almacenando y resumiendo los datos que interesan, de cada línea que se va leyendo. De esta forma, se puede garantizar que el extractor de datos funcionará para procesar correos corporativos de gran tamaño. Aún y cuando dicho procesamiento no ocurre de forma paralela, la esencia está inspirada en el concepto de map-reduce [48][47][49].

El algoritmo general empleado sigue una estructura como la mostrada en la figura 8.2.

```
función LeaListaDeCorreosEnFolders
    Obtenga lista de archivos o Folders de Posición de directorio

    Para cada archivo : En la Lista de Archivos por leer
        Si el archivo es un archivo {

            Lea Linea de Archivo  Mientras se pueda leer el archivo{

                Si la lectura empieza con "From:"
                    Remitente = Almacene dirección de quién envía el correo

                Si la lectura empieza con "To:"
                    Lista de Direcciones de Destinatarios <= Almacene todas las direcciones a las cuales es enviado el correo

                Si la lectura empieza con "Subject:" y contiene una de las palabras clave

                    Por cada Destinatario
                        Almacene el Remitente en la colección de nodos si no existe
                        Incremente en uno los correos enviados por el Remitente en la colección de nodos

                        Almacene el Destinatario en la colección de nodos si no existe
                        Incremente en uno los correos enviados por el Destinatario

                        Incremente en uno las conexiones entre el Remitente y el Destinatario
            }
        }

        Si el archivo es un directorio
            Llame de forma recursiva esta misma función con el directorio para proceder a leer sus archivos

}
```

Figura 8.2 Lógica de procesamiento de los correos electrónicos en formato MIME.

El extractor de datos tiene como entrada el directorio principal corporativo, en donde se encuentran los correos, contenidos en subdirectorios de usuario. La rutina principal del extractor producirá primordialmente dos listas de información. La primera es una lista de nodos, la cual describe las distintas personas que enviaron o bien recibieron un correo. Para cada nodo (persona) se va almacenando el volumen de correos enviados o recibidos. La segunda salida de la rutina es una lista de conexiones, la cual almacena el par remitente – destinatario y guarda un peso según la cantidad de correos enviados, que se llama frecuencia. Sobre estas salidas se producen entonces el set de datos que las distintas visualizaciones consumirán.

8.4 Definición de temas de interés

Una vez seleccionada la fuente se procede a mapear aquellos temas de interés que definirán el tema con que se realizará la investigación para probar la hipótesis.

Primeramente se procedió de hacer un análisis de aquellas palabras con una muy alta frecuencia entre los correos de la base de datos Enron. Basados en los resultados de dicha investigación, procedimos a realizar una visualización de nube de palabras para descartar aquellas palabras con una muy alta frecuencia en el correo. Para producir la visualización se usó la aplicación en línea tagxedo.com.

En la nube de palabras (ver figura 8.3) se puede observar palabras que forman parte esencialmente de dos grandes grupos:

1. Palabras normalmente usadas en cualquier corporación, tales como: Reunión, Título, Acuerdo, Actualización, Reporte, Solicitud, Problema, Conferencia, Aprobación, etc.
2. Palabras comunes de la corporación en sí: Enron, Gas, Energía, Poder, CID, FERC, etc.

Figura 8.3 Visualización “Nube de palabras” de las palabras comúnmente usadas en la base de datos Enron.

Por ser palabras comúnmente usadas, se evita el uso de estas y se ha procedido a definir tres conjuntos de palabras que pudieran representar buenos temas sobre los cuales buscar la existencia o la no existencia de conocimiento.

1. Conjunto de datos #1: Apple, IPOD

   En octubre 23 del 2001 Apple pone a disposición del mercado la primer línea de Ipod [72]. Esto ocurre en un momento en donde la marca Apple no era tan relevante en la industria. Es por esto que se considera de relevancia evaluar las personas que en ese momento tenían algún conocimiento de causa con respecto a Apple.

   Estructura de la red: 110 nodos, 100 conexiones únicas.

2. Conjunto de datos #2: "c++", "lisp", "clojure", "erlang", "prolog", "java"

Se tomó una lista de lenguajes de programación para tomarlos como parte de un conocimiento de programación, y dado a que los programadores solo son un subconjunto pequeño de la población de una corporación, interesa conocer quiénes en sus comunicaciones han mencionado dichos lenguajes.

Estructura de Red: 48 nodos, 43 conexiones únicas.

3. Conjunto de datos #3: "bribe", "suborn", "corrupt", "ilegal", "bribery", "fraud", "bankruptcy", "Rawhide", "SPE"

   Se definieron un conjunto de palabras relacionadas a corrupción y a instrumentos financieros empleados en la manipulación del estado financiero de la corporación. Se definieron sinónimos o palabras relacionadas a la palabra “corrupción” y se usaron instrumentos financieros. Dichos instrumentos financieros son SPE (“Special Purpose Entities”, por sus siglas en inglés), con los cuales la corporación ocultaba pérdidas y “Rawhide”, el cual fue uno de los últimos SPEs, cuando las prácticas de manipulación financiera salieron a la luz pública [61] [58].

   Estructura de Red: 671 nodos, 968 conexiones únicas.

Por la cantidad de nodos y conexiones con que cuenta el conjunto de datos número 3, el cual es el relacionado al tema de corrupción, se procede con el uso del mismo para efectuar gran parte del resto de la investigación. En algunas de las evaluaciones, puede que se haga uso de otros datos, como fin ilustrativo y de experimentación de casos que se quieren analizar.

8.5 Definición de Métricas de Medición

A continuación se procede a detallar aquellas métricas de medición sobre las cuales se analizarán los tres paradigmas de visualización aún por seleccionar. Se definirán cuatro métricas, cada una enfocada a valorar un área clave de la visualización y su aporte en resolver el problema de investigación.

8.5.1 Métrica #1 - Conexiones

Retomando la hipótesis y los objetivos planteados, interesa conocer en principio quién habla con quién. De ahí surge la primera métrica: Las Conexiones.

Las conexiones – Para efectos de esta investigación interesa conocer quién habla con quién en el plano individual. Al mismo tiempo interesa conocer la dirección de la comunicación. Dicha dirección puede contar con dos formas, la primera sobre la cual el mensaje es “enviado” y la segunda sobre la cual el mensaje es “recibido”.

8.5.2 Métrica #2 – Co-localización

Al mismo tiempo, interesa conocer qué personas pertenecen a un mismo grupo. Dicho grupo podría ser definido como un grupo que comparte un espacio físico, un grupo conformado por un departamento, o bien, una mezcla de ambos. La métrica que dicta el conocimiento de que dos o más personas pertenecen al mismo grupo se ha denominado con el nombre de: Co-localización.

- **Co-localización** – Es posible visualizar en una población a aquellas personas que pertenecen a un mismo grupo (localización) y poder distinguirlos claramente de otras personas pertenecientes a otros grupos. Es mediante esta métrica que se podrá comprender en la visualización qué persona pertenece a qué grupo y la distribución de su población en relación con otros grupos. Por ejemplo, la colocalización podría ayudar a comprender si el expertise de un tema se centra sobre un grupo de la corporación o si se distribuye entre varios grupos.

8.5.3 Métrica #3 – Frecuencia

El otro aspecto que interesa comprender en la visualización de las comunicaciones de los correos electrónicos es con qué frecuencia se hablan dos personas sobre un tema específico. El aspecto de frecuencia es clave para poder discriminar comunicaciones que rara vez han ocurrido, de aquellas que son frecuentes y constantes en el tiempo. Entre mayor sea la frecuencia de las comunicaciones de una persona sobre un tema, mayor es la probabilidad de que la persona tenga conocimiento relevante sobre el tema. A dicha métrica se le ha llamado: Frecuencia.

- **Frecuencia** – Es poder visualizar la frecuencia con que se habla de un tema específico, ya sea entre dos personas o en el plano general de las comunicaciones de la población. Interesa conocer en qué medida se comparte sobre un tema y poder así focalizar la atención a aquellas comunicaciones que contengan una

mayor frecuencia que pudieran demostrar grupos lógicos altamente relacionados entre sí en el intercambio de información sobre un tema

8.5.4 Métrica #4 - Relación Tema-Localización

Por último, y uno de los aspectos claves que diferencia el presente trabajo de otros trabajos previos, es el poder comprender que personas hablan entre sí, en el contexto de su localización dentro de una corporación. Dicho aspecto es fundamental para entender la sinergia de los miembros de un grupo, así como también la sinergia entre los grupos de una corporación. Por otra parte, se podrá visualizar la distribución de las comunicaciones sobre un tema, entre los diversos grupos corporativos. A dicha métrica se le ha dado el nombre de: Relación Tema-Localización.

- **Relación Tema-Localización** – Visualizar la relación tema/localización es la propiedad de visualizar la relación entre un tema y la localización de donde este tiene presencia. Comprende así aspectos de colaboración, sinergia e islas de comunicación, entre otros.

8.6 Definición de paradigmas de visualización

Para el proceso de evaluación de paradigmas de visualización se definieron tres paradigmas por su gran afinidad con los objetivos planteados. A continuación se procede a realizar un corto estudio de dichos paradigmas.

8.6.1 Circos

La visualización Circos se desarrolló como parte de una herramienta de visualización para el análisis de genomas [74]. Su autor, Martin Krzywinski, basó el desarrollo de CIRCOS en un trabajo previo de su autoría llamado Schemaball [42] para visualizar relaciones entre bases de datos.

El Circos es una composición circular pensada en poder visualizar grandes cantidades de datos a la vez. En su forma más básica, el Circos se construye a partir de un conjunto de nodos, los cuales se distribuyen en la circunferencia. Las conexiones de la visualización se forman por medio de una matriz, la cual define las frecuencias de las conexiones entre los nodos.

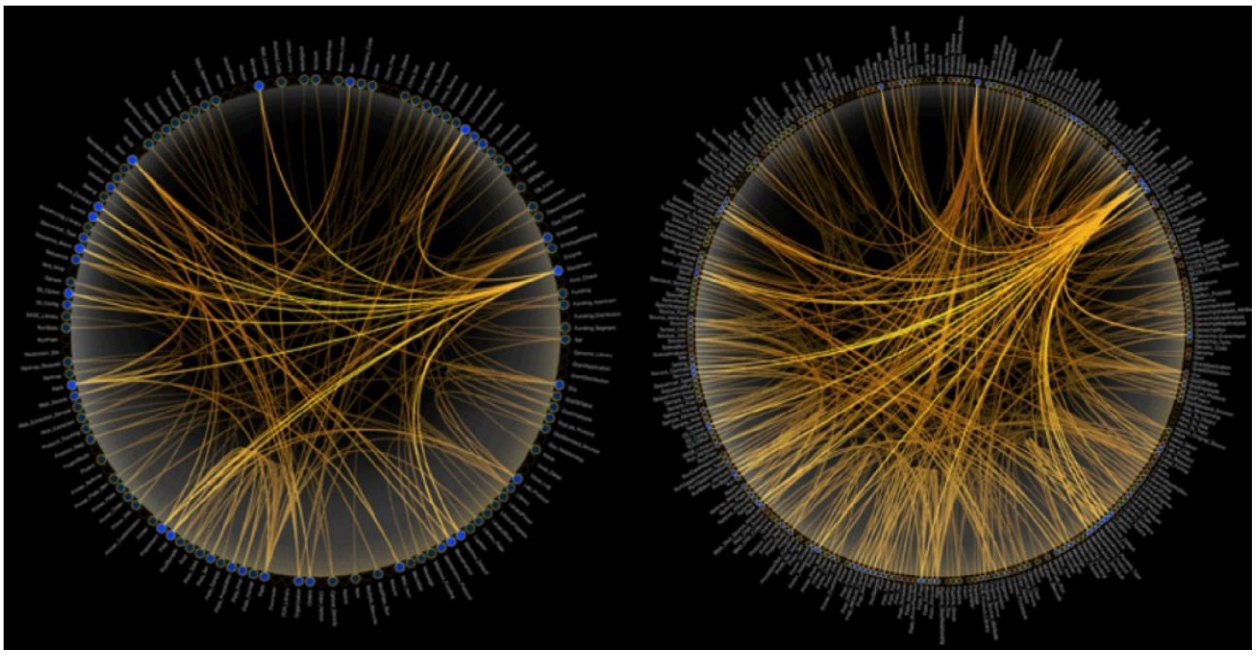


Figura 8.4 Visualización Schemaball [42], el cual es un trabajo base de la visualización Circos [74].

Gracias a una publicación por parte del New York Times titulada “Close-Ups of the Genome, Species by Species by Species” por David Constantine el 22 de enero del 2007 [43], es que la visualización y la herramienta de Circos comienzan a popularizarse y su uso comienza a expandirse a otras áreas.

La visualización del paradigma Circos es de gran trascendencia para aquellos problemas sobre los cuales se quieren ver secuencias en grandes cantidades de datos y patrones que pudieran resultar a partir de las conexiones. En el caso del problema planteado en este trabajo, se busca comprender las conexiones de comunicación entre las personas y determinar potenciales patrones que permitan distinguir la localización del conocimiento.

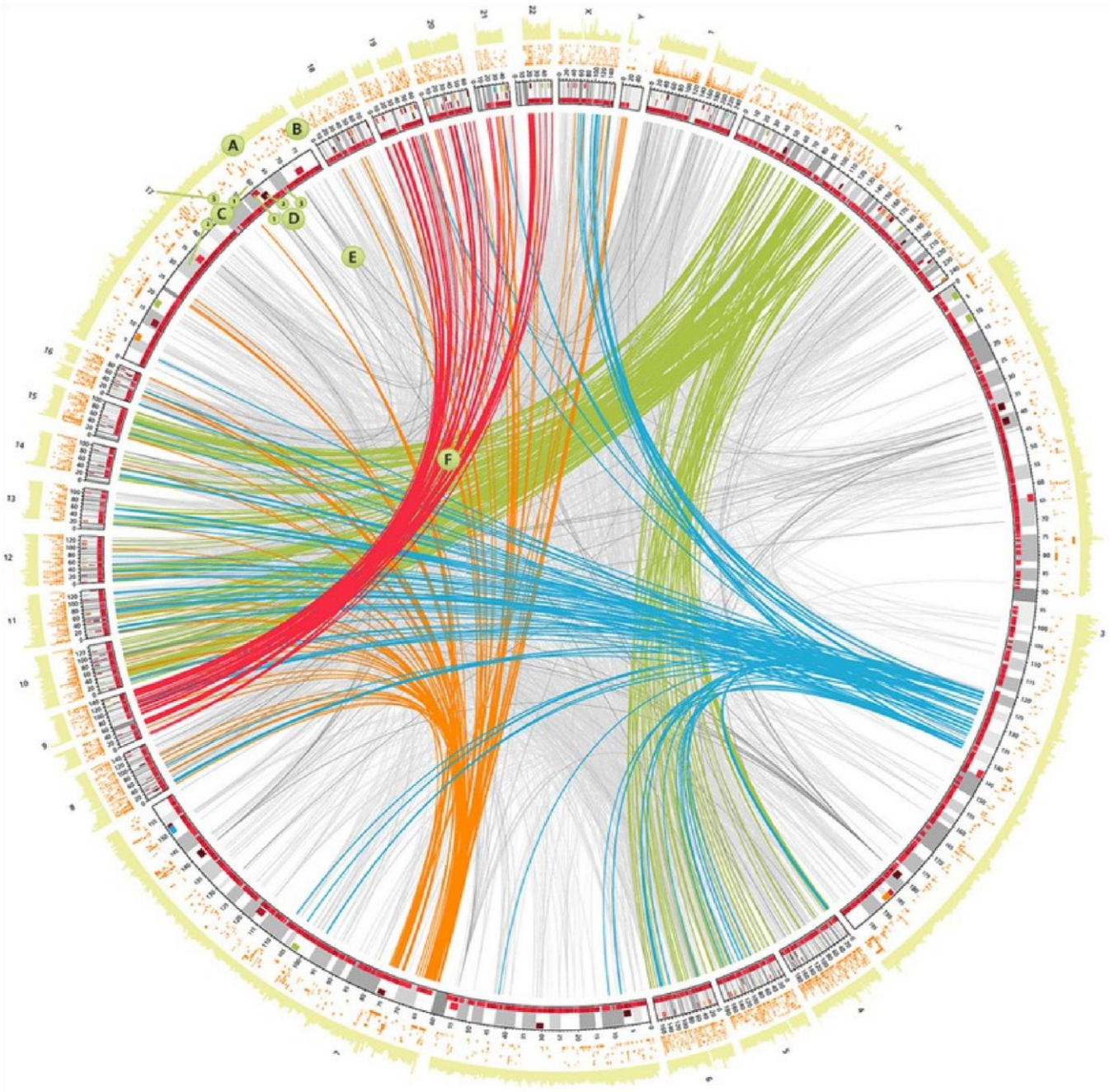


Figura 8.5 Paradigma de visualización Circos usado para el análisis de genomas [74].

8.6.2 Gmap

Gmap [44] es una visualización desarrollada por el equipo de investigación de AT&T en conjunto con la Universidad de Arizona con los objetivos primordiales de: a) poder visualizar grandes redes de datos y b) presentar la información de una forma natural que un público no técnico pudiera interpretar fácilmente.

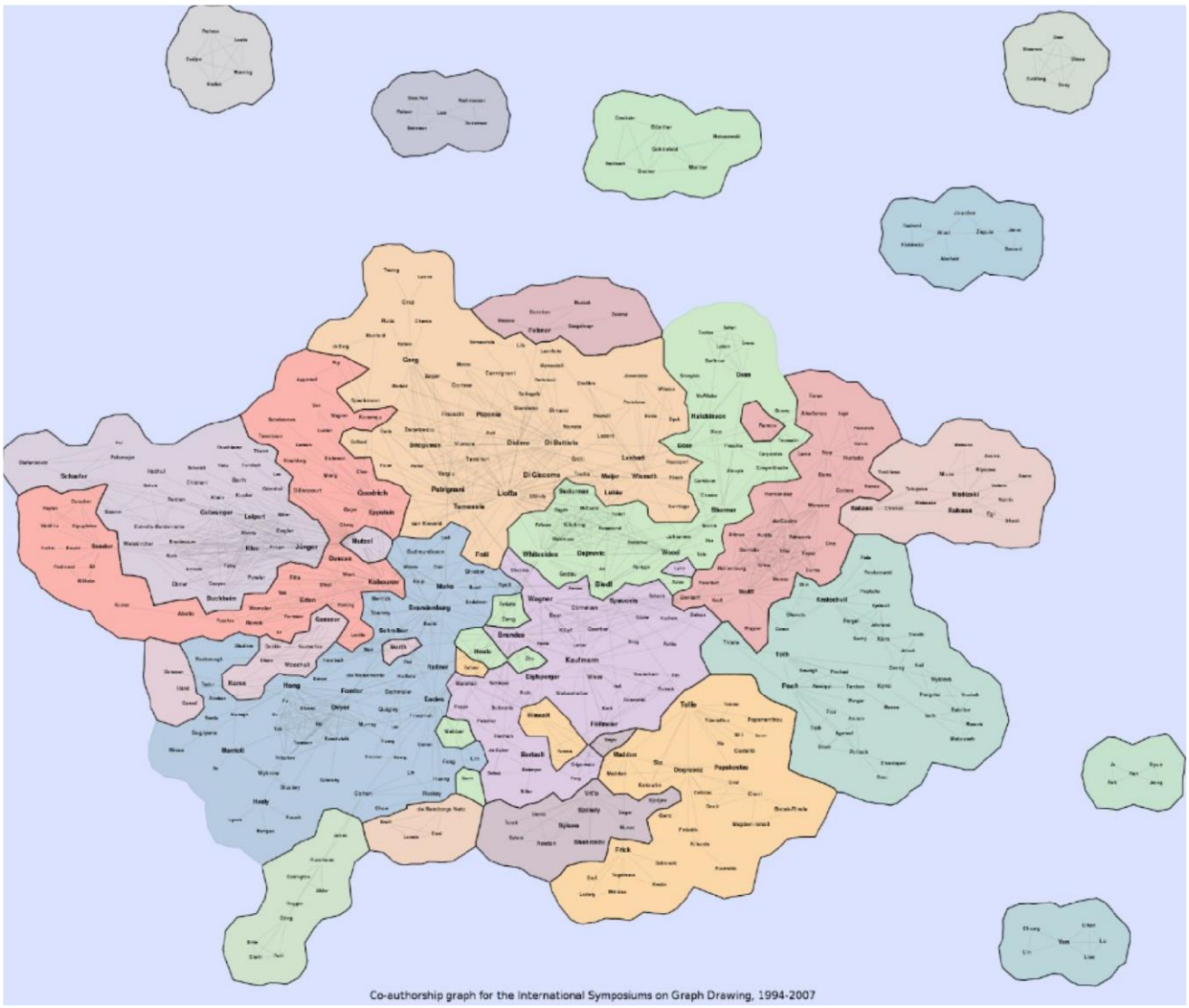


Figura 8.6 Gmap dibujado por medio de graphviz [46].

