Universidad del Valle de Guatemala

Data Science

Laboratorio No.6

Primero se importaron todas las librerías y módulos que se iban a utilizar en el laboratorio. Una vez hecho esto, se procedió con la carga de datos.

En el primer bloque se realizó la carga y normalización de los datos del archivo "trafico.txt". Para esto, se abrió el archivo usando codificación utf-16 y se revisó línea por línea. Durante este proceso se limpiaron los espacios en blanco y se verificó que cada línea tuviera la estructura correcta de un objeto JSON. Las líneas que no cumplían con este formato se descartaron automáticamente.

Cuando se intentaba convertir cada línea a formato JSON, si aparecía algún error de sintaxis se aplicaba una corrección automática para eliminar las comas extras que pudieran estar causando problemas. Después, todos los registros que pasaron la validación se guardaron en una lista, y finalmente se transfirió toda esta información a un dataframe de pandas.

Aunque ya se había comenzado con un preprocesamiento básico de los datos, todavía quedaban pendientes otros aspectos como la eliminación de enlaces y otra información irrelevante para el análisis. Sin embargo, se logró asegurar una importación correcta de los datos y se obtuvo una base de datos estructurada, mejor preparada para manejar las inconsistencias del archivo original y lista para las etapas de limpieza y análisis posteriores.

Una vez completado ese paso, se procedió con el preprocesamiento de los datos. En esta etapa se limpió y homogenizó el texto, se normalizaron los actores y se identificaron las interacciones básicas para preparar el corpus de trabajo.

Para comenzar, se descargaron las stopwords y se tradujeron al español con el fin de eliminar palabras que no aportan información relevante. Después se limpió la columna "Raw\_content" (se puede ver parte de este dataframe en el jupyter adjunto). El proceso de limpieza incluyó convertir cada tweet a minúsculas, eliminar enlaces, menciones y hashtags del texto, y quitar caracteres especiales, emojis, números y signos de puntuación. El resultado se guardó en una nueva columna llamada "texto limpio".

Al mismo tiempo, se homogenizaron las identidades de los usuarios convirtiendo sus nombres a cadenas de texto sin espacios, y se extrajeron las menciones explícitas que aparecían en los tweets.

Para evitar sesgos causados por contenido repetido, se tomó en cuenta la fecha de publicación de cada tweet y se eliminaron los duplicados usando una clave triple: usuario-texto-fecha. Este filtrado redujo la base de datos a 5,596 registros, que serían la base para todos los análisis posteriores.

