

Analítica de Redes Sociales

Data Science (CC3084) - 2026

Analítica de Redes Sociales

Semestre 02, 2026

Definición

La analítica de redes sociales es el conjunto de técnicas que estudian las relaciones entre actores sociales representándolas como un grafo, para medir su estructura y explicar el comportamiento que esa estructura produce.

La unidad de análisis no es el individuo ni el documento: es el vínculo. Un conjunto de personas descrito una por una es una tabla; ese mismo conjunto descrito por quién habla con quién es una red, y las preguntas que se pueden responder cambian por completo.

Dos cosas distintas con el mismo nombre

En la práctica el término se usa para dos actividades que conviene separar desde el inicio.

Análisis de redes sociales (SNA)	estructura relacional modelada como grafo nodos, aristas, centralidad, comunidades origen: sociología, años 30
Analítica de social media	desempeño de contenido en plataformas

alcance, impresiones, engagement, CTR
origen: marketing digital, años 2000

Las dos son legítimas y se combinan, pero responden preguntas diferentes. La primera pregunta cómo está organizado un colectivo; la segunda, cómo se comportó una publicación. Esta lección trata principalmente la primera y cierra con la segunda.

Relación con la lección anterior

La minería de texto analiza el contenido: qué se dice. La analítica de redes analiza la estructura: quién se conecta con quién.

Minería de texto	"este mensaje es negativo"	-> propiedad del contenido
Analítica de redes	"este mensaje llegó a 40,000 cuentas en 3 saltos"	-> propiedad de la estructura

Ninguna de las dos reemplaza a la otra. El caso interesante casi siempre requiere ambas: un texto tóxico importa poco si nadie lo replica, y una cascada masiva no se interpreta sin leer qué se está replicando. Cuando en esta lección aparezca el contenido de los mensajes, se asume lo ya visto (tokenización, representación vectorial, sentimiento) y no se repite.

Origen del Campo

La idea de dibujar relaciones para estudiarlas es anterior a las computadoras.

- 1934 — Jacob Moreno publica *Who Shall Survive?* e introduce el sociograma: diagramas de puntos y líneas para representar afinidades dentro de grupos escolares.

- 1967 — Stanley Milgram publica el experimento del mundo pequeño: cartas que debían llegar a un destinatario en Boston pasando solo entre conocidos.
- 1973 — Mark Granovetter publica *The Strength of Weak Ties*: los contactos lejanos aportan más información nueva que los cercanos.
- 1977 — Wayne Zachary documenta un club de karate de 34 miembros y su ruptura; el conjunto de datos sigue siendo el ejemplo estándar de detección de comunidades.
- 1998 — Watts y Strogatz formalizan las redes de mundo pequeño: alto agrupamiento y caminos cortos a la vez.
- 1999 — Barabási y Albert describen las redes libres de escala: la distribución de grados sigue una ley de potencias.
- 2004 en adelante — las plataformas digitales convierten el registro de relaciones en un subproducto automático, y el campo pasa de decenas de nodos a miles de millones.

El cambio de escala es lo que vuelve el tema parte de la ciencia de datos: los métodos son viejos, los datos son nuevos.

El Grafo como Modelo

Un grafo es un par $G = (V, E)$ donde V es el conjunto de nodos y $E \subseteq V \times V$ el conjunto de aristas.

- Nodo (vértice): el actor. Una persona, una cuenta, una organización, un hashtag, un dispositivo.
- Arista (enlace): la relación. Amistad, seguimiento, retuit, mención, correo, coautoría, transacción.

Modelar es decidir estas dos cosas, y la decisión determina todo lo que se puede medir después.

Variantes que importan

No dirigido	la relación es simétrica	amistad en Facebook, coautoría
Dirigido	la relación tiene sentido	seguir en X/Instagram, citar, retuitear
Ponderado	la arista lleva un número	cantidad de mensajes, duración total
Bipartito	dos tipos de nodo	usuarios y grupos, autores y papers
Temporal	la arista existe en un instante	quién llamó a quién y cuándo
Multicapa	varias relaciones a la vez	misma gente en WhatsApp y en X

Un error frecuente es tratar como no dirigida una relación que no lo es. En X, que A siga a B no implica que B siga a A, y colapsar eso a una amistad simétrica inventa reciprocidad que no existe en los datos.