Gmap es desarrollado como parte de la herramienta de visualización graphviz y luego una versión de este fue habilitada en línea por la Universidad de Arizona junto con cambios en los algoritmos de agrupamiento para producir grupos menos particionados y de esa forma permitir una lectura más intuitiva de la visualización [45] [89][90].

Este paradigma de mapa de relaciones es relevante para esta investigación ya que muestra un mapa lógico de las relaciones de todos los miembros en una población, y permite de forma muy visible distinguir los grupos de cada uno de los miembros, así como también su distribución en el mapa lógico de relaciones.

8.6.3 Grafo

El Grafo es escogido como un paradigma de visualización a evaluar por ser la base de las visualizaciones desde los inicios del análisis de redes sociales y ser aún el método más empleado en el estudio y la exploración de redes sociales.

Fundamentado sobre la teoría de grafos, los grafos son una excelente forma de poder observar las relaciones entre los miembros de una población. Tal y como se analiza en el apartado de antecedentes en este trabajo [41][1][19][25], tenemos evidencia de que la teoría de grafos es ampliamente utilizada en el análisis de redes sociales orientadas a comprender la comunicación y la colaboración en organizaciones.

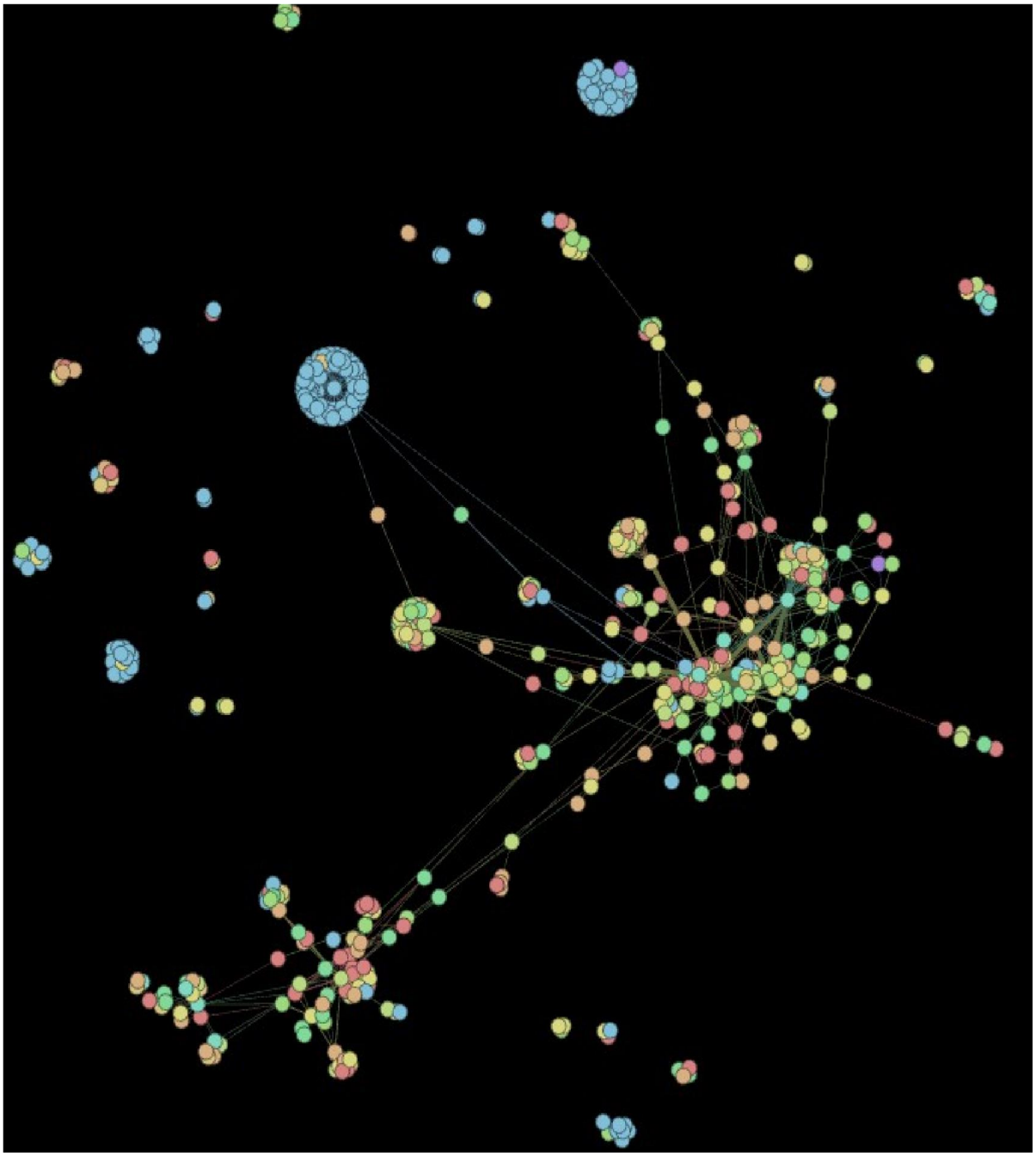


Figura 8.7 Prueba de concepto de grafo producido con Gephi [34]

8.7 Evaluación de Paradigmas de Visualización

8.7.1 Evaluación de métrica #1 – Conexiones

8.7.1.1. Hipótesis:

- Es posible conocer quién habla con quién en el plano individual.
- Es posible observar la dirección de la comunicación. Dicha dirección puede contar con dos formas, la primera sobre la cual el mensaje es “enviado” y la segunda sobre la cual el mensaje es “recibido”.

8.7.1.2 Observaciones de “conexiones” sobre el paradigma Circos

En el paradigma Circos existe una clara visualización de las conexiones (ver figura 8.8). Se puede visualizar tendencias, patrones y relaciones. Las conexiones individuales son claras gracias a la interacción que brinda la visualización. En dicha interacción, un nodo seleccionado resalta las conexiones entrantes y salientes de este. En el prototipo evaluado, no se puede observar la dirección de la comunicación. Sin embargo, el paradigma cuenta con posibilidades para observar la dirección de las conexiones (ver figura 8.9), mediante colores o bien por el nivel de donde la conexiones empiezan o terminan.

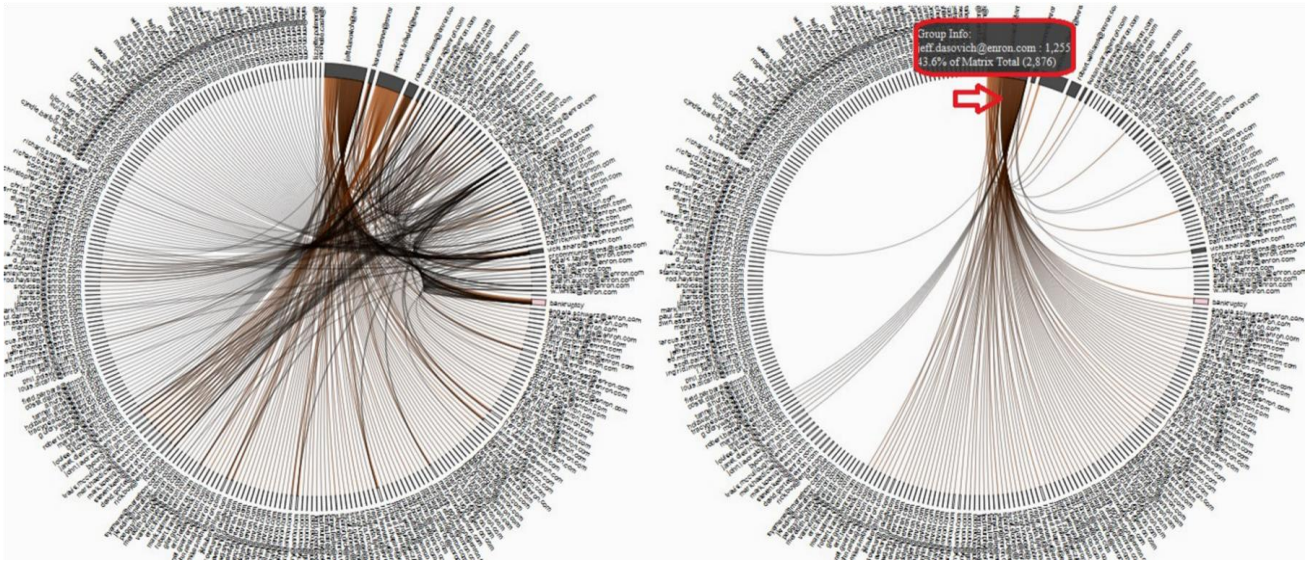


Figura 8.8 Conexiones de una persona mediante la interacción sobre la visualización.

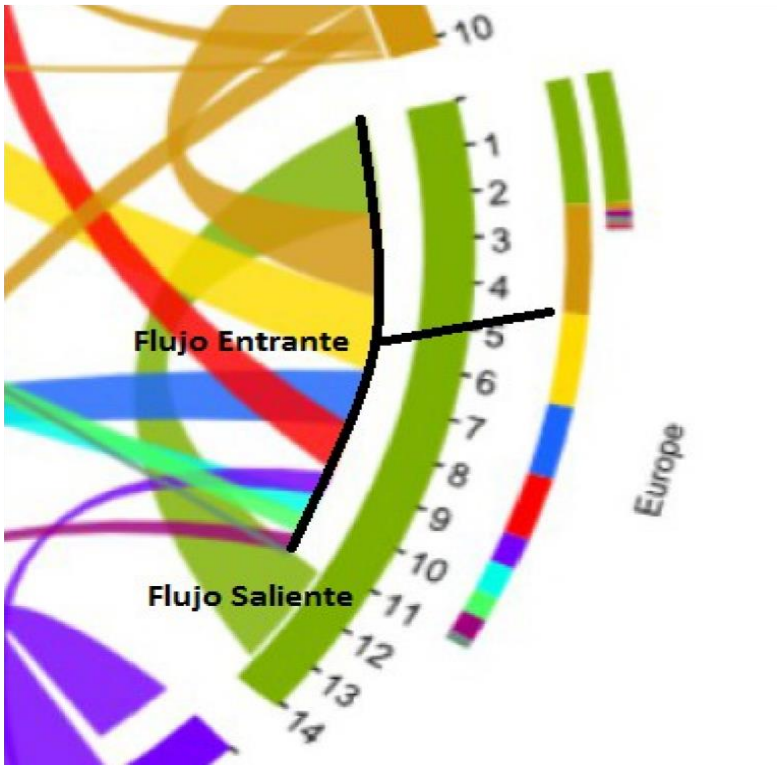


Figura 8.9 Conexiones entrantes y conexiones salientes del paradigma Circos.

8.7.1.3 Observaciones de “conexiones” sobre el paradigma de Grafo

Las conexiones en el paradigma de visualización Grafo son una parte base de visualización. Sin embargo, estas no siempre se pueden visualizar claramente. En algunos casos, es necesario navegar por el Grafo para observar en detalle las conexiones y comprender quién habla con quién.

La claridad de las conexiones variará según la cantidad de información que se visualiza y el algoritmo de ordenamiento aplicado sobre el Grafo. Las conexiones son difíciles de visualizar cuando varios nodos se sobreponen entre sí. Esto puede ocurrir por modificaciones en las variables de atracción, repulsión y gravedad del algoritmo de ordenamiento, que produzcan una fuerte atracción entre los nodos, ocultando así las conexiones entre estos (ver figura 8.10).

En la figura 8.10, se puede observar el Grafo correspondiente a las comunicaciones de correo electrónico de la corporación Enron con palabras claves relacionadas a “corrupción”. En dicho ejemplo, se alteran los pesos de gravedad, repulsión y atracción, con el fin de producir un Grafo cuyos nodos estén cercanos entre sí. La figura 8.11 muestra el mismo set de datos, desplegado con el mismo algoritmo (ForceAtlas), pero usando los atributos que la herramienta Gephi [34] brinda por defecto para el algoritmo de ForceAtlas.

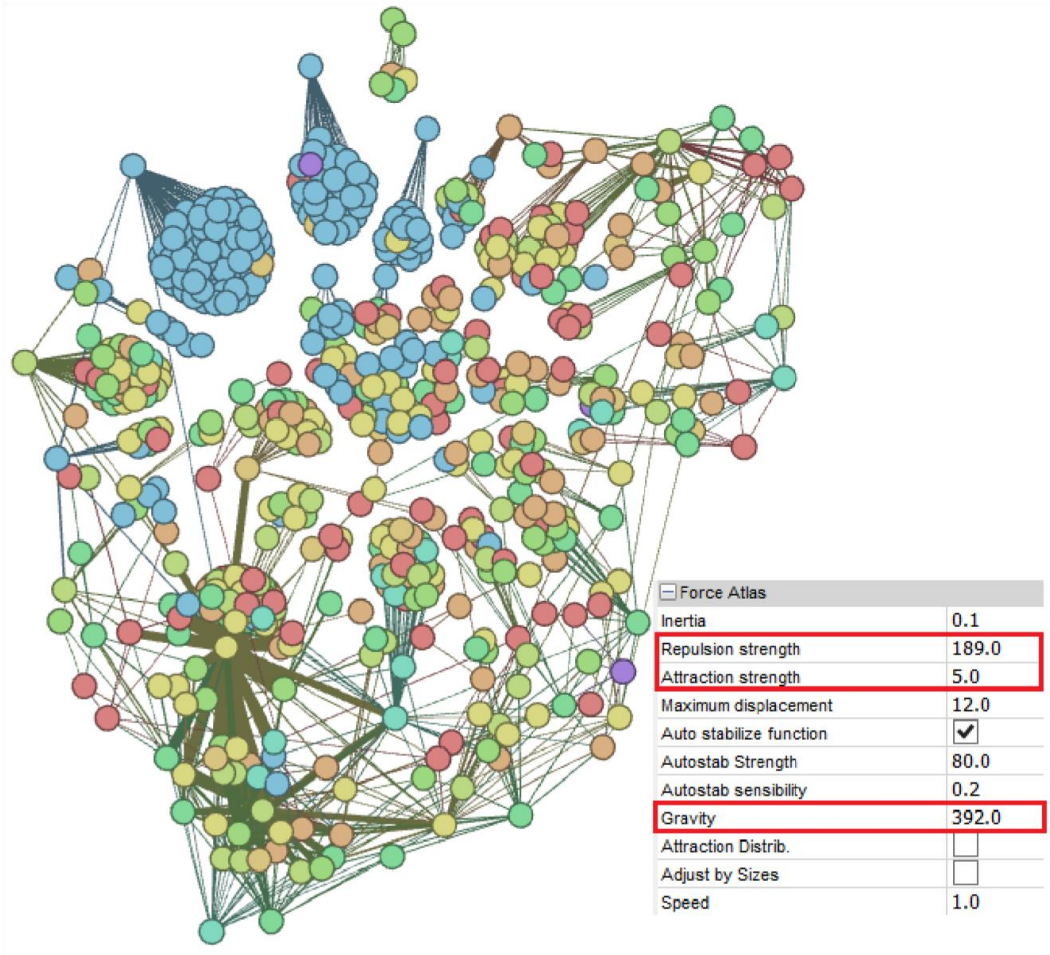


Figura 8.10 Grafo no dirigido con algoritmo de ordenamiento ForceAtlas.

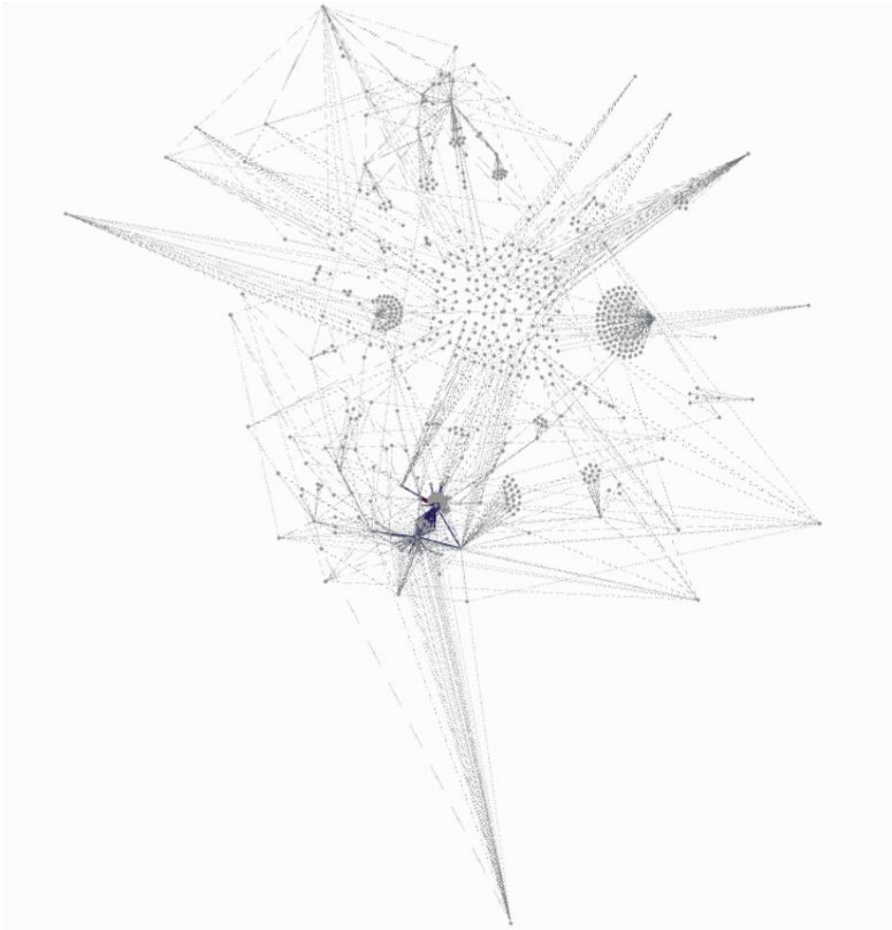


Figura 8.11 Grafo no dirigido de la base de datos Enron con las palabras claves de corrupción y un algoritmo de ForceAtlas aplicado para el agrupamiento de los nodos.

Para los casos donde ningún o pocos nodos se sobreponen entre sí, se observó que puede ser difícil leer el nombre de las personas en los nodos cuando el Grafo es muy grande. Esto dificulta comprender quién está hablando con quién (ver figura 8.12).