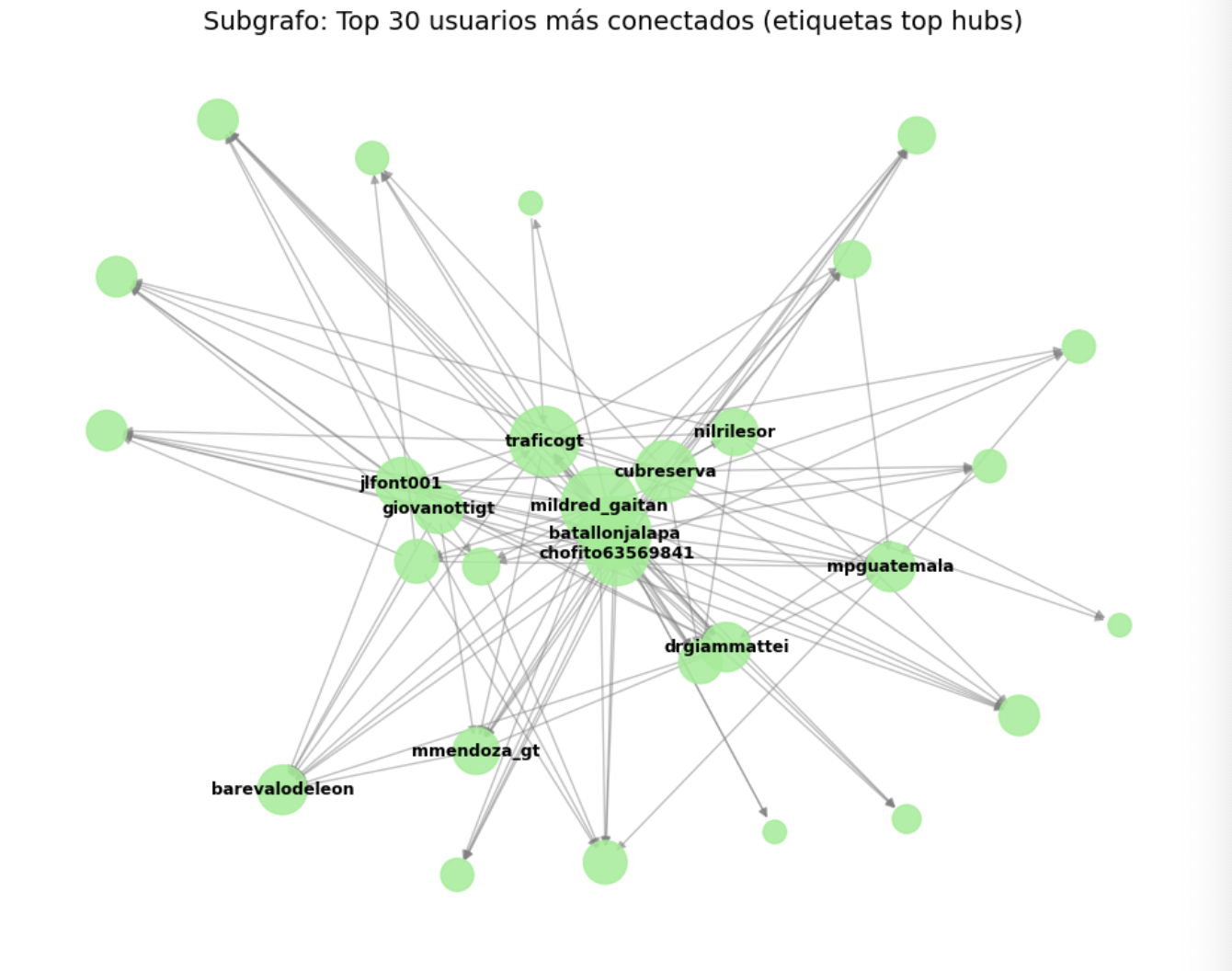
Finalmente, se completó la clasificación de interacciones identificando los tweets que contenían menciones. Con esta información se creó un dataframe que incluye las columnas mentions, hashtags, is\_retweet e is\_reply, datos necesarios para construir el grafo que se solicitaba en el laboratorio.

Una vez completado el preprocesamiento, se procedió con el análisis exploratorio de datos. Esta etapa tenía como objetivo obtener una primera aproximación descriptiva del comportamiento de los usuarios y del contenido que publicaban. Las estadísticas generales de los datos fueron las siguientes:

* 5,596 tweets en total
* 2,071 usuarios únicos
* 11,009 menciones
* 592 hashtags

Estas cifras reflejan que existe una conversación amplia y muy activa en Twitter (X) sobre el tráfico en Guatemala. Al analizar las cuentas más mencionadas, se encontró que @traficogt acumula más de 4,000 menciones, confirmando su papel central en esta red de conversación. Pero a partir de esto es natural preguntarse ¿Qué usuarios aparecen como los más mencionados y qué papel juegan en la conversación?

De hecho el análisis reveló que unos pocos usuarios concentran la mayoría de las menciones. Esto sugiere que estas son cuentas centrales dentro de la red, probablemente medios de comunicación, instituciones oficiales o figuras públicas.  Entre estas menciones están cuentas importantes como @barevalodeleon (presidente electo de Guatemala 2024-2028), @drgiammattei (expresidente de la república) y @amilcarmontejo (vocero de @traficogt), entre otros. El gráfico correspondiente muestra las principales cuentas dentro de todas las interacciones, destacando aquellas con mayor relevancia o impacto.



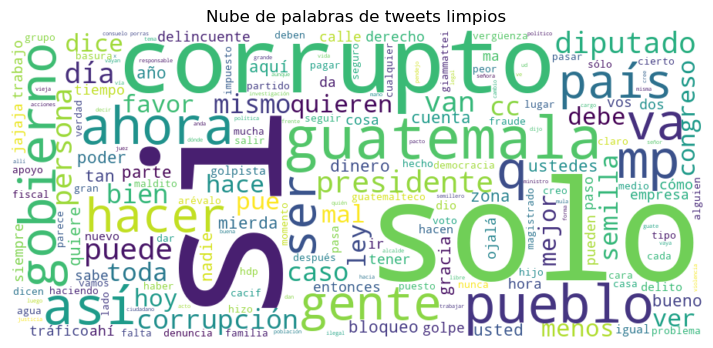
Estos resultados sugieren que la conversación en esta red social involucra actores políticos, institucionales y algunos ciudadanos, generalmente influencers que tienden a participar activamente en este tipo de plataformas.