Elegir qué es una arista

Con los mismos datos de una plataforma se pueden construir redes distintas, y cada una responde otra pregunta.

Red de seguimiento	A sigue a B	audiencia potencial, estructura estable
Red de retuits	A retuitea a B	difusión efectiva, qué se propagó
Red de menciones	A menciona a B	conversación, confrontación incluida
Red de respuestas	A responde a B	diálogo real, hilos
Red de co-hashtag	A y B usan #x	temas compartidos sin contacto directo
Red de co-audiencia	A y B comparten seguidores	cercanía ideológica o temática

Las redes de seguimiento son estables y grandes; las de interacción son dispersas y cambian por día. Un influencer con muchos seguidores puede tener una red de retuits casi vacía, y ese contraste ya es un hallazgo.

Representación en memoria

Matriz de adyacencia	$A[i][j] = 1$ si hay arista consulta de arista $O(1)$	$O(n^2)$ memoria buena para redes pequeñas y densas
Lista de adyacencia	cada nodo guarda sus vecinos recorrer vecinos es directo	$O(n + m)$ memoria la opción por defecto
Lista de aristas	pares (origen, destino) formato de intercambio	$O(m)$ memoria CSV, streaming

Las redes sociales reales son extremadamente dispersas: un usuario típico de una red con cientos de millones de cuentas conoce a unos cientos. Una matriz de adyacencia densa para un millón de nodos pediría del orden de 10^{12} celdas, así que en la práctica se trabaja con listas de adyacencia o matrices dispersas.

Métricas de Nodo

Las medidas de centralidad responden a la pregunta de quién es importante en la red. No hay una sola respuesta porque no hay una sola forma de ser importante.

Grado

El número de vecinos. En redes dirigidas se separa en grado de entrada y de salida.

$$k_i = \sum_j A_{ij}$$

- Grado de entrada: cuánta atención recibe. Seguidores, menciones recibidas, citas.
- Grado de salida: cuánta actividad emite. A quién sigue, a quién menciona.

Es la medida más barata y la más fácil de inflar artificialmente. Comprar seguidores sube el grado de entrada sin cambiar nada más de la estructura, así que mide popularidad y no influencia: una cuenta con muchísimos seguidores puede no generar ninguna difusión.

Intermediación (betweenness)

Fracción de caminos más cortos entre pares de nodos que pasan por el nodo.

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

donde σ_{st} es el número de caminos más cortos entre s y t , y $\sigma_{st}(v)$ los que pasan por v .

Identifica puentes: nodos que conectan regiones que de otro modo estarían separadas. Un nodo puede tener grado bajo y intermediación altísima si es el único vínculo entre dos comunidades. Es la medida que detecta a quien controla el flujo de información, no a quien tiene más público.

Su costo es alto: el algoritmo de Brandes la calcula en $O(nm)$ para grafos no ponderados, lo que en redes grandes obliga a estimarla por muestreo de nodos fuente.

Cercanía (closeness)

Inverso de la distancia media a todos los demás nodos.

$$C_C(v) = \frac{n-1}{\sum_{u \neq v} d(v, u)}$$

Mide qué tan rápido puede alcanzar el nodo al resto de la red. Alta cercanía significa buena posición para difundir. Solo tiene sentido dentro de una componente conexa: si el grafo está partido, la distancia a los nodos inalcanzables es infinita y hay que calcularla por componente.

Centralidad de vector propio y PageRank

La idea es recursiva: un nodo es importante si está conectado a nodos importantes.

$$x_i = \frac{1}{\lambda} \sum_j A_{ij} x_j$$

PageRank es la variante con amortiguación que Google usó para ordenar páginas web, y funciona igual sobre grafos sociales:

$$PR(i) = \frac{1 - d}{n} + d \sum_{j \in M(i)} \frac{PR(j)}{L(j)}$$

donde $d \approx 0.85$ es el factor de amortiguación, $M(i)$ los nodos que apuntan a i y $L(j)$ el grado de salida de j . Distingue a quien recibe mil enlaces de cuentas irrelevantes de quien recibe diez de cuentas centrales.