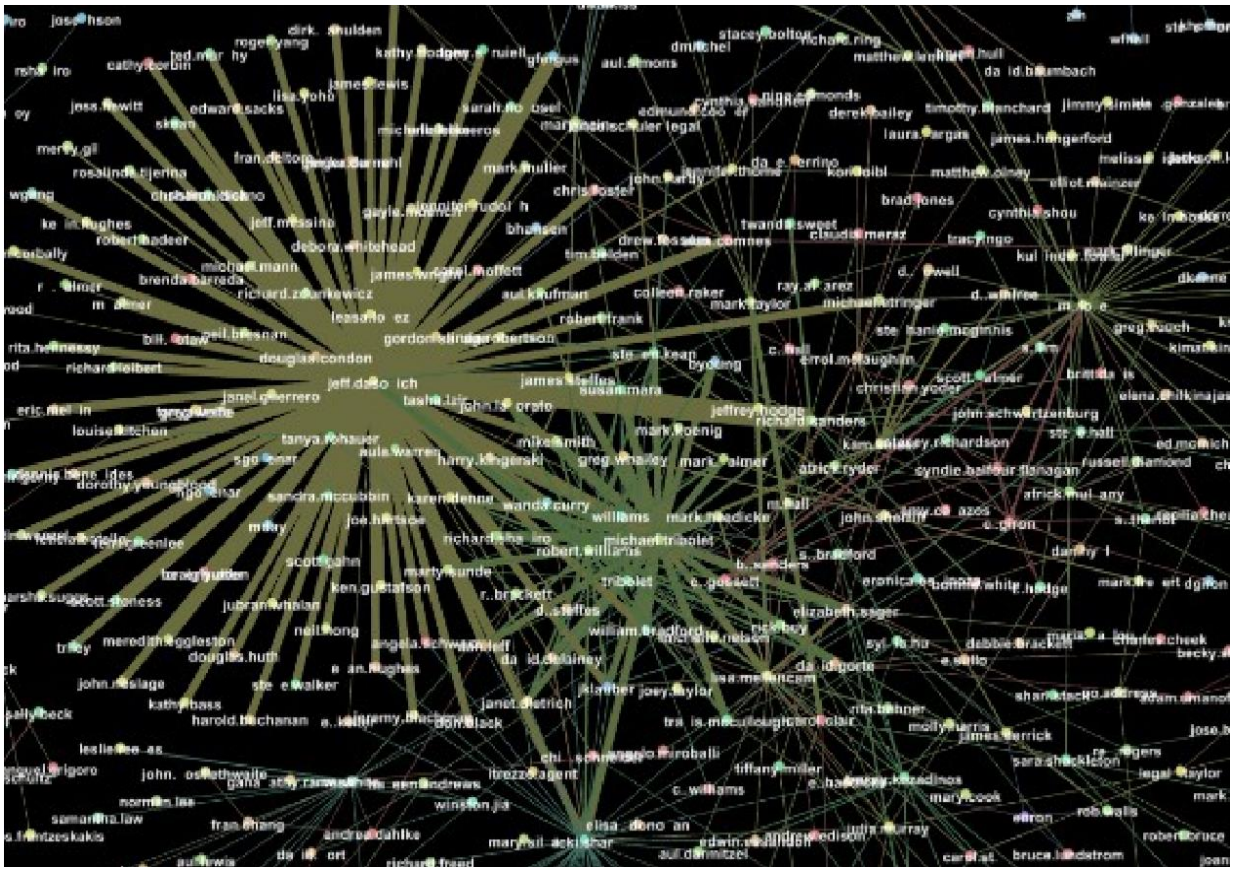


Figura 8.12 Acercamiento sobre Grafo no dirigido, sobre el cual sigue siendo difícil observar los nombres de los nodos.

El paradigma de visualización de Grafo también comprende la capacidad de resaltar las conexiones de nodos seleccionados, mientras se mantiene el contexto del resto conexiones (ver figura 8.13). Herramientas como Gephi [34] permiten dicha capacidad de interacción.

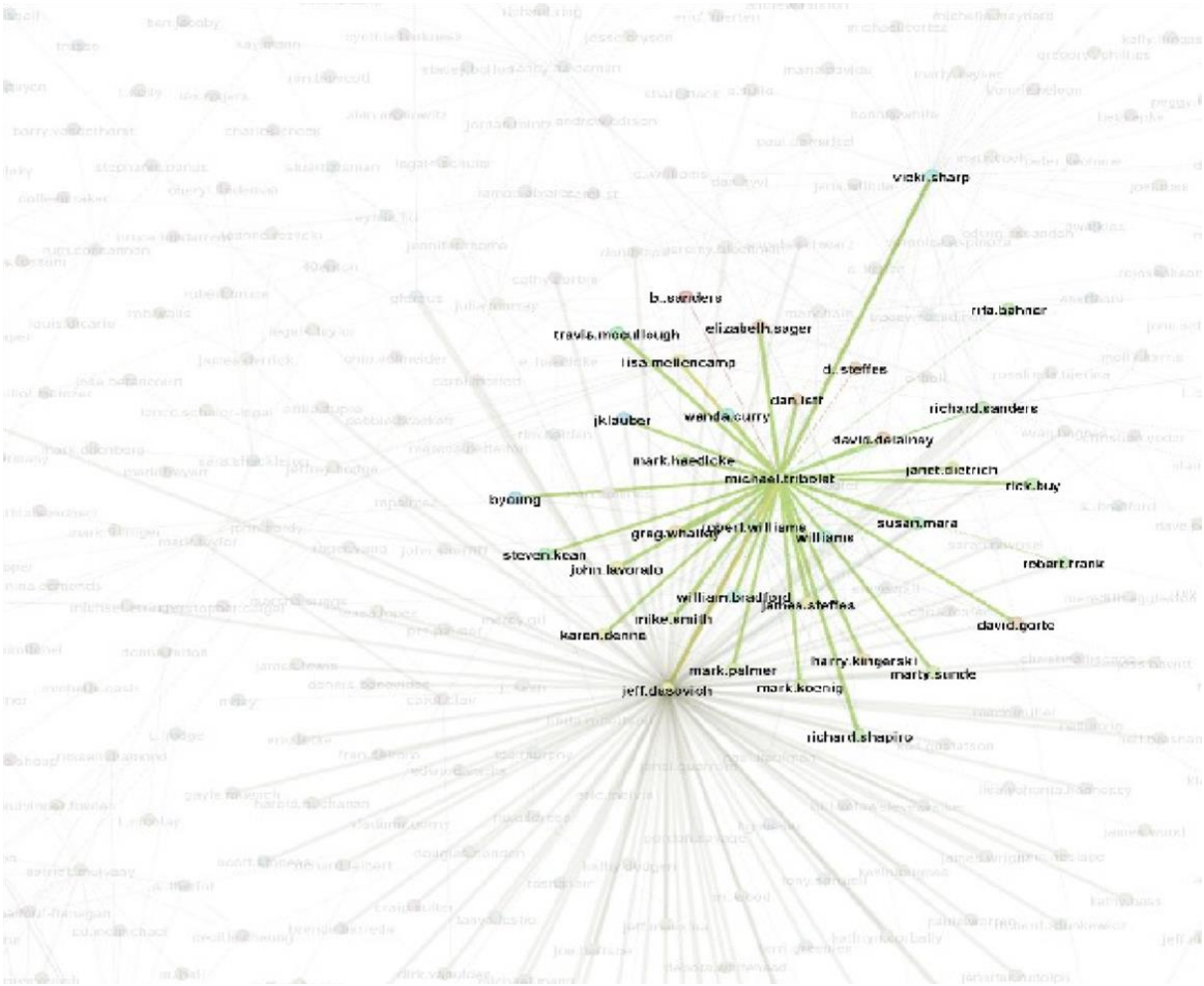


Figura 8.13 Un nodo y sus conexiones resaltadas en un Grafo no dirigido

8.7.1.4 Observaciones sobre el paradigma Gmap

Si bien las conexiones existen en esta visualización, es difícil poder concluir mediante estas quién se comunica con quién en aquellos casos donde existe una alta densidad de conexiones. En casos donde la densidad acumulada de conexiones no sea tan alta, es posible observar las conexiones. Aun así, las conexiones pueden llegar a perderse (ver figura 8.15). Este problema puede ser

resuelto por un método de interacción, que permita resaltar las conexiones de un nodo seleccionado.

En algunos casos se observa (ver figura 8.14) que existen entes contiguos, que no cuentan con conexiones entre sí. Dicha observación demuestra que la cercanía entre dos nodos no siempre implica una conexión entre estos.

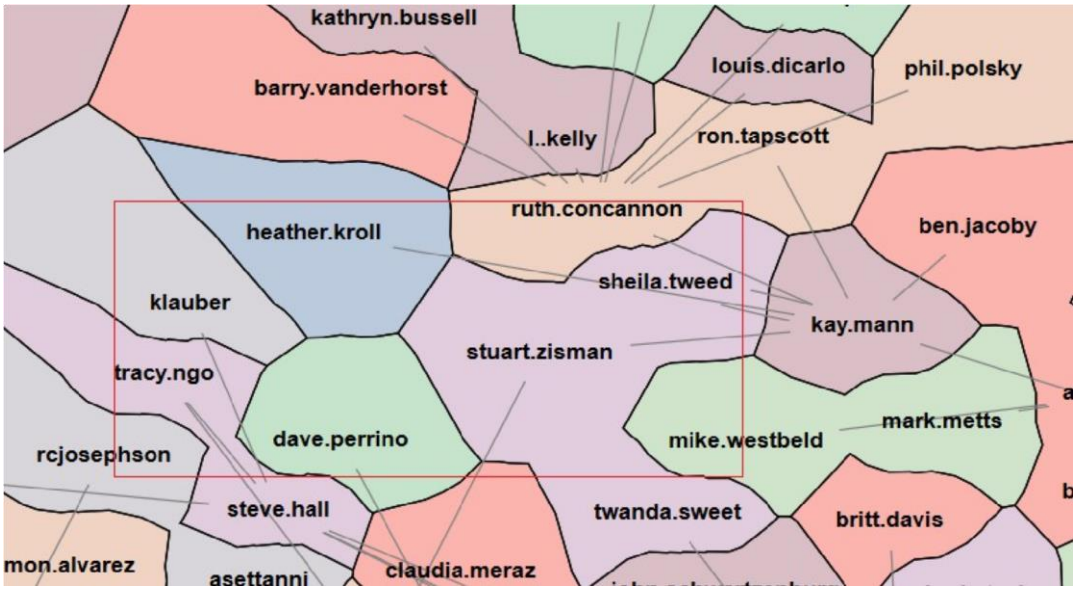


Figura 8.14 Gmap con personas cercanas entre sí, que no cuentan con comunicación entre ellos.

Por otra parte, se ha observado que las conexiones en el mapa no aportan a comprender la dirección de la comunicación.

Se concluye entonces que la cercanía de los entes en el mapa no siempre implica una conexión entre estos, mas sí evidencia cierta relación, por los nodos con los que comparten conexiones. Las conexiones aportan a entender quién habla con quién, mas estas pueden ser un poco difíciles de visualizar cuando

existe una alta densidad de conexiones. Técnicas de interacción y de focus+context podrían ayudar a una clara comprensión de “quién habla con quién” y en “qué dirección” ocurre la comunicación.

Figura 8.15 Alta densidad en las conexiones de un Gmap.

8.7.1.5 Calificación de la métrica #1 – Conexiones

Según las observaciones hechas de la métrica de conexiones, se califican los paradigmas de la siguiente forma:

1. El paradigma Circos y Grafos con una calificación de **10,** por contar con mecanismos que permiten una clara comprensión de quién habla con quién 2. El paradigma Gmaps con una calificación de **8**, por cuanto se puede perder claridad

en quién habla con quién cuando existen una gran cantidad de conexiones sobrepuestas entre sí.

8.7.2 Evaluación de métrica #2 – Co-localización

8.7.2.1 Hipótesis:

- Es posible visualizar en una población a aquellas personas que pertenecen a un mismo grupo (localización) y poder distinguirlos claramente de otras personas pertenecientes a otros grupos.

8.7.2.2 Observaciones sobre el paradigma Circos

El paradigma de visualización Circos permite agrupar un conjunto de personas por proximidad de posición sobre el perímetro de la circunferencia, por los colores de sus conexiones, por los colores de grupo o bien por niveles adicionales alrededor del perímetro de la circunferencia que permite visualizar y comprender atributos extras con respecto a la visualización.

En el caso de la visualización sobre flujos migratorios (figura 8.16), se puede observar las proporciones de flujos entrantes y salientes sobre cada una de las regiones.

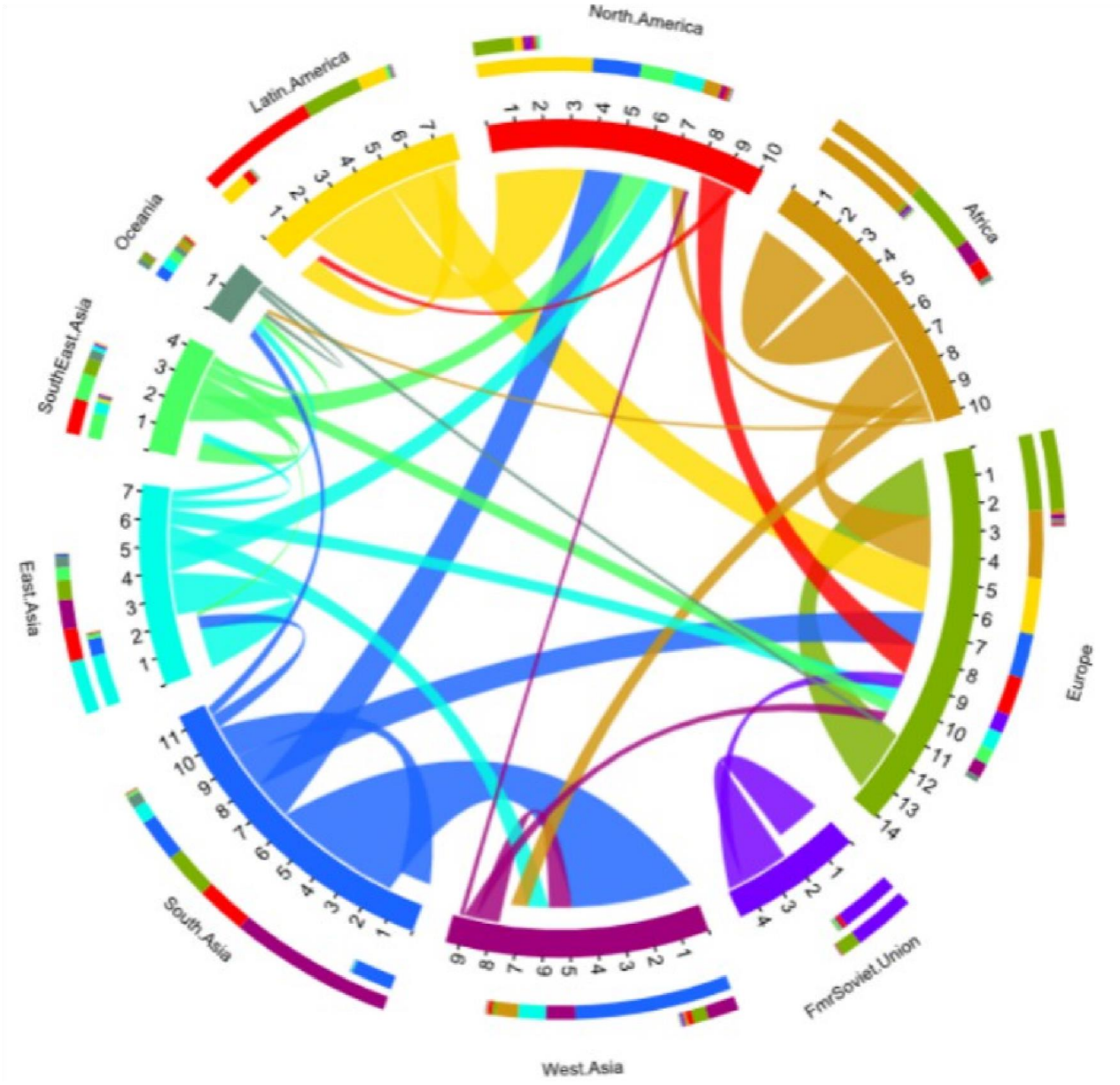


Figura 8.16 Circos de flujos migratorios evidencian una forma de agrupamiento [88].

8.7.2.3 Observaciones sobre el paradigma de Grafo

Coloreando los nodos es posible observar qué nodo pertenece a qué grupo en el Grafo y al mismo tiempo comprender la distribución del grupo sobre la población en general (ver figura 8.17). Sin embargo, en el plano general del Grafo, solo se puede observar de forma clara quién pertenece a quién explorando en detalle las distintas partes de este.

Debido a cómo funcionan los algoritmos de agrupamiento, en especial sobre grandes redes de datos, como lo es el correo corporativo de Enron, no es posible diferenciar los colores al observar el Grafo en su totalidad (ver figura 8.10). El comprender la distribución de los grupos requiere de una amplia exploración e interacción sobre el Grafo observándolo desde un nivel donde se puedan distinguir los colores de los nodos claramente.

Variaciones sobre la estructura del Grafo, sus colores y mediante un algoritmo de ordenamiento como lo es el frunchterman Reingold [82], pareciera que es más sencillo categorizar la co-localización de los puntos y definir a qué grupos pertenecen (ver figura 8.20). Sin embargo, mucho se encuentra en una fuerte ilusión causada por los colores predominantes del nodo fuente de la comunicación, los cuales hacen pensar que existe una alta concentración de ese grupo en esa área. Lo anterior, puede ser engañoso para el lector de la visualización.

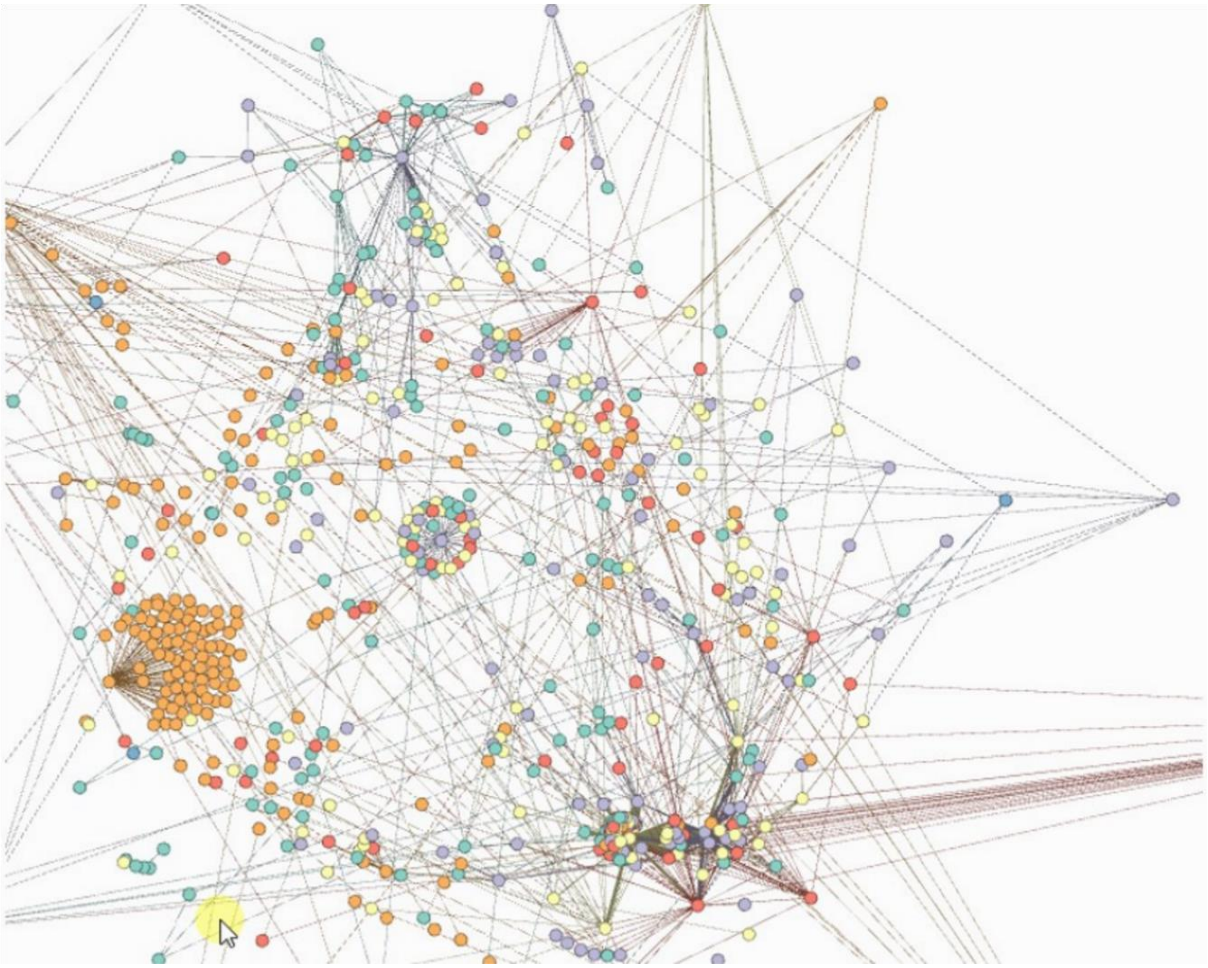


Figura 8.17 Grafo de comunicaciones relacionadas a palabras clave sobre corrupción en el correo corporativo de Enron.