Es importante aclarar que si se mostraran todas las interacciones de manera completa, el grafo resultaría completamente ilegible. Por esta razón, se presentó de forma simplificada para facilitar la lectura de este informe. Sin embargo, sí se analizó la estructura completa para caracterizar la topología de la red de interacciones y fundamentar las conclusiones sobre su conectividad.

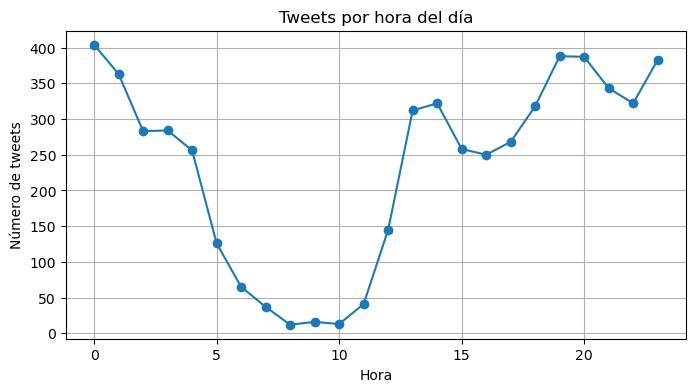
Los resultados muestran que la red tiene muy poca densidad, con un valor de 0.0010, lo cual significa que existen pocas conexiones en comparación con todas las que teóricamente podrían establecerse. El diámetro del componente principal es de 7, lo que indica que a pesar de esta baja densidad, la distancia máxima entre cualquier par de usuarios dentro de la red principal es relativamente corta, facilitando así la circulación de información.

Por otra parte, se aplicó el mismo análisis a los hashtags, pues a partir de ver las cuentas con más interacción nos preguntamos ¿Qué tema revelan los hashtags? Se halló a partir de este análisis que hay una fuerte carga política. Al observar tanto las personas que participaban como los hashtags que utilizaban, queda claro que este ya no es un problema menor: se trata de un tema que preocupa a muchos ciudadanos y que los políticos deben atender para apoyar al pueblo, pues según los análisis realizados, definitivamente es un asunto con alta carga política.

Adicionalmente, se creó una nube de palabras usando los textos limpios, lo que reforzó esta observación. Las palabras más frecuentes incluyen "sí", "solo", "corrupto", "gobierno", etc., que representan claramente un reclamo dirigido a las instituciones gubernamentales y al gobierno mismo. Reforzando la idea de un fuerte discurso que denuncia la corrupción dentro del país.



Finalmente, se preguntó ¿En qué horarios se concentra la actividad en Twitter sobre el tráfico?. Considerando eso se analizó la distribución horaria de la actividad. Se observa una caída marcada durante las horas de madrugada, pero aparecen picos de actividad al inicio del día (cerca de medianoche) y durante la tarde-noche.



Esto tiene sentido, ya que muchas personas suelen usar su celular antes de dormir y aprovechan para hacer estas publicaciones, o bien durante las horas de la tarde, que son momentos oportunos para expresar su descontento con las soluciones gubernamentales porque están experimentando los problemas de tráfico en tiempo real. De hecho, durante este período de tarde-noche hay bastante actividad, alcanzando hasta 350 publicaciones por hora.

Adicionalmente, regresando al grafo. Luego de pasar ciertos análisis se reveló un coeficiente de agrupamiento de 0.2301 que representa un nivel moderado de cohesión local. Esto significa que los usuarios tienden a formar comunidades o grupos relativamente bien conectados entre sí por lo que se forman pequeños grupos o pequeños núcleos de interacción más intenta dentro de la red general.

Los resultados muestran que después de filtrar la red por interacciones fuertes (con un peso mínimo de 3) y aplicar una poda con k-core de nivel 2, el grafo se redujo a 103 nodos y 214 enlaces. Esta simplificación permitió identificar las comunidades de manera más clara y precisa.