8.7.2.4 Observaciones sobre el paradigma Gmap

El gmap genera una excelente forma de visualizar a aquellas personas que comparten un mismo grupo. Nos permite comprender la proporción que tiene un grupo con respecto a los otros grupos. En la figura 8.18 se puede observar cómo el concepto de “mapa” contribuye a una clara y natural comprensión de los grupos.

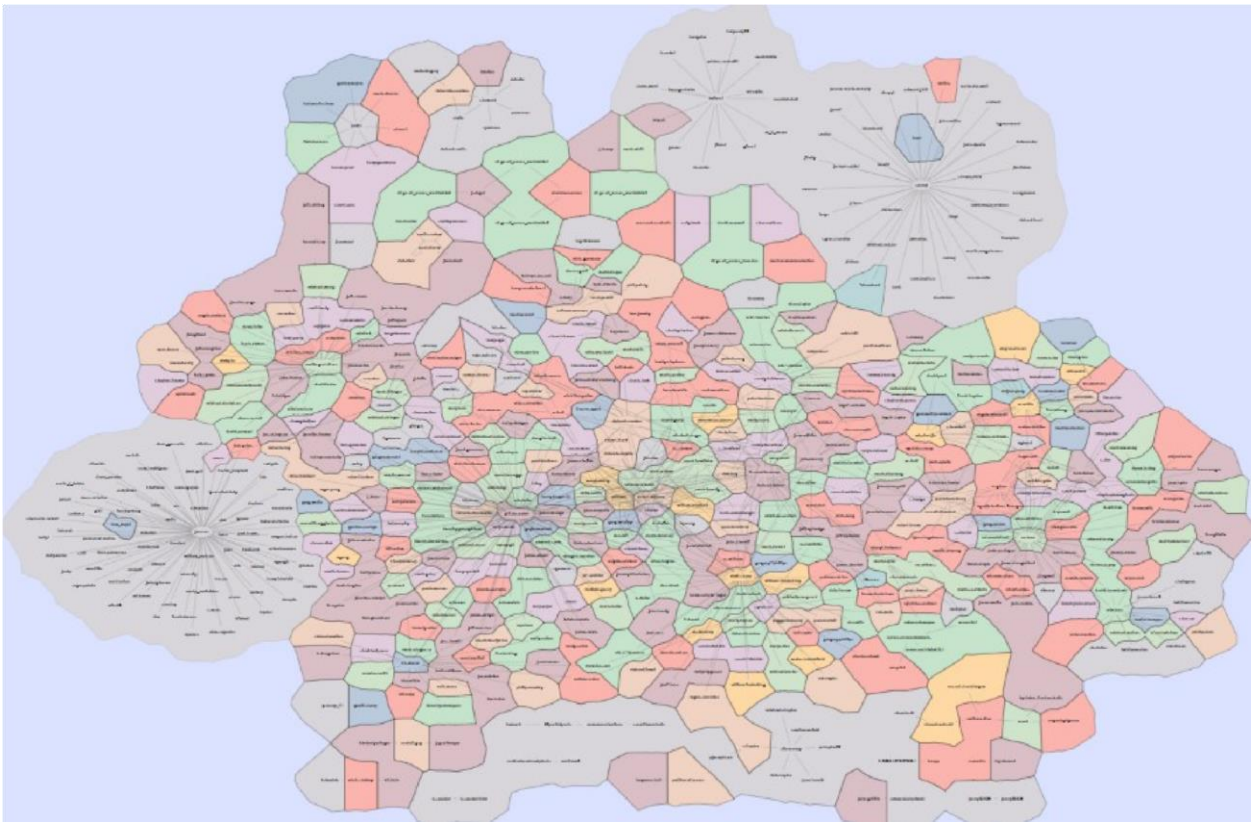


Figura 8.18 Gmap generado con el conjunto de palabras relacionado a corrupción

8.7.2.5 Calificación de métrica #2 – “Co-localización”

Según las observaciones hechas de la métrica de co-localización, se califican los paradigmas de la siguiente forma:

1. El paradigma Circos y Gmap con una calificación de **10,** por cuanto permiten de forma clara comprender qué persona pertenece a qué grupo.
2. El paradigma Grafo con una calificación de **7,** por cuanto es difícil distinguir, en grafos de gran tamaño, los colores de los nodos que caracterizan la pertenencia de un nodo con un grupo.

8.7.3 Evaluación de métrica #3 – Frecuencia

8.7.3.1 Hipótesis:

- Es posible visualizar la frecuencia con que se habla de un tema específico, ya sea entre dos personas o en el plano general de las comunicaciones de la población.

8.7.3.2 Observaciones sobre el paradigma Circos

Variaciones a la visualización como la implementada por el grupo de MIT y el grupo de investigación de General Electric [66] (ver figura 8.19) permiten una visualización de los nodos relevantes. Sin embargo, no se puede distinguir la frecuencia de una comunicación entre dos puntos y la concentración del volumen de las conexiones. Se puede entonces decir que en el paradigma Circos, el volumen de las comunicaciones sobre un tema se puede comprender de forma muy limitada.

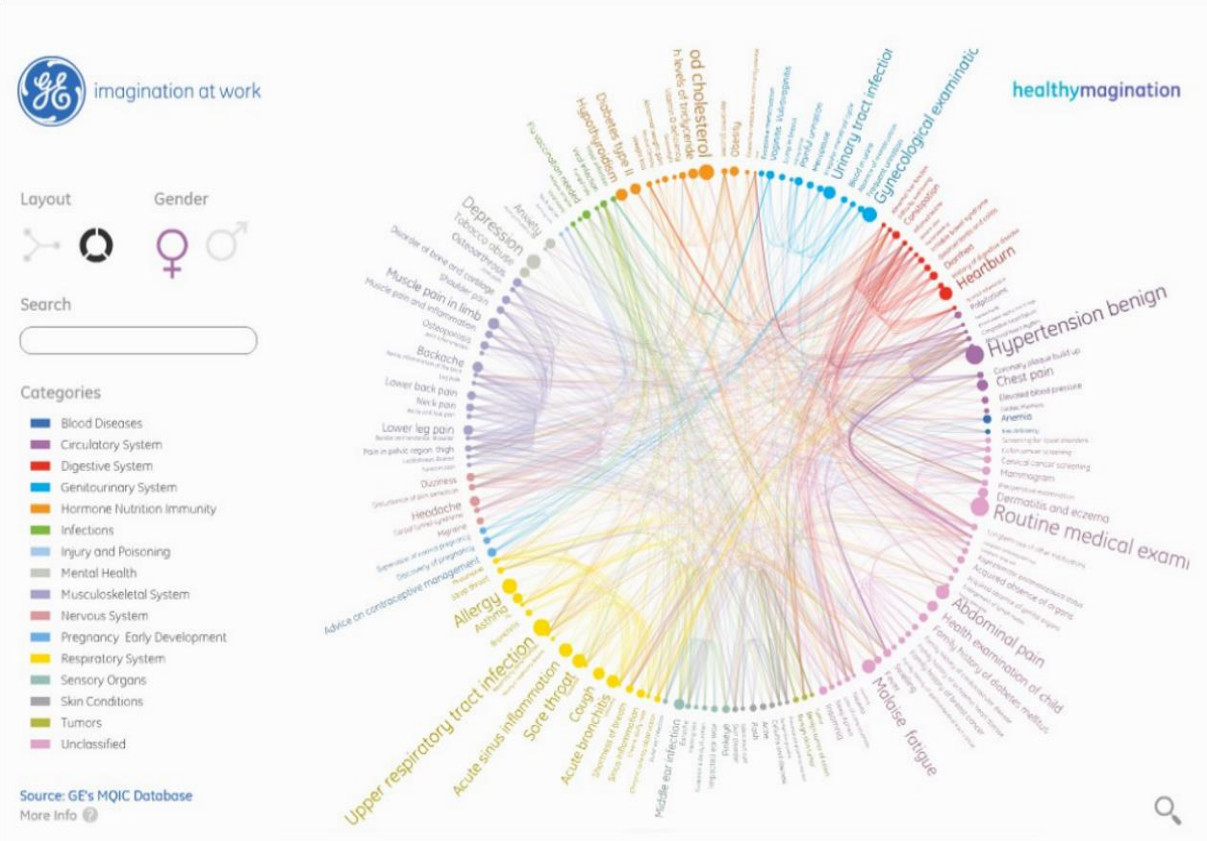


Figura 8.19 Relaciones entre enfermedades y síntomas [66].

8.7.3.3 Observaciones sobre el paradigma de Grafo

Es posible visualizar aquellos nodos con una gran cantidad de comunicaciones y que estos resalten del resto por la frecuencia de sus comunicaciones. En la figura (ver figura 8.20) se puede observar un Grafo no dirigido con un algoritmo de ordenamiento frunchterman Reingold [82] que lo transforma prácticamente en una visualización hiperbólica sobre la cual es claro distinguir las frecuencias.

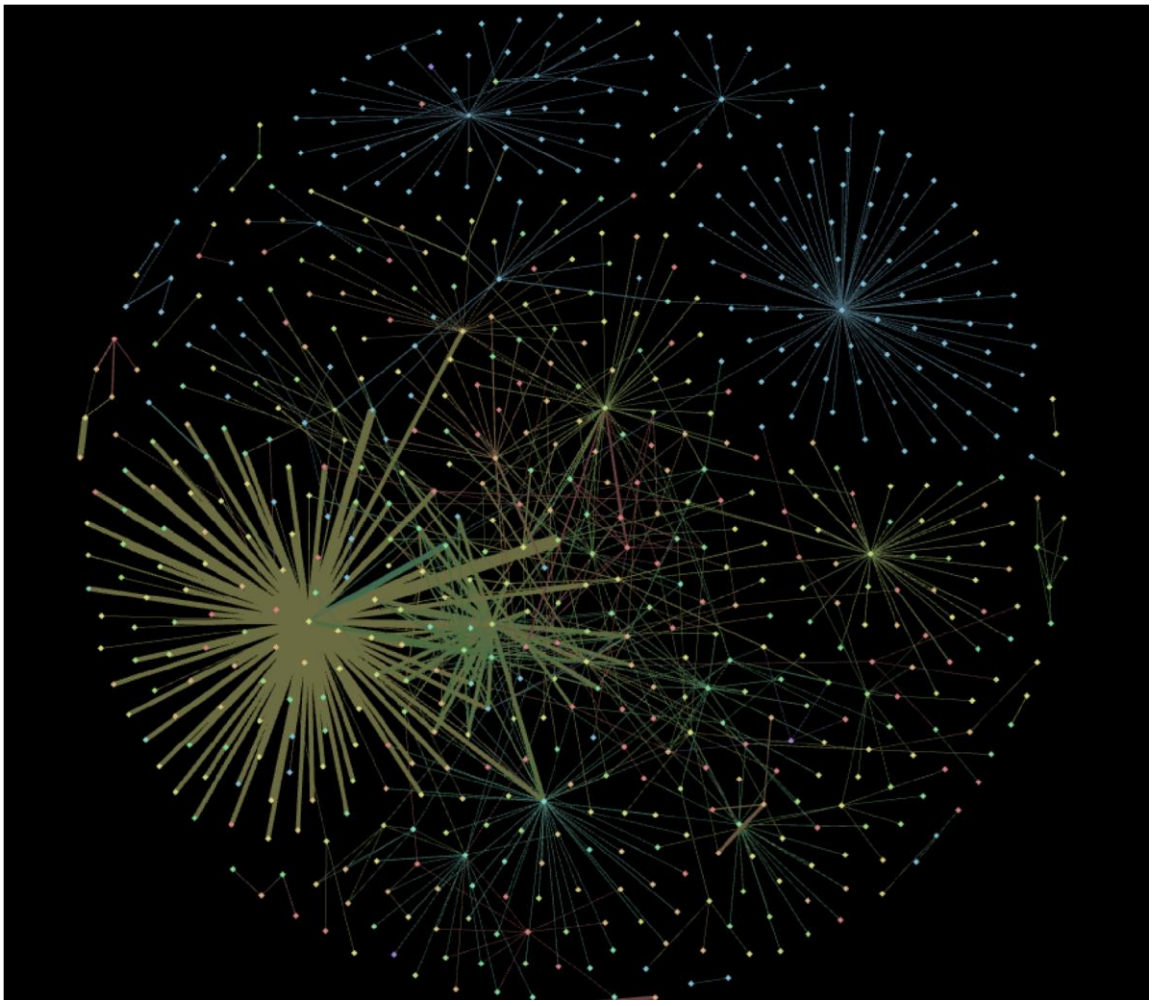


Figura 8.20 Grafo con ordenamiento frunchterman Reingold [82], de los correos con las palabras asociadas a “corrupción” en el correo electrónico Enron.

8.7.3.4 Observaciones sobre el paradigma Gmap

A diferencia de otras visualizaciones, la frecuencia es una combinación entre la cercanía de las personas en el mapa y las conexiones. Dichos aspectos permiten visualizar de forma natural quién habla con quién, mas no se observa con qué frecuencia lo hacen.

Claramente, en el contexto general, se podría modificar el grosor de las conexiones para resaltar aquellas conexiones con muy alta frecuencia sobre aquellas con muy poca frecuencia. Otra alternativa, que se puede observar en trabajos previos [18] sobre el Gmap, es modificar el tamaño de la etiqueta del nodo para así diferenciar la frecuencia con que una persona habla con respecto al resto de las personas que tienen conocimiento en un tema (ver figura 8.21).

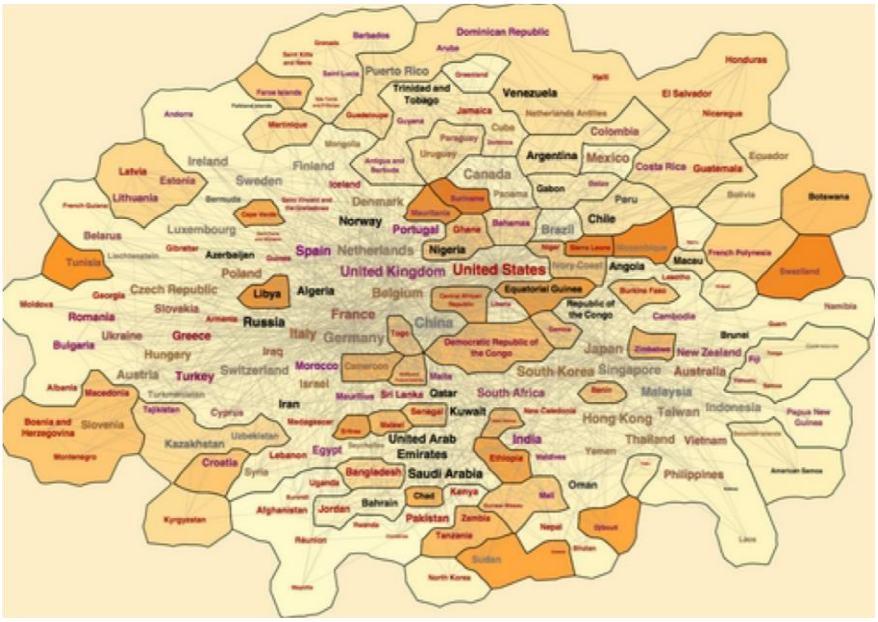


Figura 8.21 Mapa de tratados entre naciones [18].

8.7.3.5 Calificación de métrica #3 – “Frecuencia”

Según las observaciones hechas de la métrica de frecuencia, se califican los paradigmas de la siguiente forma:

1. El paradigma Circos con una calificación de **10,** por cuanto cuenta con excelentes opciones para poder distinguir las conexiones fuertes de las conexiones débiles.
2. El paradigma Gmap y Grafo con una calificación de **9**, por cuanto ambos muestran la frecuencia de forma satisfactoria. En el caso del Grafo, la frecuencia es dada principalmente por el grosor de las conexiones, mientras que en el Gmap, cuenta con una combinación entre la frecuencia dada por la cercanía de los nodos en el mapa, y las conexiones. Cabe resaltar, que el Gmap, cuenta con mecanismos de su implementación base en grahviz, sobre el cuál se puede modificar el grosor de las líneas, lo cual mejoraría la visualización de la frecuencia.

8.7.4 Evaluación de métrica #4 – Relación Tema-Localización

8.7.4.1 Hipótesis:

- Es posible observar la relación entre un tema y la localización de dónde este tiene presencia. Se puede comprender así aspectos de colaboración, sinergia e islas de comunicación, entre otros.

8.7.4.2 Observaciones sobre el paradigma Circos

La relación tema-localización no existe en este paradigma. Si bien se pueden observar patrones en las conexiones, nada se puede concluir con respecto a grupos lógicos que pudieran surgir según las conexiones, sus frecuencias y su ubicación en un espacio geo-localizado.

Por ejemplo, en la figura 8.22, no es posible conocer qué personas del grupo rojo tienen una cercana relación con las personas del grupo verde en un tema específico. Como se observó anteriormente (ver figura 8.16), se podría comprender qué porcentaje del grupo rojo tiene relación con el grupo verde, pero no se lograría visualizar quiénes componen ese porcentaje o con qué otros grupos tienen relación.

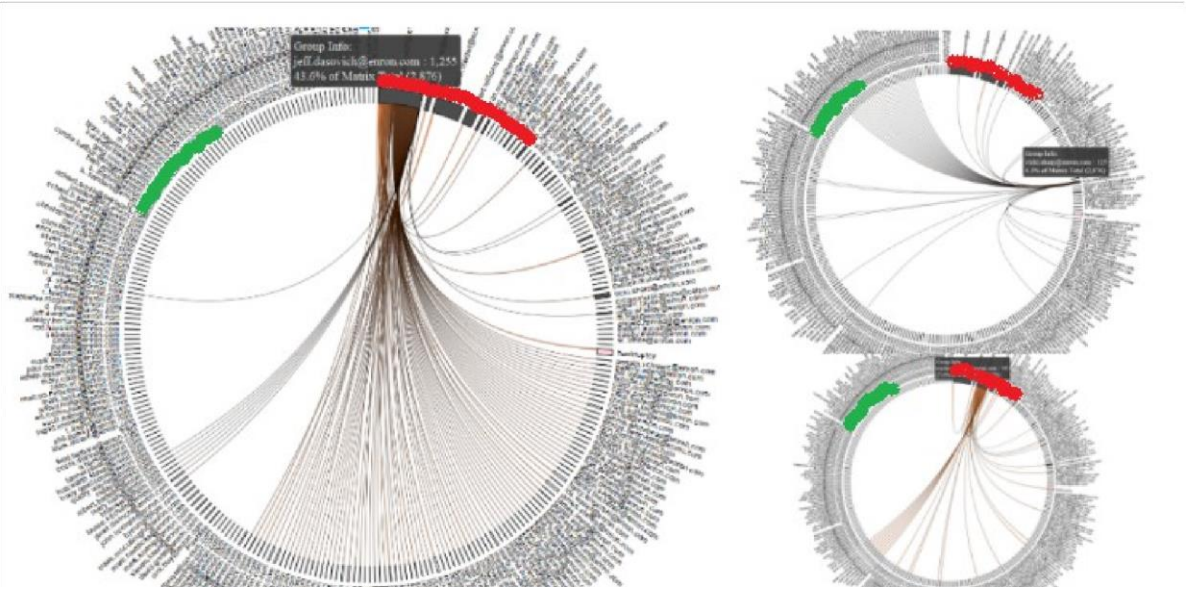


Figura 8.22 Conexiones de correo electrónico de personas con mayor volumen de comunicación sobre un tema.

8.7.4.3 Observaciones sobre el paradigma de Grafo

No es posible mediante el Grafo observar la relación entre un tema y la localización de dónde este tiene presencia, esto principalmente por cuanto es difícil distinguir los grupos a los cuales pertenecen los nodos cuando se observa el Grafo en el contexto de toda la red.

Probando diferentes algoritmos y modificando los pesos sobre las variables de gravedad, de repulsión, de atracción, etc., se procuró replicar una representación similar a la claridad que Gmap provee para evidenciar la relación tema-localización. Las distintas pruebas fueron negativas por cuanto siempre quedan un gran conjunto de nodos altamente dispersos y que se escapan del enfoque necesario para poder comprender a qué grupos pertenecen los nodos y por ende la relación tema-localización (ver figura 8.23).

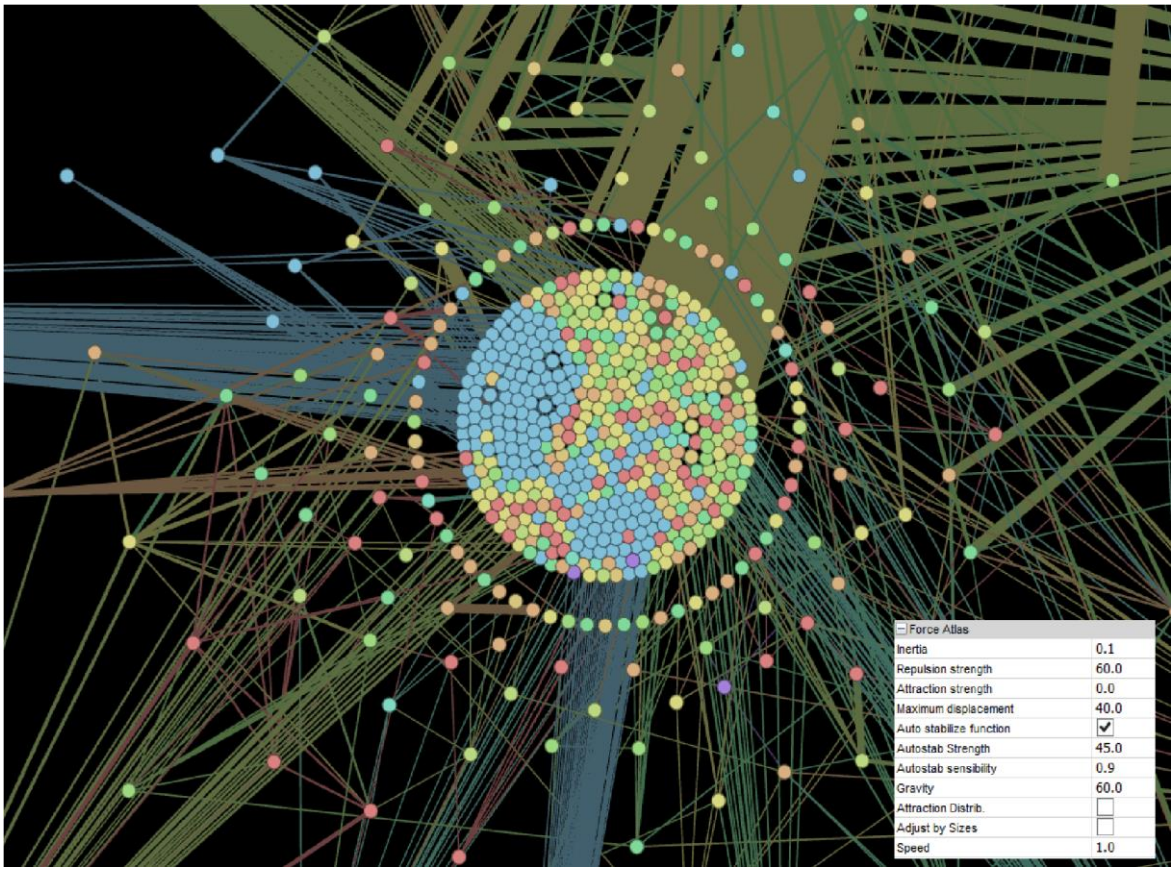


Figura 8.23 Grafo No Dirigido de las comunicaciones de correo de la base de datos Enron que contiene el conjunto de palabras de corrupción con un algoritmo de agrupamiento Force Atlas con valores modificados para acercar los nodos dentro de una sola imagen.

8.7.4.4 Observaciones sobre el paradigma Gmap

Se pueden distinguir los grupos y cómo la relación entre estos sobre un tema domina su lugar dentro del mapa con respecto a los otros nodos. En el caso del Gmap que visualiza el conjunto de palabras relacionadas a corrupción, podemos apreciar dos ámbitos de comunicación, con patrones muy distintos con respecto a la localización de las personas que los conforman (ver figura 8.24). En el conjunto número uno, se aprecia una composición variada de grupos, mientras

en el conjunto número dos, predomina un grupo representado por el conjunto de “personas externas” a la corporación Enron. También se puede observar que en el conjunto de personas externas, existe una persona que pertenece a uno de los grupos internos de la corporación Enron, la cual pareciera tener en el tema una mayor relación con las personas externas a la empresa.

Figura 8.24 Conjuntos en el Gmap que muestran las diferentes composiciones de localización del conocimiento sobre un tema.

8.7.4.5 Calificación de métrica #4 – “Relación Tema-Localización”

Según las observaciones hechas de la métrica de relación tema-localización, se califican los paradigmas de la siguiente forma:

1. El paradigma Circos con una calificación de **10,** por cuanto la forma de mapa permite una clara e intuitiva visualización de la geo-localización del conocimiento.

2. El paradigma Grafo con una calificación de **3**, por cuanto su limitante en poder visualizar los colores de los nodos en grafos de gran tamaño, hace difícil comprender siempre el aspecto de geo-localización que se busca.

3. El paradigma Circos con una calificación de **1**, por cuanto no existe forma de comprender la geo-localización en este paradigma.

8.8 Análisis de los Resultados

Sobre las observaciones realizadas, se analiza y concluye la relación con las hipótesis planteadas para cada métrica

8.8.1 Circos

La visualización Circos no es necesariamente intuitiva a primera vista. Sin embargo, una vez que se conoce la visualización se puede distinguir claramente entre los patrones claves para establecer conclusiones sobre la información. Dos de los factores claves de este paradigma de visualización son:

- La interacción, la cual permite un excelente focus+context para explorar la información, ya sea por medio de la selección de los nodos, sus conexiones o bien mediante filtros que permitan dar un perspectiva correcta del tema que se quiere visualizar.
- La escalabilidad de la visualización, la cual permite múltiples variaciones para poder agregar la visualización de otros atributos.

La aplicabilidad de este paradigma de visualización sobre el problema de investigación a resolver es baja, debido a la imposibilidad de poder visualizar de forma clara la relación entre tema-localización en el contexto de las comunicaciones relacionadas entre múltiples grupos.

Conforme a las observaciones y el análisis realizado, el investigador procede a evaluar al paradigma de visualización **Circos** de la siguiente manera: (ver cuadro 8.1):

Cuadro 8.1 Evaluación de Paradigma Circos

METRICAS	Hipótesis	Calificación Circos
Métrica #1 - Conexiones	Es posible conocer quién habla con quién en el plano individual	9 - Muy bueno
Métrica #1 - Conexiones	Es posible observar la dirección de la comunicación. Dicha dirección puede contar con dos formas, la primera sobre la cuál el mensaje es "enviado" y la segunda sobre la cuál el mensaje es "recibido"	9 - Muy bueno
Métrica #2 - Co-localización	Es posible visualizar en una población, aquellas personas que pertenecen a una mismo grupo (localización) y poder distinguirlos claramente de otras personas pertenecientes a otros grupos	9 - Muy bueno
Métrica #3 - Frecuencia	Es posible visualizar la frecuencia que se habla de un tema específico, ya sea entre dos personas o en el plano general de la comunicaciones de la población	3 - Malo
Métrica #4 - Tema-Locaclización	Es posible observar la relación entre un tema y la localización de donde este tiene presencia. Comprendiendo así aspectos de colaboración, sinergia, islas de comunicación, entre otros.	1 - Pésimo

8.8.2 Gmap

Para el paradigma de visualización Gmap, se puede concluir que es una excelente herramienta para observar aquellas personas con una fuerte relación entre sí sobre un tema. Es sencillo distinguir a qué localización pertenecen los distintos nodos y cómo los grupos se encuentran distribuidos en el mapa de la comunicación sobre un tema. Permite visualizar aspectos donde las personas tienden más a pertenecer a círculos de comunicación propios que a círculos de comunicación de corporativos, así como también se pueden observar temas comunes entre personas de un mismo grupo.

Encontramos mediante las observaciones que en comparación a otros paradigmas, el Gmap cuenta con áreas qué podrían mejorarse en pro de tener una mejor comprensión de las conexiones y las frecuencias con que estas ocurren.

Conforme a las observaciones y el análisis realizado se procede a evaluar al paradigma de visualización **Gmap** de la siguiente manera: (ver cuadro 8.2):

Cuadro 8.2 Evaluación de Paradigma Gmap

METRICAS	Hipótesis	Calificacion Gmaps
Métrica #1 - Conexiones	Es posible conocer quién habla con quién en el plano individual	7 - Bueno
Métrica #1 - Conexiones	Es posible observar la dirección de la comunicación. Dicha dirección puede contar con dos formas, la primera sobre la cuál el mensaje es "enviado" y la segunda sobre la cuál el mensaje es "recibido"	6 - Regular
Métrica #2 - Co-localización	Es posible visualizar en una población, aquellas personas que pertenecen a una mismo grupo (localización) y poder distinguirlos claramente de otras personas pertenecientes a otros grupos	10 - Excelente
Métrica #3 - Frecuencia	Es posible visualizar la frecuencia que se habla de un tema específico, ya sea entre dos personas o en el plano general de la comunicaciones de la población	8 - Bueno
Métrica #4 - Tema-Locaclización	Es posible observar la relación entre un tema y la localización de donde este tiene presencia. Comprendiendo así aspectos de colaboración, sinergia, islas de comunicación, entre otros.	10 - Excelente

8.8.3 Grafo

Sobre el paradigma de visualización de Grafo, podemos concluir que es una excelente visualización para la exploración de una red. Nos permite comprender claramente aquellos nodos relevantes y las comunidades alrededor de estos. Permite visualizar grupos con alta relación en sus comunicaciones. Sin embargo, el paradigma no muestra con claridad la pertenencia de los nodos a los grupos con los que se

comprende la variable de localización del conocimiento. Esto por cuanto, para redes de gran tamaño como lo es el correo corporativo de Enron, se dificulta la compresión de la distribución de los grupos, diferenciados entre sí por sus colores. Por lo tanto, no se puede distinguir de forma clara cómo el conocimiento se encuentra distribuido entre los diversos grupos.

Conforme a las observaciones y el análisis realizado se procede a evaluar al paradigma de visualización **Grafo** de la siguiente manera: (ver cuadro 8.3):

Cuadro 8.3 Evaluación de Paradigma Grafo

METRICAS	Hipótesis	Calificación Grafo
Métrica #1 - Conexiones	Es posible conocer quién habla con quién en el plano individual	8 - Bueno
Métrica #1 - Conexiones	Es posible observar la dirección de la comunicación. Dicha dirección puede contar con dos formas, la primera sobre la cuál el mensaje es "enviado" y la segunda sobre la cuál el mensaje es "recibido"	9 - Muy bueno
Métrica #2 - Co-localización	Es posible visualizar en una población, aquellas personas que pertenecen a una mismo grupo (localización) y poder distinguirlos claramente de otras personas pertenecientes a otros grupos	4 - Regular
Métrica #3 - Frecuencia	Es posible visualizar la frecuencia que se habla de un tema específico, ya sea entre dos personas o en el plano general de la comunicaciones de la población	10 - Excelente
Métrica #4 - Tema-Locaclización	Es posible observar la relación entre un tema y la localización de donde este tiene presencia. Comprendiendo así aspectos de colaboración, sinergia, islas de comunicación, entre otros.	3 - Malo

8.9 Comparación de los Resultados de los Paradigmas

El cuadro de comparación de las evaluaciones sobre los paradigmas de visualización muestra cómo el paradigma de visualización Gmap fue el que mejor puntuó sobre la calificación general de las cuatro métricas evaluadas.

Cuadro 8.4 Puntuación comparativa de los resultados de las evaluaciones sobre los paradigmas de visualización.

Métricas	**Circos**	**Gmap**	**Grafo**
Conexiones	10	8	10
Co-localización	10	10	7
Frecuencia	10	9	9
Relación Tema-Localización	1	10	3
Puntuación Total	*31*	***37***	*29*

En los resultados de las evaluaciones, cabe resaltar el paradigma de visualización Circos, que encuentra su fortaleza en el poder evidenciar patrones, conexiones y secuencias sobre gran cantidad de datos. Por su parte, en el Grafo no dirigido, mediante el uso de la herramienta gephi [34], se logra evidenciar el gran potencial que el Grafo como visualización tiene para la exploración y comprensión de redes sociales.

Tanto el Circos, como el Grafo no Dirigido no fueron aptos para poder mostrar la distribución entre un tema y los diversos grupos que lo tratan. Caso contrario ocurre con el Gmap, sobre el cual, tal y como sus autores describen [90], permite una visualización

natural de tipo geo-espacial sobre la cual se puede fácilmente comprender qué grupos hablan de qué tema y con qué relación entre estos.

9 Prototipo Final

9.1 Mejoras y Diseño del Prototipo Final de Visualización

Tal y como se puede observar en el cuadro 8.4 de comparación de las evaluaciones de los paradigmas de visualización, existen dos métricas sobre las cuales se pueden implementar mejoras orientadas a lograr una mejor visualización para resolver el problema de investigación planteado:

9.1.1 Frecuencia

La relación existente entre dos nodos en el Gmap se define por la conexión y por la distancia que existe en el mapa, entre uno y otro. Existen escenarios sobre los cuales, pueden existir relaciones fuertes entre nodos, no cercanos entre sí. Dicho escenario se puede visualizar por medio de otras visualizaciones tales como el Grafo no dirigido con una estructura hiperbólica (ver figura 9.1) usando el set de palabras relacionadas al tema “corrupción” existentes dentro de la base de datos Enron. Por medio de dicha visualización, se puede observar nodos que tienen una fuerte relación con el nodo principal, mas no necesariamente se encuentran cerca de este. Para observar este tipo de relaciones sobre el Gmap se implementará un atributo “penwidth” que la herramienta graphviz soporta. Dicho atributo determina el grosor de las líneas que representan la conexión sobre los nodos. De igual forma, interesa sobresaltar aquellos nodos con una gran frecuencia de comunicaciones en general de aquellos con una baja frecuencia de

comunicaciones. Para este efecto se hará uso del atributo de nodo “fontsize”, empleado en ejemplos de Gmap [18], pero que no fue utilizado durante las pruebas del prototipo. Por medio de este atributo se logrará incrementar el tamaño de la etiqueta del nodo, de tal forma que este resalte por su tamaño sobre otros con una menor frecuencia de comunicación.

9.1.2 Interacción

Por medio de las evaluaciones de los paradigmas de visualización, se pudo observar cómo los paradigmas Grafo y Circos son capaces de mostrar la relevancia de la interacción en una visualización. El focus+context se encontraba claramente aplicado en ambos ejemplos mediante la interacción sobre los nodos o sus conexiones, sobre los cuales una selección lograba enfocar la atención sobre un nodo (o bien conexión), mientras el resto, no seleccionado, pasa a un segundo plano.

El Gmap por su parte, según lo estudiado, carece de interacción y no cuenta con un mecanismo focus+context que nos permita enfocar la atención sobre nodos que interesen y sus respectivas conexiones. Para esto, se ha implementado un código (ver figura 9.1), haciendo uso de jquery, el cual modifica los atributos y eventos para poder dotar al Gmap de una interacción que nos permita lograr una mejor comprensión de los datos visualizados y así permitir conclusiones más certeras sobre estos.

```
$( ".node" )
.mouseover(function() {
var thisElementTitle = $( this ).find( "title" ).text();
//Color connections
$( "title:contains('--" + thisElementTitle + "')" ).filter(function () {
    return this.innerHTML.match("--" + thisElementTitle + "$");
}).parent().find("path").attr('stroke', '#e57373');
$( "title:contains('" + thisElementTitle + "--')" ).filter(function () {
    return this.innerHTML.match("^" + thisElementTitle + "--");
}).parent().find("path").attr('stroke', '#d1a900');

$( "title:contains('" + thisElementTitle + "--')" ).filter(function () {
    return this.innerHTML.match("^" + thisElementTitle + "--");
}).parent().find("path").attr('opacity', '0.2');


//Change Opacity
$('.edge').children('path').attr('opacity', 0.2);
$( "title:contains('--" + thisElementTitle + "')" ).filter(function () {
    return this.innerHTML.match("--" + thisElementTitle + "$");
}).parent().find("path").attr('opacity', 12);
$( "title:contains('" + thisElementTitle + "--')" ).filter(function () {
    return this.innerHTML.match("^" + thisElementTitle + "--");
}).parent().find("path").attr('opacity', 12);
})


.mouseout(function() {
$('.edge').children('path').attr('opacity', 1.2);
$('.edge').children('path').attr('stroke', 'white');

});
```

Figura 9.1 Código para manejar el focus+context de las conexiones sobre un Gmap

9.1.3 Composición Cromática

Uno de los problemas identificados en la visualización Gmap es que la composición cromática dificulta la lectura de la visualización. A raíz de este problema el profesor Franklin Hernández produjo un artículo [67] sobre el cual se indica cómo generar a partir de la manipulación dinámica del modelo HSB esquemas de colores con armonía ideales para visualizaciones con muchos colores o matices, como ocurre con el Gmap.

Sobre esta teoría, se generó el siguiente esquema cromático: #6d998e, #6c98ad, #498a8c, #548054, #2c5870, #2c855a, #869396, #bcc6cc, #95cbcc, #4e7599, #e8e8e8, #81e6de, #86b59f, #677073, #ffc466 y #e57373 (ver Figura 9.2). Este se empleará en el atributo de nodo “clustercolor” de graphviz.

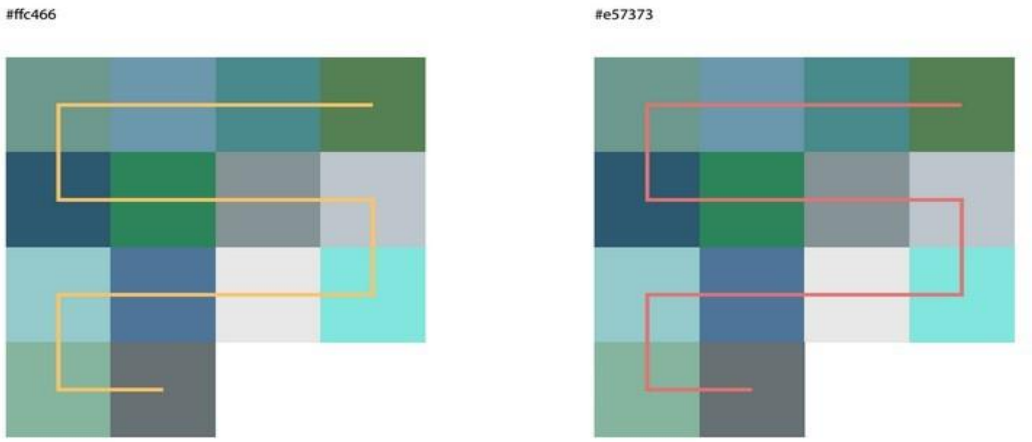


Figura 9.2 Composición cromática propuesta.