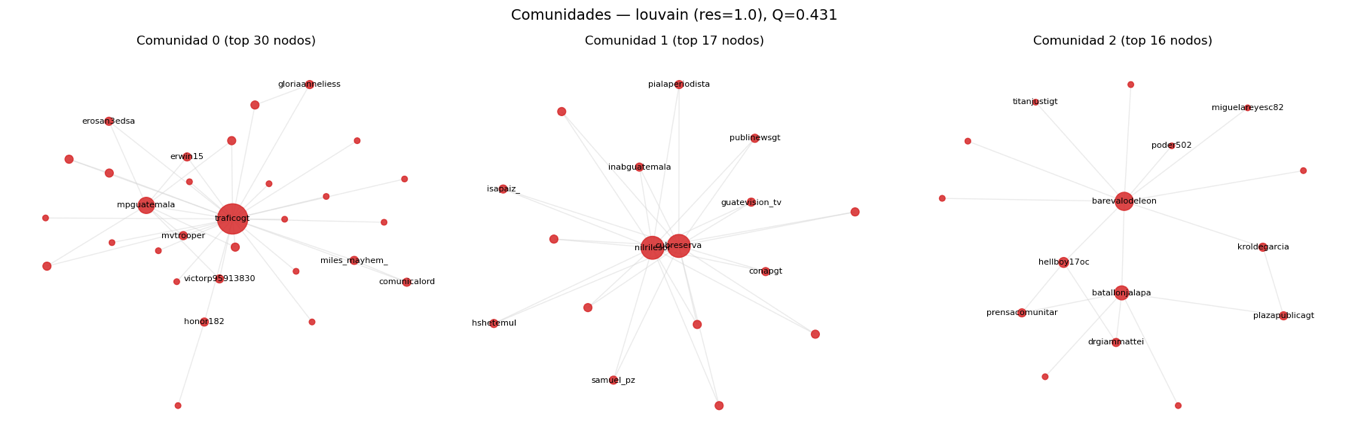
Al comparar diferentes algoritmos de detección de comunidades, el método de Louvain con resolución 1.0 fue el que obtuvo la modularidad más alta (Q≈0.431), lo que indica una división adecuada entre grupos que son densos internamente pero tienen conexiones más débiles entre ellos. Esta comparación confirmó que Louvain supera en calidad de partición a otras alternativas como greedy, label propagation o girvan-newman para este conjunto de datos.

Con esta configuración se identificaron 7 comunidades principales. La más grande cuenta con 30 nodos, seguida por dos comunidades de 17 y 16 nodos respectivamente, mientras que las comunidades menores agrupan entre 2 y 11 usuarios cada una.

Las visualizaciones revelan que cada comunidad se organiza alrededor de un nodo central que funciona como polo de interacción:

* En la comunidad 0 destaca @traficogt, vinculado principalmente a cuentas institucionales y de monitoreo
* En la comunidad 1 sobresalen perfiles como @nilrilesreserva, que conecta principalmente medios y periodistas
* En la comunidad 2 se concentra la actividad alrededor de @barevalodeleon, junto con figuras políticas y organizaciones

La modularidad obtenida indica que la red presenta una fragmentación temática clara, donde los usuarios se agrupan alrededor de referentes específicos. Esto confirma la existencia de núcleos de opinión diferenciados que estructuran y organizan la conversación digital sobre el tema del tráfico.



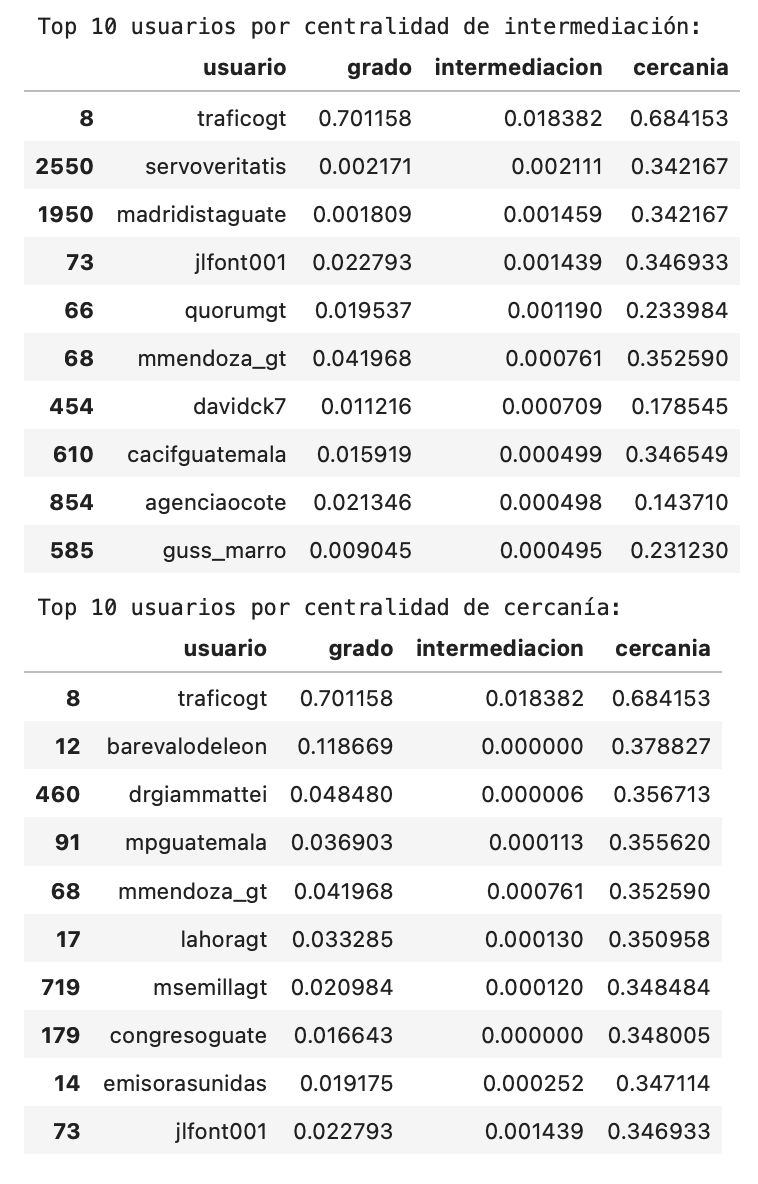
El análisis de centralidad confirma que @traficogt es el actor dominante en la red, liderando las tres métricas principales. Su centralidad de grado extremadamente alta refleja la enorme cantidad de conexiones directas que mantiene con otros usuarios. Además, es el nodo más cercano en términos de distancia promedio y ocupa una posición estratégica como intermediario en la red.