9.2 Definición de Casos de Uso

Con el fin de poder determinar si la propuesta de visualización final cumple o no con los objetivos planteados y por ende comprender si se resuelve o no la hipótesis, se define la siguiente lista de casos de uso. Esta representa una lista de diez preguntas (ver cuadro 9.1) que abarcan escenarios a los cuales se espera el paradigma de visualización final pueda dar respuesta.

Dichos casos de uso se encuentran agrupados en 3 grandes grupos, los cuales se describen a continuación:

1. Influencia: A partir de la visualización se quiere inferir aquellos individuos que pudieran ser factores de influencia en un tema específico en sus grupos o bien en la población en general.
2. Conocimiento: Se quiere comprender quiénes tienen conocimiento de un tema y en qué medida. Al mismo tiempo se quiere comprender la presencia o carencia del conocimiento sobre un tema objetivo.
3. Organización: De la visualización se busca adquirir información de peso para decisiones ejecutivas que impliquen estrategias de negocios relacionadas a reorganizaciones, aperturas o cierres de grupos, establecimiento de nuevos negocios o bien reforzamiento de las políticas de colaboración y sinergia de la corporación.

Cuadro 9.1 Definición de Casos de Uso para la evaluación del prototipo final

Influencia	*#1 - Es posible poder concluir qué personas pudieran ser agentes de influencia sobre la población en general en un tema específico* *#2 - Es posible poder concluir, qué personas pudieran ser en un tema, agentes de influencia en el grupo al que pertenecen*
Conocimiento	*#3 - Es posible poder observar grupos informales, no organizacionales, formados por un conocimiento en común alrededor de un tópico específico* *#4 - Es posible poder concluir quién pudiera tener cierto conocimiento relacionado a un tema.* *#5 - Es posible observar la distribución del conocimiento relacionado a un tema entre los diferentes grupos*
Organización	*#6 - Es posible poder concluir qué personas pudieran pertenecer a otros grupos de negocios por su afinidad de comunicación sobre un tema* *#7 - Es posible visualizar oportunidades de sinergia entre grupos que tienen buena comunicación sobre un tema* *#8 - Es posible identificar fallas de sinergia en grupos con poca comunicación entre estos sobre un tema sobre los cuales por función organizacional se espera una mayor comunicación* #9 - Es posible deducir la existencia o no de un conocimiento, en pro de invertir en un nuevo nicho negocio asociado a un grupo *#10 - Es posible poder comprender el nivel de sinergia que existe sobre toda la población que habla sobre un tópico específico*

9.2.1 Casos de Uso

A continuación se analizan los distintos casos de uso en detalle

Caso de uso #1: Es posible poder concluir qué personas pudieran ser agentes de influencia sobre la población en general en un tema específico.

Interesa poder conocer aquellas personas que por sus números de conexiones, claramente por encima de otras personas, pudiesen entonces ser considerados como potenciales agentes de influencia en una corporación. Se puede definir como agente de influencia a aquella persona que en sus comunicaciones sobre un tema específico pudiera llegar a muchos otros más por encima de la persona promedio de la población. Por lo general, sería de esperar que dichas personas de influencia fuesen personas altamente reconocidas entre la población y en especial por la gerencia. Sin embargo, en corporaciones de gran tamaño y altamente distribuidas, la comprensión por parte de la gerencia sobre la localización del conocimiento es limitada. Por lo general, se restringe a un pequeño número de personas, lejano al grupo que de manera forma, formal o informal, tienen conocimiento relevante sobre un tema. Por lo tanto, interesa sobretodo conocer a aquellos potenciales agentes de influencia, que no son conocidos de forma previa como tales.

Caso de uso #2: Es posible poder concluir qué personas pudieran ser, en un tema, agentes de influencia en el grupo al que pertenecen.

Mediante la visualización se espera poder comprender qué personas tienen conocimiento sobre un tema y a qué organización estas pertenecen. Poder conocer a aquellas personas que pudieran tener conocimiento relevante sobre un tema es de importancia, dado que estas pudiesen ser potenciales agentes de influencia sobre sus respectivos grupos.

Caso de uso #3: Es posible poder observar grupos informales, no organizacionales, formados por un conocimiento en común alrededor de un tema específico.

En toda organización, surgen grupos informales comúnmente relacionados por intereses comunes. Interesa por lo tanto conocer aquellos grupos, nichos de personas formados por el intercambio de comunicaciones relacionados con un tema específico. Por ejemplo, por la visualización, podríamos comprender aquellas personas con un interés sobre “minería de datos”, buscando aquellas palabras claves que rodean el tema y visualizando los grupos lógicos que se formar a partir de dicha visualización. Se espera que en este tipo casos se pueda observar grupos conformados por los grupos al cual las personas pertenecen, sin embargo, al mismo tiempo, se podría observar personas altamente relacionadas entre sí, las cuales pertenecen a distintos grupos organizacionales.

Caso de uso #4: Es posible poder concluir quién pudiera tener cierto conocimiento relacionado a un tema.

Una de las primeras deducciones a las que se quiere llegar es quiénes pudieran tener un conocimiento sobre un tema sin importar en primera instancia la cantidad de conexiones o bien la frecuencia de las comunicaciones que se pudieran visualizar. Consiguientemente, sobre aquella población que potencialmente pudiera tener un conocimiento de causa sobre un tema, interesa conocer en qué medida dicho conocimiento existe y en qué relación con respecto a la población en general y a sus respectivos grupos.

Caso de uso #5: Es posible observar la distribución del conocimiento relacionado a un tema entre los diferentes grupos.

Como parte de la visualización, se espera comprender la distribución de un conocimiento sobre un tema en específico. Este puede ser el caso de una tecnología disruptiva o bien de un nuevo proceso, el cual pudiera afectar a la corporación en general. En dicho caso, interesa conocer quiénes en la corporación tienen conocimiento sobre el tema y cómo se encuentra distribuido entre los diferentes grupos. El otro ejemplo podría ser el de una metodología que se viene implementando y reforzando su uso por el último año. Interesa en dicho caso poder observar la distribución del conocimiento sobre las comunicaciones de la población, para comprender en qué medida ha sido incorporada la metodología por las personas y en qué grupos existe una baja recepción del tema, con el fin de darle seguimiento.

Caso de uso #6: Es posible poder concluir qué personas pudieran pertenecer a otros grupos de negocios por su afinidad de comunicación sobre un tema.

Las corporaciones están conformadas por localizaciones y sobre cada localización existen grupos de negocios que en ocasiones trabajan de forma conjunta o en ocasiones trabajan de forma aislada. Resulta normal que por asignaciones de mediano o largo plazo miembros de un grupo pudiesen estar más relacionados con personas de otros grupos que con aquellos de su propio grupo. En este caso, la visualización nos permitiría observar aquellas personas cuya comunicación en un tema dado es más fuerte con personas de otro grupo que con aquellas de su propio grupo.

Ante la carencia de comunicación con las personas de su propio grupo, se podría considerar si una persona estaría mejor ubicada organizacionalmente como miembro de otro grupo.

Caso de uso #7: Es posible visualizar oportunidades de sinergia entre grupos que tienen buena comunicación sobre un tema.

En dicho caso de uso, el conocimiento de qué grupos tienen más miembros que se comunican entre sí, ayuda comprender con qué grupos se podrían establecer posible sinergias sobre un tema. Entre mayor sinergia exista entre los grupos sobre un tema, mayor es la probabilidad de que estos sean exitosos siguiendo un fin en común.

Caso de uso #8: Es posible identificar fallas de sinergia entre grupos, que por expectativa corporativa, deben estar altamente relacionados en un tema.

Relacionado al caso anterior, pero visto desde la otra perspectiva en donde el objetivo ya fue dado a dos grupos por mandato organizacional o corporativo, interesa entonces conocer cómo funcionan dichos grupos en pro del objetivo en común. Interesa por ende confirmar la existencia de comunicación entre los grupos, caso contrario se podría deducir que estos trabajan de forma aislada.

Caso de uso #9: Es posible deducir la existencia o no de un conocimiento, en pro de invertir en un nuevo negocio asociado a un grupo.

La visualización que se busca debería permitir comprender de forma clara y concisa el volumen de personas que tienen conocimiento sobre un tema y en qué cantidad. De esta forma podríamos visualizar y deducir qué tan bien preparado o no se encuentra un grupo con respecto a un tema en específico para así plantear estrategias que apoyen a levantar el conocimiento del grupo en el tema de interés para el nuevo negocio. Es relevante entender en cuáles grupos existe una mayor concentración del conocimiento en un tema, con el fin de poderlo usarlo como base para decisiones de inversión en nuevos negocios, que tomen provecho del conocimiento ya existente en los grupos.

Caso de uso #10: Es posible poder comprender el nivel de sinergia que existe sobre toda la población que habla sobre un tema específico.

Uno de los principales objetivos y motivaciones iniciales de esta investigación es el poner los fundamentos para comprender en qué medida existe o no colaboración en una corporación y romper aquellas islas organizacionales que detienen a la corporación de ser más ágil y eficiente. Mediante la visualización se quiere comprender la interacción entre los miembros de la corporación que tienen un conocimiento sobre un tema y poder concluir así que nivel de sinergia existe entre los individuos y entre los grupos en relación con un tema en específico.

10. Evaluación de propuesta de visualización final

La propuesta final de visualización está desarrollada con base en el paradigma Gmaps. Sobre este, se mejoran sus aspectos de frecuencia, interacción y focus+context analizados en la capítulo 9 de este trabajo.

Las mejoras se realizan mediante el uso de jquery (ver figura 9.1), el cual se inserta mediante consola en el buscador web, una vez que la visualización del Gmap ha sido desplegada en su versión web [65]. Dicha implementación facilita la portabilidad del aporte de este trabajo, dado a que el código en github [65] es incompleto, lo que dificulta así todas las capacidades implementadas en su versión web [90].

Al implementar las mejoras analizadas en el capítulo 9, se logra una visualización mucho más apta para poder comprender el problema objetivo y así poder probar si la hipótesis planteada es cierta.

Tal como se muestra en la figura 10.1, se puede observar cómo los cambios efectuados sobre el grosor de las conexiones, así como el tamaño de los nodos, permiten una buena visibilidad de los nodos relevantes en el conocimiento del tema estudiado. Por otra parte, la armonía de los colores es una clara mejora en producir una visualización fácil de leer y que gusta como representación visual que es. Por último, la implementación de la interacción sobre los nodos permite una clara visualización de las conexiones entrantes y salientes, resaltando las salientes (coloreadas en un tono amarillo) sobre aquellas conexiones entrantes (coloreadas en un tono rosado).

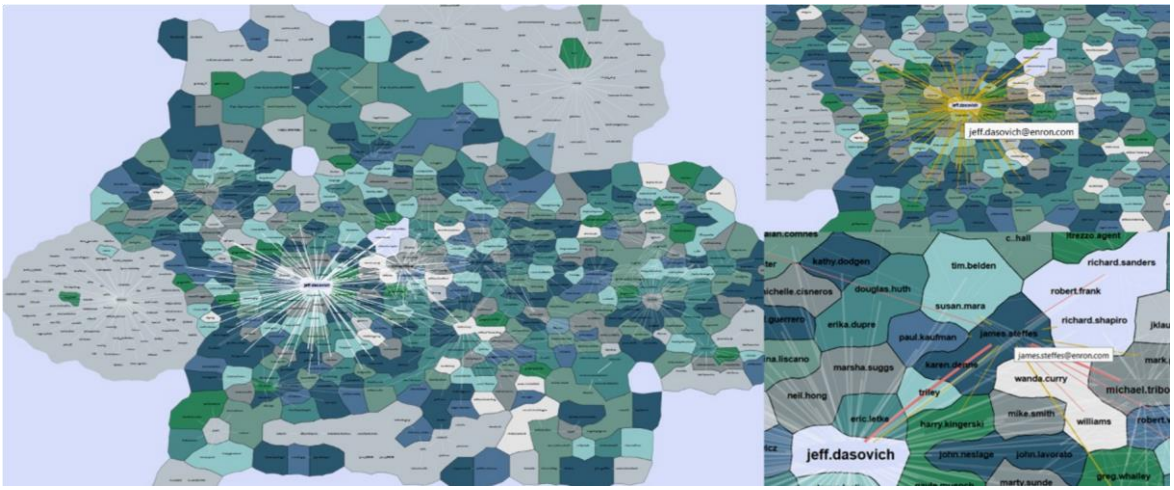


Figura 10.1 Propuesta final con mejoras implementadas.

A continuación se hace un análisis de la visualización propuesta contra cada uno de los casos de uso definidos previamente en la etapa de análisis de la investigación.

10.1 Casos de uso sobre el aspecto de “Influencia”

#1 - Es posible poder concluir qué personas pudieran ser agentes de influencia sobre la población en general en un tema específico (ver figura 10.2).

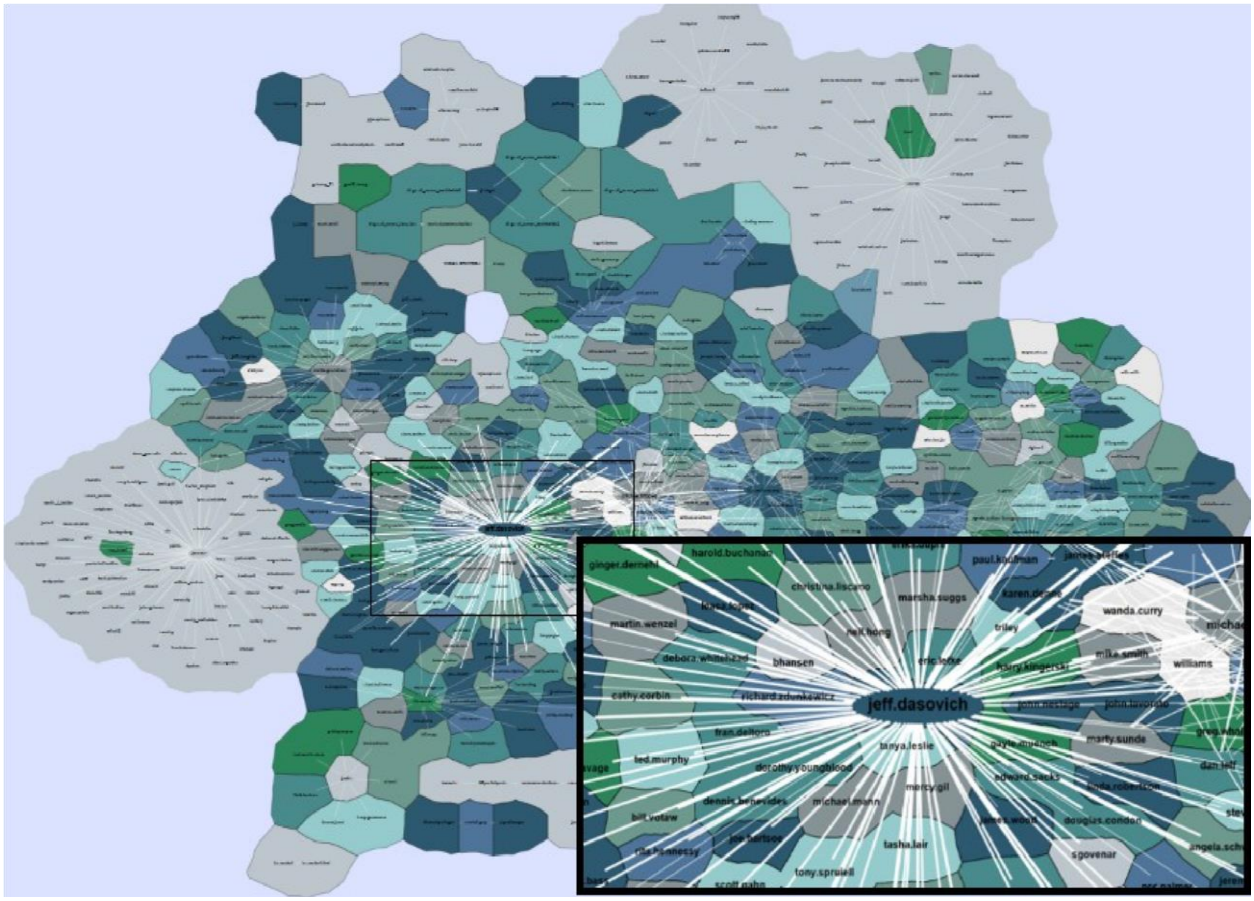


Figura 10.2 Jeff Dasovich (Responsable de Asuntos Gubernamentales) se distingue sobre los demás por la alta frecuencia en sus comunicaciones. Jeff es claramente un agente de influencia en el caso de defensa del caso Enron.

#2 - Es posible poder concluir qué personas pudieran ser agentes de influencia, en el grupo al que pertenecen, en un tema (ver Figura 10.3).

Ciertamente se pueden concluir quiénes son los agentes de influencia en el plano general de las comunicaciones (ver figura 10.2). De igual forma, se pueden observar los agentes de influencia sobre tanto grupos informales, conformados por miembros de varios grupos formales, o bien las personas que pudiesen ser agentes de influencia en

sus grupos (ver figura 10.3), caracterizados por el color con que cada grupo en el mapa se diferencia.

En el caso de los datos relacionados al conocimiento del tema de “corrupción” de Enron, se puede claramente visualizar la predominancia del señor Jeff Dasovich con respecto al resto de los actores en el mapa. Según [58][85], el señor Dasovich era en ese entonces el ejecutivo de asuntos gubernamentales, y actor central en el manejo y la defensa de Enron ante los cuestionamientos por parte del gobierno de los Estados Unidos con respecto a las prácticas financieras de la corporación. De esta forma se puede claramente comprobar que la visualización es efectiva en identificar a aquellas personas con conocimiento relevante que pudiesen ser agentes de influencia.

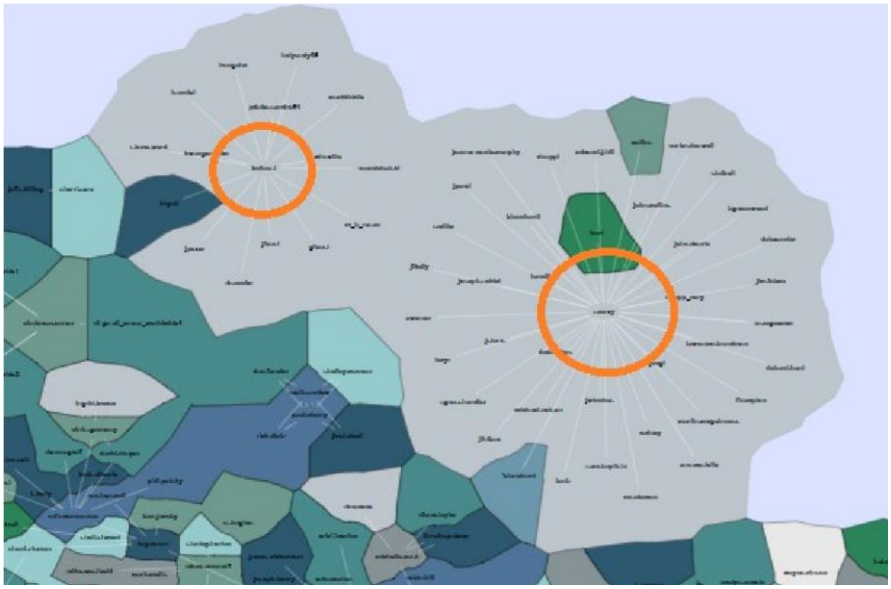


Figura 10.3 Claros agentes de influencia del grupo gris.

10.2 Casos de uso sobre el aspecto de “Conocimiento”

#3 - Es posible poder observar grupos informales, no organizacionales, formados por un conocimiento en común alrededor de un tema específico (ver Figura 10.4).

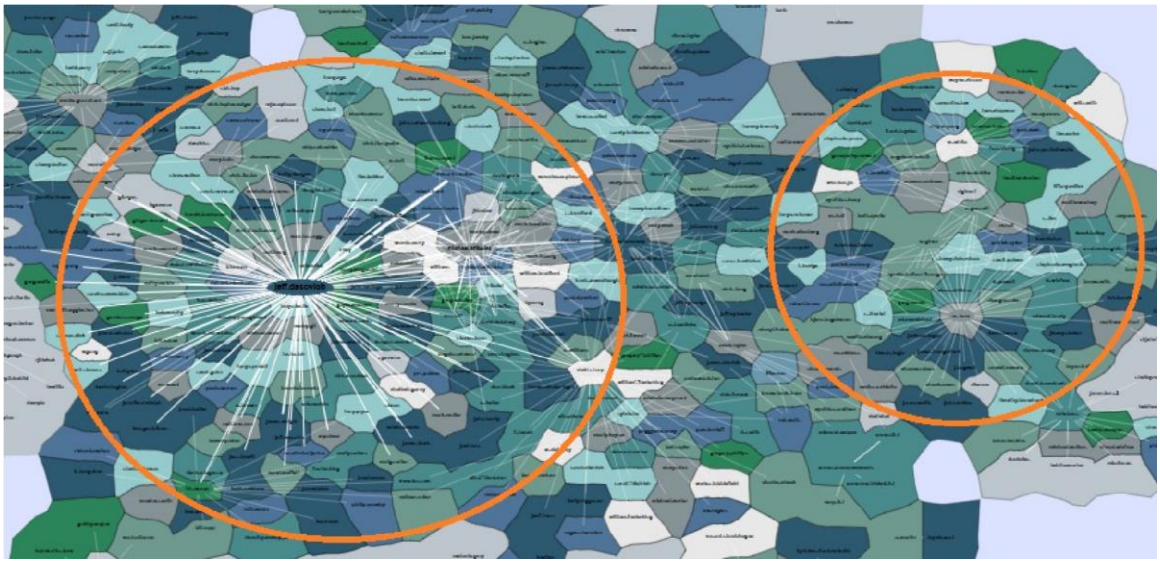


Figura 10.4 Grupos informales formados por la alta relación en la comunicación de sus miembros.

#4 - Es posible poder concluir quién pudiera tener cierto conocimiento relacionado a un tema (ver figura 10.5).

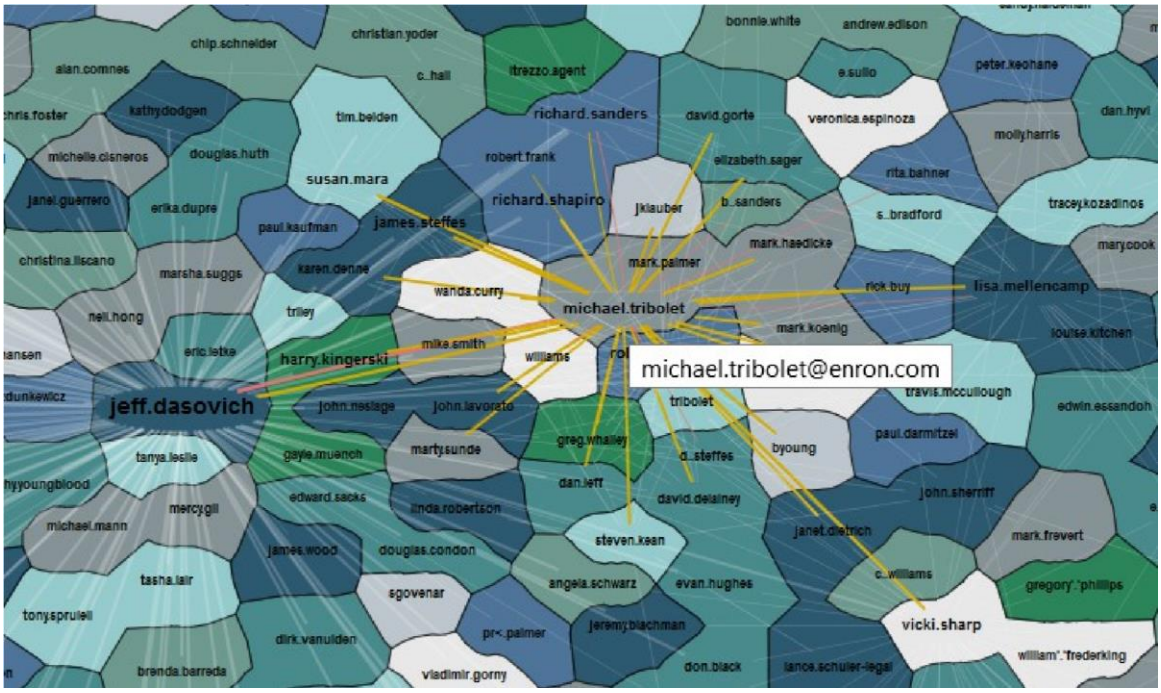


Figura 10.5 Persona con pocas, pero muy fuertes conexiones con personas claves.

#5 - Es posible observar la distribución del conocimiento relacionado a un tema entre los diferentes grupos (ver figura 10.6).

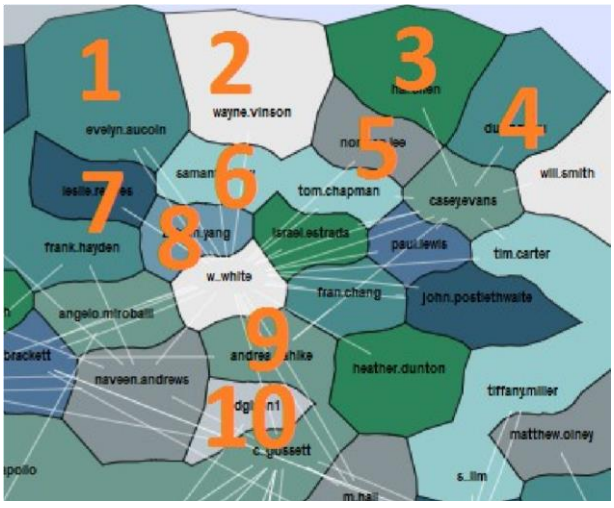


Figura 10.6 Comunidad conformada por personas de 10 diferentes grupos.

Al evaluar los casos de uso con respecto al conocimiento, encontramos que ciertamente es posible inferir en dónde existe conocimiento y en qué medida. Se pueden claramente observar grupos informales conformados por personas de distintos grupos (ver figura 10.4).

10.3 Casos de uso sobre el aspecto de “Organización”

#6 - Es posible poder concluir qué personas pudieran pertenecer a otros grupos de negocios por su afinidad de comunicación sobre un tema (ver figura 10.7).

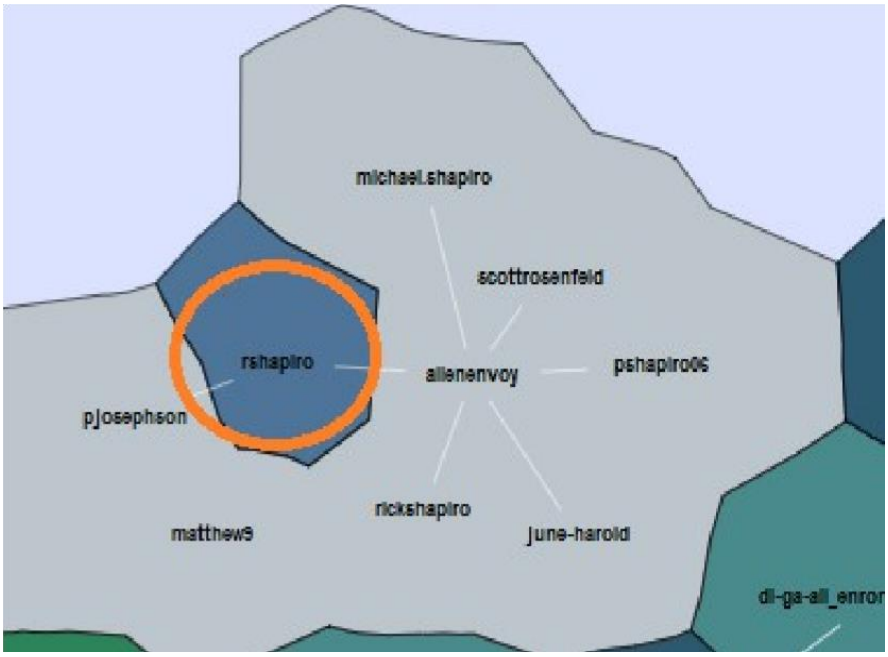


Figura 10.7 Persona que pudiera pertenecer al grupo gris.

#7 - Es posible visualizar oportunidades de sinergia entre grupos que tienen buena comunicación sobre un tema (ver figura 10.8)

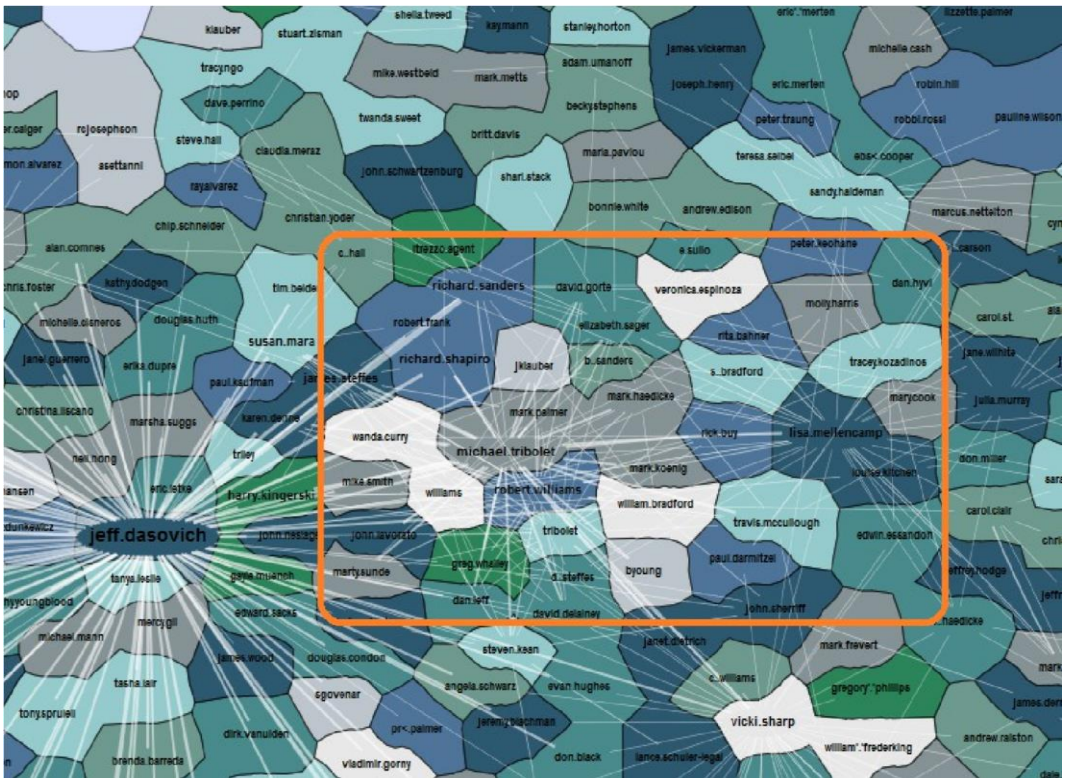


Figura 10.8 Miembros de diversos grupos que evidencia una fuerte sinergia entre ellos.

#8 - Es posible identificar fallas de sinergia entre grupos, que por su función organizacional, deben estar altamente relacionados en un tema (ver figura 10.9).

Figura 10.9 Se pueden evidenciar pocas comunicaciones, entre miembros del mismo grupo y con miembros de otros grupos.

#9 - Es posible deducir la existencia o no de un conocimiento, en pro de invertir en un nuevo nicho de negocio asociado a un grupo (ver figura 10.10).

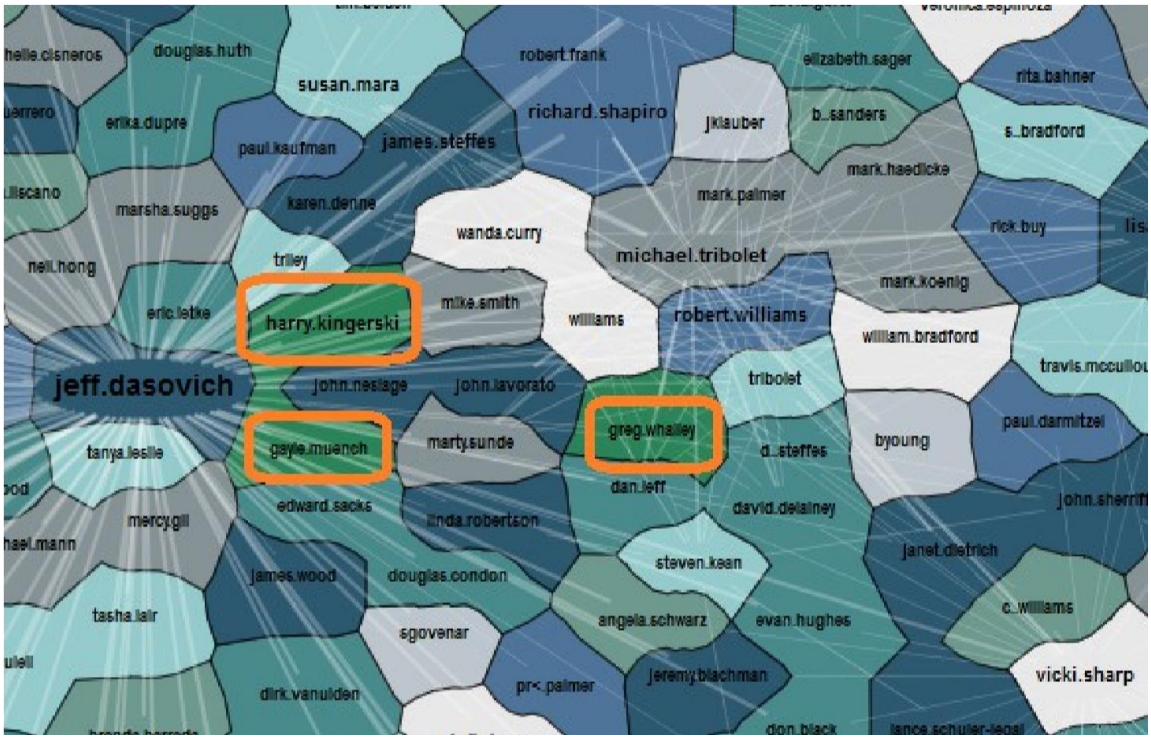


Figura 10.10 Tres personas del grupo verde se encuentran en el centro de las comunicaciones fuertes de la red.

#10 - Es posible poder comprender el nivel de sinergia que existe sobre toda la población que habla sobre un tema específico (ver Figura 10.11).

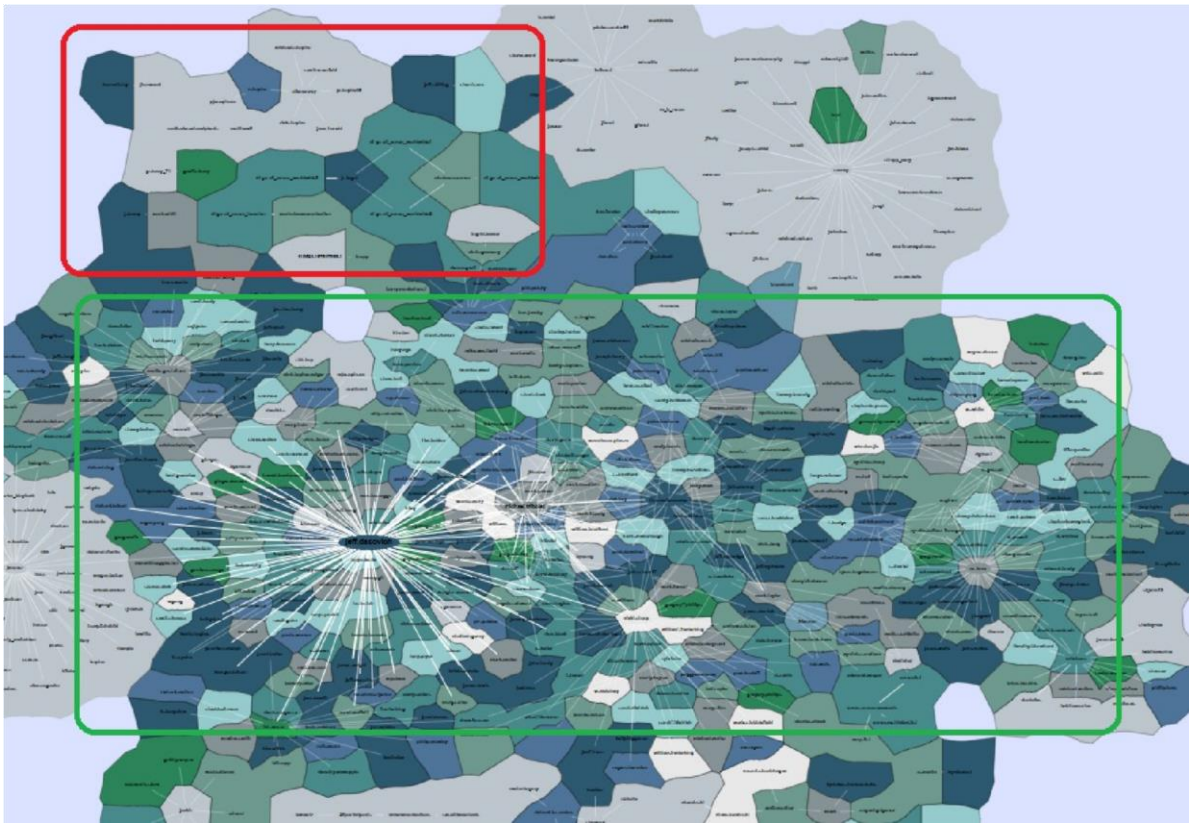


Figura 10.11 Sectores que evidencian sinergia, donde rojo representa poca sinergia y verde significa una importante sinergia.

La visualización puede aportar grandemente a comprender personas con gran afinidad en su comunicación con otros grupos (ver figura 10.7). Permite visualizar la medida de sinergia entre grupos (ver figuras 10.8, 10.9 y 10.11) y ayuda a descubrir personas con conocimiento en grupos, que pudiesen justificar la apertura de un nuevo negocio (ver figura 10.10). Es entonces posible comprender aspectos de organización y sinergia con respecto a la localización del conocimiento para tomar decisiones que pudiesen afectar la estructura organizacional de una corporación.

10.4 Conclusiones sobre la evaluación final

Todos los casos de uso evaluados sobre la visualización propuesta fueron satisfactorios (ver cuadro 10.1). De esta forma, se confirma el cumplimiento de los objetivos propuestos en este trabajo. Concluimos entonces que sí es posible identificar la localización del conocimiento dentro de la corporación en temas específicos a través de la visualización del correo electrónico corporativo.

Cuadro10.1 Resultados de las evaluaciones de los casos de uso sobre el prototipo final.