Detrás de este actor principal aparecen otros usuarios con diferentes tipos de relevancia según la métrica analizada:

En centralidad de grado y cercanía destacan @barevalodeleon, @drgiammattei, @mpguatemala, @mmendoza\_gt, @lahoragt y @amilcarmontejo. Todos estos perfiles están vinculados a la política, instituciones o medios de comunicación, lo que evidencia que no solo concentran atención sino que también tienen una gran capacidad para difundir información en la red.

En centralidad de intermediación surgen perfiles menos visibles en términos de conexiones directas, pero que cumplen un papel fundamental como puentes. Usuarios como @servoveritatis, @madridistaguate o @quorumgt, aunque no son los más mencionados, logran conectar diferentes comunidades y facilitar el flujo de mensajes entre grupos distintos.

En conjunto, estos resultados permiten caracterizar a los influencers desde tres dimensiones complementarias: los grandes hubs que concentran la conversación, los actores centrales en la difusión de información, y los nodos que, aunque con menor grado de conexión, cumplen funciones cruciales de enlace entre comunidades que de otra manera estarían fragmentadas.



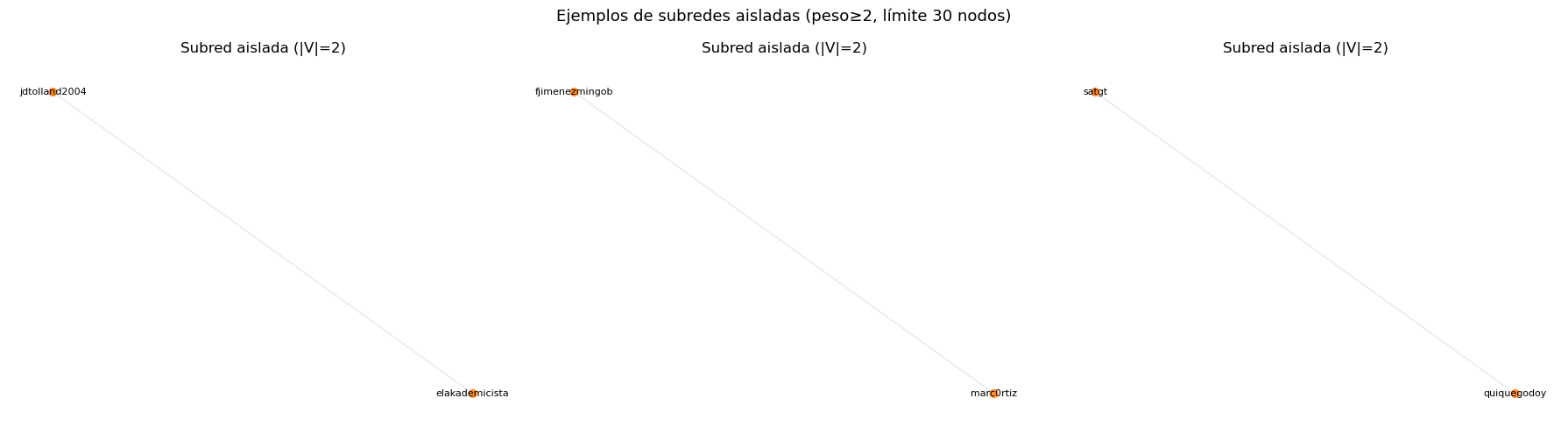
Ahora, para la detección y análisis de grupos aislados se vio nuevamente que en la red de interacciones en Twitter presenta una fuerte centralización alrededor de la cuenta @traficogt, que domina todas las métricas de centralidad calculadas.