Influencia	*#1 - Es posible poder concluir qué personas pudieran ser agentes de influencia sobre la población en general en un tema específico*	Verdadero
	#2 - Es posible poder concluir, qué personas pudieran ser agentes de influencia, en el grupo al que pertenecen, en un tema.	Verdadero
Conocimiento	*#3 - Es posible poder observar grupos informales, no organizacionales, formados por un conocimiento en común alrededor de un tópico específico*	Verdadero
	#4 - Es posible poder concluir quién pudiera tener cierto conocimiento relacionado a un tema.	Verdadero
	#5 - Es posible observar la distribución del conocimiento relacionado a un tema entre los diferentes grupos	Verdadero
Organización	*#6 - Es posible poder concluir qué personas pudieran pertenecer a otros grupos de negocios por su afinidad de comunicación sobre un tema*	Verdadero
	#7 - Es posible visualizar oportunidades de sinergia entre grupos que tienen buena comunicación sobre un tema	Verdadero
	#8 - Es posible identificar fallas de sinergia entre grupos, que por su función organizacional, deben estar altamente relacionados en un tema.	Verdadero
	#9 - Es posible deducir la existencia o no de un conocimiento, en pro de invertir en un nuevo nicho negocio asociado a un grupo	Verdadero
	#10 - Es posible poder comprender el nivel de sinergia que existe sobre toda la población que habla sobre un tópico específico	Verdadero

11. Conclusiones

Como conclusiones principales del trabajo realizado se pueden citar las siguientes:

1. El paradigma de visualización Gmaps es una excelente visualización para comprender la localización del conocimiento de una población.

2. Es posible abstraer, de los títulos de los correo electrónicos corporativos, las personas con conocimiento relevante sobre un tema.

3. Los cambios sobre el grosor de las líneas de conexión y el tamaño del nombre de la persona, aportan claridad para observar las personas de influencia y distribución del conocimiento en el Gmap.

4. La interacción propuesta en este trabajo aporta focus+context a las conexiones del Gmaps, lo que permite una mejor comprensión de las comunicaciones entre las personas en el mapa.

5. La selección de las palabras clave que componen un tema deben ser específicas al tema y que no formen parte del lenguaje común de la corporación.

6. El mecanismo de extracción de datos implementado, siguiendo la filosofía de map-reduce, es capaz de funcionar en bases de datos corporativas de gran tamaño.

7. La metodología empleada permitió seleccionar el paradigma que permite resolver el problema objetivo y enriquecerlo con fortalezas identificadas sobre los otros paradigmas.

12. Recomendaciones

1. La interacción sobre el Gmap podría mejorarse, si adicional al resalte de las conexiones, se resaltan los nodos partícipes en la comunicación.

2. Interesa aplicar los aportes de esta investigación sobre un correo electrónico corporativo sobre el cual se tenga acceso directo a los actores de la corporación, y así poder identificar otras conclusiones que se pudieran derivar de la visualización.

3. Es deseable dotar a la visualización Gmap con mecanismos que permitan un mayor nivel de detalle sobre los correos que componen las conexiones

4. Tal como se evidenció en las evaluaciones de los paradigmas, cada paradigma de visualización tiene sus propias virtudes. Resultaría interesante una aplicación que mezclase las distintas visualizaciones con el fin de lograr abstraer un mayor número de conclusiones a partir de la visualización

13. Bibliografía

[1] Luukkonen, T., et al. "The measurement of international scientific collaboration." Scientometrics, vol. 28, no. 1, pp. 15-36, 1993.

[2] Bjork, S. and J. Redstrom. "Redefining the focus and context of focus+ context visualization." In Information Visualization, 2000. InfoVis 2000. IEEE Symposium, pp. 85-89. IEEE, 2000.

[3] Herman, I., G. Melançon, and M. S. Marshall. "Graph visualization and navigation in information visualization: A survey." Visualization and Computer Graphics, IEEE Transactions, vol. 6, no. 1, pp. 24-43, 2000.

[4] Kosara, R., S. Miksch, and H. Hauser. "Focus+ context taken literally." Computer Graphics and Applications, IEEE, vol. 22, no. 1, pp. 22-29, 2002.

[5] Keim, D. A. "Information visualization and visual data mining". Visualization and Computer Graphics, IEEE Transactions, vol. 8, no. 1, pp. 1-8, 2002.

[6] Gloor, P.A., et al. "Visualization of communication patterns in collaborative innovation networks-analysis of some w3c working groups." In Proceedings of the twelfth international conference on Information and knowledge management, pp. 56-60. ACM, 2003.

[7] Mueller-Prothmann, T. and I.Finke. "SELaKT-Social Network Analysis as a Methodfor Expert Localisation and Sustainable Knowledge Transfer." J. UCS, vol.10, no. 6, pp.691-701, 2004.

[8] Burkhard, R. A. "Learning from architects: the difference between knowledge visualization and information visualization." In Information Visualisation, 2004. IV 2004.

Proceedings. Eighth International Conference, pp. 519-524. IEEE, 2004.

[9] Heer, J. and D. Boyd. "Vizster: Visualizing online social networks." In Information Visualization, 2005. INFOVIS 2005. IEEE Symposium, pp. 32-39. IEEE, 2005. [10] Bird, C., et al. "Mining email social networks." In Proceedings of the 2006 international workshop on Mining software repositories, pp. 137-143. ACM, 2006. [11] Yang, C.C., N. Liu, and M. Sageman. "Analyzing the terrorist social networks with visualization tools." In Intelligence and security informatics, pp. 331-342. Springer Berlin Heidelberg, 2006.

[12] Hauser, H. "Generalizing focus+ context visualization." In Scientific visualization: The visual extraction of knowledge from data, pp. 305-327. Springer Berlin Heidelberg, 2006.

[13] Shneiderman, B. and A. Aris. "Network visualization by semantic substrates." Visualization and Computer Graphics, IEEE Transactions, vol. 12, no. 5, pp. 733-740, 2006.

[14] Wen, Z. and M.X. Zhou. "Evaluating the use of data transformation for information visualization." Visualization and Computer Graphics, IEEE Transactions, vol. 14, no. 6, pp.1309-1316, 2008.

[15] Simoff, S.J. "Form-Semantics-Function–A Framework for Designing Visual Data Representations for Visual Data Mining." In Visual Data Mining, pp. 30-45. Springer Berlin Heidelberg, 2008.

[16] Zhang, J., C. Chen, and J. Li. "Visualizing the intellectual structure with paper-reference matrices." Visualization and Computer Graphics, IEEE Transactions, vol. 15, no. 6, pp.1153-1160, 2009.

[17] Gansner, E., et al. "Putting recommendations on the map: visualizing clusters and relations." In Proceedings of the third ACM conference on Recommender systems, pp. 345-348. ACM, 2009.

[18] Gansner, E.R., Y. Hu, and S.G. Kobourov. "Gmap: Drawing graphs as maps." InGraph Drawing, pp. 405-407. Springer Berlin Heidelberg, 2010.

[19] Nankani, E., et al. "Enterprise university as a digital ecosystem: Visual analysis ofacademic collaboration." In Digital Ecosystems and Technologies, 2009. DEST'09. 3rd IEEE International Conference, pp. 727-732. IEEE, 2009.

[20] Wang, S., et al. "2.5 D Focus+ Context Map Visualization." In Computer-AidedDesign and Computer Graphics (CAD/Graphics), 2011 12th International Conference, pp. 389-396. IEEE, 2011.

[21] Aigner, W., et al. "Survey of Visualization Techniques." In Visualization of TimeOriented Data, pp. 147-254. Springer London, 2011.

[22] Lam, H., et al. "Empirical studies in information visualization: Seven scenarios". Visualization and Computer Graphics, IEEE Transactions, vol. 18, no. 9, pp.1520-1536, 2012.

[23] McGuffin, M.J. "Simple algorithms for network visualization: A tutorial."Tsinghua Science and Technology, vol.17, no. 4, pp. 383-398, 2012.

[24] Alsukhni, M. Interactive visualization of the collaborative research network. PhDdiss., University of Ontario Institute of Technology, 2012.

[25] Alsukhni, M. and Y. Zhu. "Interactive visualization of the social network of research collaborations." In Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference, pp. 247-254. IEEE, 2012.

[26] Pfeffer, J. "Fundamentals of visualizing communication networks." Communications, China, vol.10, no. 3, pp. 82-90, 2013.

[27] Stewart, S.A. and S.S. Raza Abidi. "Using Social Network Analysis to Study the Knowledge Sharing Patterns of Health Professionals Using Web 2.0 Tools." In Biomedical Engineering Systems and Technologies, pp. 335-352. Springer Berlin Heidelberg, 2013.

[28] Kosara, R. Semantic Depth of Field-Using Blur for Focus+ Context Visualization. PhD Thesis, Vienna University of Technology, Austria, 2001.

[29] Iba, T., et al. "Analyzing the creative editing behavior of Wikipedia editors: through dynamic social network analysis." Procedia-Social and Behavioral Sciences, vol. 2, no. 4, pp. 6441-6456, 2010.

[30] Kidane, Y.H. and P. A. Gloor. "Correlating temporal communication patterns of the Eclipse open source community with performance and creativity." Computational and mathematical organization theory, vol.13, no. 1, pp.17-27, 2007.

[31] Schoder, D., P. A. Gloor, and P.T. Metaxas. "Social Media and Collective Intelligence—Ongoing and Future Research Streams." KI-Künstliche Intelligenz, vol. 27, no. 1, pp. 9-15, 2013.

[32] PA Gloor, Measuring Creative Performance of Teams Through Dynamic Semantic Social Network Analysis (Video of Presentation at Conference), COSM2013. <https://www.youtube.com/watch?v=AutfpHry7Dc, 2013>, [Consulta: 11/11/2014].

[33] Cross, R., A. Parker, and S. P. Borgatti. "A bird's-eye view: Using social network analysis to improve knowledge creation and sharing." IBM Institute for Business Value, pp.1669-1600, 2002.

[34] Bastian, M., S. Heymann, and M. Jacomy. "Gephi: an open source software for exploring and manipulating networks." ICWSM, vol. 8, pp.361-362, 2009.

[35] Dadzie, A.S., and M. Rowe. "Approaches to visualising linked data: A survey." Semantic Web, vol. 2, no. 2, pp. 89-124, 2011.

[36] Adamic, L. and E. Adar. "How to search a social network." Social Networks, vol. 27,no. 3, pp. 187-203, 2005.

[37] Hanneman, R. A., and M. Riddle. Introduction to social network methods. 2005.

[38] Pancho, D.P., et al. "FINGRAMS: visual representations of fuzzy rule-based inference for expert analysis of comprehensibility." Fuzzy Systems, IEEE Transactions, vol. 21, no. 6, pp.1133-1149, 2013.

[39] Viégas, F.B., S. Golder, and J. Donath. "Visualizing email content: portraying relationships from conversational histories." In Proceedings of the SIGCHI conference on Human Factors in computing systems, pp. 979-988. ACM, 2006.

[40] Viégas, F.B., et al. "Digital artifacts for remembering and storytelling: posthistoryand social network fragments." In System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference , pp. 10-pp. IEEE, 2004.

[41] Tichy, N.M., M.L. Tushman, and C. Fombrun. "Social network analysis for organizations." Academy of management review , vol. 4, no. 4, pp. 507-519, 1979.

[42] Schemaball http://mkweb.bcgsc.ca/schemaball/

[43] Constantine, D. “Close-ups of the genome, species by species by species”. NewYork Times F 4, 2007.

[44] Gansner, E.R., Y. Hu, and S. Kobourov. "GMap: Visualizing graphs and clusters as maps." In Pacific Visualization Symposium (PacificVis), 2010 IEEE, pp. 201-208. IEEE, 2010.

[45] Visualizing Data with Graphs and Maps, Yifan Hu, AT&T Labs Research, May 7,2012. Gaithersburg <http://math.nist.gov/mcsd/Seminars/2012/2012-05-07-Hupresentation.pdf>, [Consulta: 11/11/2014].

[46] Cluster relations in a graph highlighted using gvmap, <http://www.graphviz.org/Gallery/undirected/gd_1994_2007.html>, [Consulta: 11/11/2014].

[47] Cheng, C. et al. "Map-reduce for machine learning on multicore." Advances inneural information processing systems, vol. 19, pp. 281, 2007.

[48] Dean, J. and S. Ghemawat. “Map reduce: Simplified data processing on large clusters”, in Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation, vol.6, pp. 137–149, San Francisco, CA. 2004.

[49] Ekanayake, J., S. Pallickara, and G. Fox. "Map reduce for data intensive scientific analyses." In eScience, 2008.eScience'08. IEEE Fourth International Conference on, pp. 277-284. IEEE, December, 2008.

[50] Hernández-Castro, F. y J. Monge. “Estrategias de Representación de Estructuras Jerárquicas”, Tiempo Compartido, Vol. 7, No 4, pp. 31, October 21, 2007

[51] B. J. “Simple Mail Transfer Protocol”, Posted in august 1982. <http://tools.ietf.org/html/rfc821>, [Consulta: 11/11/2014].

[52] Bostock, M., V. Ogievetsky and J. Heer. “D3-InfoVis”. October 2011. <http://vis.stanford.edu/files/2011-D3-InfoVis.pdf>, [Consulta: 11/11/2014].

[53] Clark, J. “Movement in Manhattan”, May 7, 2012 <http://vimeo.com/41703644 >, [Consulta: 11/11/2014].

[54] Clark, J. “Movement in Manhattan Video”, May 8, 2012 <neoformix.com/2012/MovementInManhattanVideo.html>, [Consulta: 11/11/2014].

[55] Clark, J. “Movement in Manhattan”, Apr 18, 2012

<http://neoformix.com/2012/MovementInManhattan.html>, [Consulta: 11/11/2014]. [56] Dot (Graph Description Language),<http://en.wikipedia.org/wiki/DOT_%28graph_description_language%29>, [Consulta: 11/11/2014].

[57] D3, Bundle Layout, <https://github.com/mbostock/d3/wiki/Bundle-Layout>,[Consulta: 11/11/2014].

[58] Eichenwald, K. W., “We Have a Problem ..”.The New York Times, march 20, 2005, <http://www.nytimes.com/2005/03/20/business/yourmoney/20book.html?pagewanted=1&_r=0>, [Consulta: 11/11/2014].

[59] Enron Scandal, <http://en.wikipedia.org/wiki/Enron_scandal>, [Consulta: 11/11/2014].

[60] Enron Email Dataset, <https://www.cs.cmu.edu/~./enron/ >, [Consulta: 11/11/2014].

[61] Enron Corporation, <https://www.tshaonline.org/handbook/online/articles/doe08>

[62] Floridi, L. “La Infosfera” <http://axisvega.wordpress.com/la-infoesfera-lucianofloridi/>, [Consulta: 11/11/2014].

[63] Gansner, E., E. Koutsofios and S. North. “Drawing graphs with dot”. Nov 2, 2010 <http://www.graphviz.org/pdf/dotguide.pdf>, [Consulta: 11/11/2014].

[64] Gansner, E. “Using Graphviz as a Library (cgraph version)”. published online August 21, 2014. <http://www.graphviz.org/doc/libguide/libguide.pdf>, [Consulta: 11/11/2014].

[65] Gmap github, <https://github.com/spupyrev/gmap>, [Consulta: 11/11/2014].

[66] Health infoscape, <http://senseable.mit.edu/healthinfoscape/>, [Consulta: 11/11/2014].

[67] Hernández-Castro, F. “Combinando colores desde código”, [Documento en línea], <http://skizata.com/combinando-colores.html>, [Consulta: 11/11/2014].

[68] Hierarchical Edge Bundling TreeMap, Dec. 19, 2012.<http://bl.ocks.org/mbostock/4341134>, [Consulta: 11/11/2014].

[69] Hodgman, R. “Chord Layout”, edited Nov 15, 2014 <https://github.com/mbostock/d3/wiki/Chord-Layout>, [Consulta: 11/11/2014].

[70] US Court Enron Case <http://www.gpo.gov/fdsys/pkg/USCOURTS-txsb-4_03-ap-03721/mods.xml>, [Consulta: 11/11/2014].

[71] Infosphere, <http://en.wikipedia.org/wiki/Infosphere>, [Consulta: 11/11/2014].

[72] Ipod, <http://en.wikipedia.org/wiki/IPod>, [Consulta: 11/11/2014].

[73] Kreil, M. “Fireflies HD. Iphone Location Data Localization”, July 12, 2011 <http://crowdflow.net/2011/07/12/fireflies-hd/>, [Consulta: 11/11/2014].

[74] Krzywinski, M.I et al. “Circos: An information aesthetic for comparative genomics”. Genome Res. published online June 18, 2009. <http://circos.ca/>, [Consulta: 11/11/2014].

[75] Multipurpose Internet Mail Extensions(MIME) RFCs

Part 1 Format of Internet Message Bodies, <http://tools.ietf.org/html/rfc2045>

Part 2 Media Types, <http://tools.ietf.org/html/rfc2046>

Part 3 Message Header Extensions for Non-ASCII Text, <http://tools.ietf.org/html/rfc2047>

Media Type Specifications and Registration Procedures, <http://tools.ietf.org/html/rfc4288>

Part 4 Registration Procedures, <http://tools.ietf.org/html/rfc4289>

Part 5 Conformance Criteria and Examples, <http://tools.ietf.org/html/rfc2049>, [Consulta: 11/11/2014].

[76] Personal Storage Table, <http://en.wikipedia.org/wiki/Personal_Storage_Table>, [Consulta: 11/11/2014].

[77] Thomas, T.L. “Infosphere Threats” in Military Review, September-October,1999. <http://fmso.leavenworth.army.mil/documents/infosphere/infosphere.htm>, [Consulta: 11/11/2014].

[78] USA Wind Maps, <http://hint.fm/wind/>, [Consulta: 11/11/2014].

[79] Visualizing information flow in science, <http://well-formed.eigenfactor.org/map.html#/?id=4592>, [Consulta: 11/11/2014].

[80] Wenceslao Galán, “La infosfera y las nuevas patologías”. Lectura de La Fábrica dela Infelicidad, de Franco Berardi (Bifo), Traficantes de Sueños (2002). <http://www.espaienblanc.net/La-infosfera-y-las-nuevas.html>, [Consulta: 11/11/2014].

[81] Yifan, H. “Visualizing Data with Graphs and Maps”. AT&T Labs Research. National Institute of Standards and Technology (NIST), Gaithersburg. May 7, 2012 <http://math.nist.gov/mcsd/Seminars/2012/2012-05-07-Hu.html>, [Consulta: 11/11/2014].

[82] Fruchterman, T. M. J., & Reingold, E. M. “Graph Drawing by Force-Directed Placement” Software: Practice and Experience, vol. 21(1 1), pp. 1129-1164, 1991

[83] Brandon, H., et al. "Visualizing collaboration and influence in the open-source software community." in Proceedings of the 8th Working Conference on Mining Software Repositories, pp. 223-226. ACM, 2011.

[84] Hernández-Castro, F., J Monge Visualización tridimensional de estructuras jerárquicas. Tesis de Maestria, Instituto Tecnológico de Costa Rica, Enero, 2006.

[85] Grieve, T. “The decline and fall of the Enron empire” http://www.salon.com/2003/10/14/enron_22/, Oct 14, 2003, [Consulta: 11/11/2014].

[86] Hierarchical Edge Bundling Circos, June 24, 2011.<http://bl.ocks.org/mbostock/1044242>, [Consulta: 11/11/2014].

[87] Klimt, B., and Yiming Y. "The enron corpus: A new dataset for email classification research." In Machine learning: ECML 2004, pp. 217-226. Springer Berlin Heidelberg, 2004.

[88] Sander, N., et al. Visualising migration flow data with circular plots. No. 2/2014. Vienna Institute of Demography Working Papers, 2014.

[89] Kobourov, S. et al. "Visualizing graphs as maps with contiguous regions."EuroVis14, Accepted to appear (2014).

[90] Kobourov, S. et al. “Gmap Web Version”, Department of Computer Science, University of Arizona, Tucson, AZ, USA <gmap.cs.arizona.edu> [Consulta: 11/11/2014].