Esta cuenta se posiciona como el nodo con mayor grado (acumula la mayor cantidad de vínculos directos), con mayor cercanía (se encuentra accesible en pocos pasos desde casi cualquier otro usuario) y con mayor intermediación (actúa como punto de paso crucial en el flujo de información entre distintas comunidades).

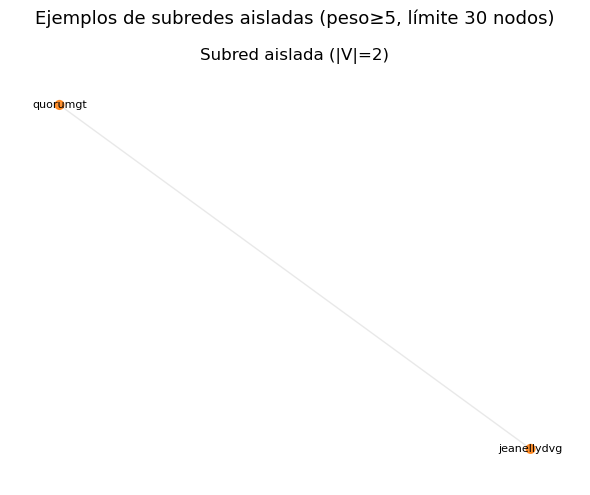
Alrededor de este liderazgo emergen otros actores relevantes que cumplen diferentes papeles. Por un lado, cuentas como @barevalodeleon, @drgiammattei, @mpguatemala, @mmendoza\_gt, @lahoragt y @amilcarmontejo muestran altos valores de grado y cercanía, posicionándose como amplificadores capaces de difundir mensajes de manera eficiente. Por otro lado, usuarios como @servoveritatis @madridistaguate o @quorumgt destacan en intermediación, actuando como puentes que conectan comunidades y permiten que los mensajes de ciertos grupos alcancen audiencias más amplias.

En conjunto, este patrón refleja un ecosistema donde un actor central organiza la conversación y distintos nodos clave se reparten entre roles de amplificación y conexión entre comunidades.

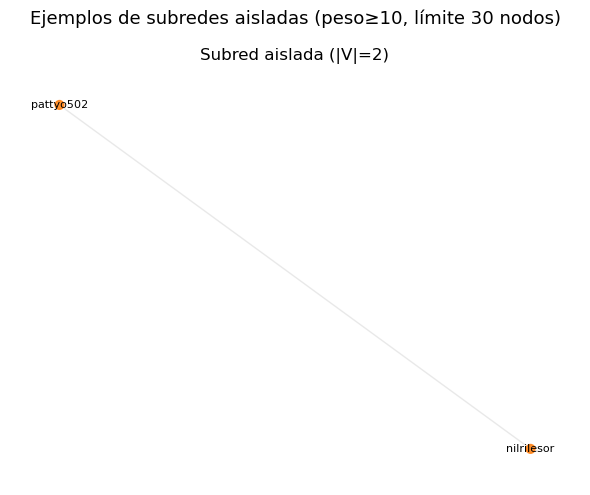
El análisis de subredes aisladas confirma esta estructura jerárquica. Para comprender mejor la organización de la red, se analizaron diferentes umbrales de peso de las conexiones:



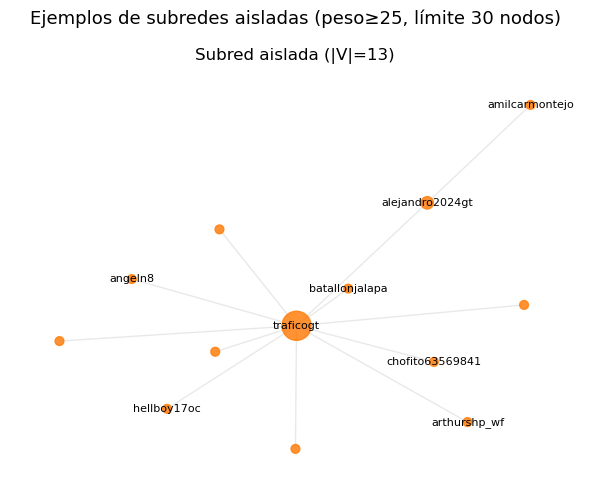
Al establecer un umbral bajo de peso (**W = 2**), la red mantiene un tamaño considerable con un gran componente central, pero aparecen múltiples microgrupos y díadas desconectadas del núcleo. Esto indica interacciones puntuales o relaciones efímeras entre usuarios.



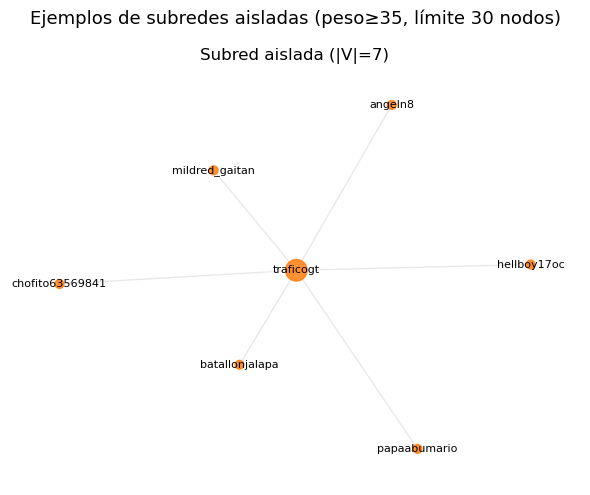
Cuando el umbral se eleva a **W = 5**, la red comienza a contraerse, reduciéndose a poco menos de doscientos nodos con vínculos más frecuentes. Esto evidencia un filtrado de interacciones esporádicas y una concentración en relaciones más recurrentes.



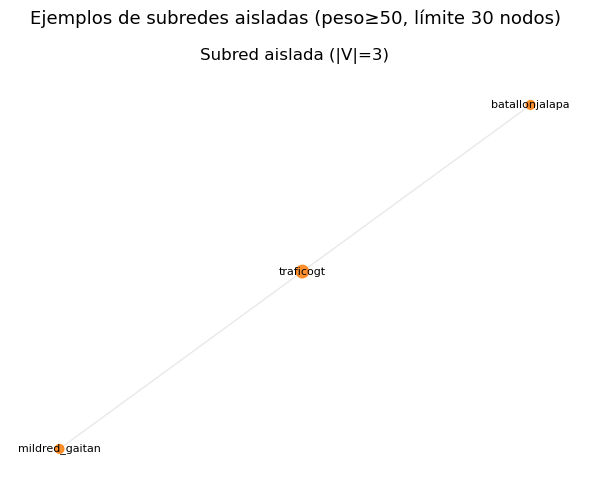
Con **W = 10**, la estructura se simplifica aún más. Emerge con claridad un componente principal más reducido y denso que concentra la mayoría de las interacciones significativas, mientras que las pequeñas subredes periféricas casi desaparecen.



A partir de **W = 25**, la red se transforma en un núcleo compacto compuesto por un grupo muy pequeño de cuentas que interactúan reiteradamente entre sí. El grafo muestra que la periferia se ha perdido casi por completo y solo persiste un núcleo donde las interacciones son especialmente frecuentes.



Con **W = 35**, se observa muy claramente una estructura estelar alrededor de **@traficogt**, en la cual varios usuarios mantienen vínculos muy reiterados con esta cuenta, reforzando la centralidad de su papel en el sistema.



Finalmente, con **W = 50**, la red queda reducida a un núcleo mínimo de usuarios donde la interacción se sostiene únicamente en lazos muy fuertes y repetidos. En esta etapa prácticamente desaparece toda la complejidad periférica, quedando expuesta la base más fiel y reiterativa de la conversación.

La comparación de umbrales revela un patrón consistente: la conversación en Twitter es amplia y diversa cuando se consideran interacciones de baja frecuencia, pero al exigir mayor repetición de vínculos, la red se contrae hacia un núcleo muy reducido, cohesionado y fuertemente organizado en torno a @traficogt.

Este hallazgo refuerza la idea de que, aunque muchos actores participan ocasionalmente, el peso de la discusión recae en unos pocos usuarios centrales que concentran tanto las menciones como los vínculos más persistentes, definiendo así la dinámica estructural de la red.

Por otra parte se aplicaron dos enfoques distintos de análisis de sentimiento utilizando un léxico en español, lo que permitió comparar cómo cambia la clasificación de los tweets según la metodología empleada.

En el primer método se calculó un puntaje basado en la diferencia entre palabras positivas y negativas, dividido por la longitud del texto. Se establecieron umbrales específicos para clasificar los tweets en positivo, negativo o neutral.

Los resultados mostraron que:

* **75.3%** de los mensajes fueron clasificados como neutrales
* **20.2%** como negativos
* **4.6%** como positivos

Esta distribución indica que el discurso predominante es informativo o descriptivo, sin cargar muchas palabras con polaridad. Sin embargo, existe un volumen considerable de expresiones críticas en comparación con las pocas manifestaciones positivas encontradas.

En el segundo método se adoptó una regla más estricta y directa: cualquier tweet que contenga al menos una palabra negativa se clasifica como negativo; si no hay palabras negativas pero sí positivas, se clasifica como positivo; y en ausencia de ambas, se considera neutral.

Esta lógica, que prioriza la detección de expresiones negativas, produjo los siguientes resultados:

* **70.5%** de tweets neutrales (reducción del 4.8%)
* **24.4%** de tweets negativos (aumento del 4.2%)
* **5.1%** de tweets positivos (aumento del 0.5%)

La comparación de ambos métodos confirma que el vocabulario de crítica y denuncia está más presente de lo que reflejaba el primer análisis. Este hallazgo refuerza la idea de que las conversaciones en torno a temas como el tráfico y la política tienden a expresarse con una carga significativa de descontento ciudadano.

El segundo enfoque, al ser más sensible a la presencia de palabras negativas, logra capturar mejor la naturaleza crítica del discurso, revelando que aproximadamente una cuarta parte de los tweets contienen expresiones de insatisfacción o crítica hacia las instituciones y situaciones mencionadas.

|  |  |
| --- | --- |
|  |  |

La evidencia obtenida a lo largo del análisis muestra con claridad cómo la red de Twitter en torno a temas de política y tráfico en Guatemala está profundamente estructurada por la acción de influencers y comunidades organizadas.

Las cuentas más influyentes, como **@traficogt**, diversos medios de comunicación y actores políticos, concentran la mayoría de las interacciones y funcionan como nodos de referencia para toda la discusión. Sus altos valores de centralidad de grado confirman que son los más mencionados y retuiteados, mientras que su centralidad de intermediación evidencia que algunos cumplen un papel crucial como puentes, conectando distintas comunidades y permitiendo que la información circule más allá de nichos locales específicos.

En este sentido, los influencers no solo difunden contenidos, sino que también organizan la arquitectura misma de la conversación digital, definiendo qué temas se discuten y cómo se abordan.

La detección de comunidades mediante el algoritmo Louvain mostró que la red se fragmenta en grupos temáticos bien definidos. Algunos se organizan alrededor de tópicos políticos y expresan críticas al gobierno, al congreso y a las instituciones, mientras que otros giran en torno a problemas de movilidad como bloqueos, accidentes o emergencias viales.

La aparición de subredes aisladas, especialmente visible al aumentar los umbrales de peso, confirma que existen pequeños grupos de usuarios que interactúan intensamente entre sí pero muy poco con el resto, formando auténticas burbujas de información. Este patrón refleja una conversación que no es homogénea, sino polarizada y distribuida en espacios de afinidad temática y relacional.

El análisis de contenido y sentimiento complementa esta visión al revelar un predominio de mensajes neutrales de carácter descriptivo, acompañados por una proporción significativa de publicaciones negativas asociadas a denuncias de corrupción, críticas al gobierno o reportes de caos vial. Los mensajes positivos, en cambio, son marginales y suelen limitarse a agradecimientos o casos puntuales de experiencias satisfactorias.

La nube de palabras, con la prominencia de términos como "corrupto", "bloqueo", "caos" y "congreso", refuerza la idea de que la conversación se articula principalmente en torno a la crítica y la queja ciudadana.

En conjunto, estos hallazgos permiten afirmar que Twitter funciona en este contexto como un espacio de vigilancia social y denuncia. Los influencers definen la agenda y marcan los ejes de discusión, mientras que las comunidades refuerzan y amplifican los mensajes alineados con sus intereses o posturas.

La interacción de ambos elementos configura la opinión pública digital: los líderes concentran la atención, las comunidades reproducen y amplifican los discursos, y el contenido negativo se convierte en la base emocional que articula gran parte de la conversación. Esta dinámica convierte a la plataforma en un instrumento fundamental para canalizar el descontento ciudadano y ejercer presión sobre las instituciones públicas.