Cuál usar

Grado	popularidad inmediata, alcance directo
Intermediación	control del flujo, puentes entre grupos
Cercanía	velocidad de difusión desde ese nodo
Vector propio	prestigio: importancia por asociación
PageRank	prestigio robusto en redes dirigidas

No coinciden, y esa discrepancia es informativa. Elegir la métrica antes de tener clara la pregunta es la forma más común de producir un ranking que no significa nada.

Métricas de Red

Además de describir nodos, se describe la red completa.

- Densidad: proporción de aristas existentes sobre las posibles. En redes sociales grandes es minúscula.

$$\rho = \frac{2m}{n(n-1)}$$

- Grado medio: $\langle k \rangle = 2m / n$. Cuántos vecinos tiene un nodo típico.
- Distribución de grados: cómo se reparte la popularidad. Casi nunca es normal.
- Coeficiente de agrupamiento: probabilidad de que dos vecinos de un nodo sean vecinos entre sí.
- Longitud media de camino y diámetro: qué tan lejos están los nodos entre sí.
- Componentes conexas: cuántos pedazos separados tiene la red; casi siempre hay una componente gigante y muchos fragmentos.
- Reciprocidad: en redes dirigidas, qué fracción de aristas tiene su contraparte.
- Asortatividad: si los nodos de grado alto se conectan entre sí o con nodos de grado bajo.

La distribución de grados no es normal

Distribución normal (intuición equivocada)

la mayoría cerca del promedio, extremos raros
ejemplo: estatura de una población

Ley de potencias (lo que se observa)

$$P(k) \sim k^{-\gamma}$$

la mayoría con muy pocas conexiones
unos pocos con órdenes de magnitud más

Barabási y Albert propusieron el mecanismo de acoplamiento preferencial: los nodos nuevos se conectan con preferencia a los que ya tienen muchas conexiones, y el resultado es una red libre de escala.

La consecuencia metodológica es fuerte: el promedio deja de ser un resumen útil. Decir que una cuenta tiene un número de seguidores igual al promedio no describe a nadie, porque la media está dominada por la cola. Con estas distribuciones se reportan medianas, cuantiles y la forma de la cola.

Fuente: Barabási y Albert, *Emergence of Scaling in Random Networks*, Science 286, 1999. <https://www.science.org/doi/10.1126/science.286.5439.509>

Mundo pequeño

Cuando cada quien se relaciona solo con los que tiene cerca, llegar al otro extremo de la red cuesta muchos saltos. Unos pocos vínculos que crucen de un grupo a otro derrumban esa distancia.

Conviene verlo con un anillo de doce nodos donde cada uno solo conoce a sus dos vecinos. Ir del nodo de arriba al de abajo obliga a recorrer medio anillo: seis saltos. Basta con agregar una sola arista que cruce de un lado al otro para que el mismo recorrido baje a tres saltos.

Anillo sin atajos	A -> 1 -> 2 -> 3 -> 4 -> 5 -> B	6 saltos
Anillo con un atajo	A -> 1 ----- atajo -----> 7 -> B	3 saltos

Lo importante es que nadie perdió a sus vecinos: el atajo se suma, no reemplaza nada. Por eso la red sigue igual de apiñada en grupos y aun así todos quedan cerca.

Eso es lo contraintuitivo que formalizaron Watts y Strogatz. La expectativa sería tener que elegir entre una red apiñada en grupos (con caminos largos) y una red revuelta al azar (con caminos cortos y sin grupos). Con unos pocos atajos se obtienen las dos propiedades a la vez: agrupamiento alto y caminos cortos. Es la razón de que el mundo se sienta pequeño aunque casi todos nuestros vínculos sean locales.

Milgram lo había medido antes con cartas físicas: pidió a personas en el medio oeste de Estados Unidos hacer llegar un sobre a un destinatario en Boston, entregándolo solo a conocidos directos. Las cadenas que se completaron tuvieron una mediana de alrededor de cinco intermediarios, de donde salió la expresión seis grados de separación.

En 2012, Backstrom, Boldi, Rosa, Ugander y Vigna repitieron la medición sobre el grafo completo de usuarios activos de Facebook: unos 721 millones de usuarios y 69 mil millones de aristas de amistad. La distancia media resultó 4.74, es decir 3.74 intermediarios.

Milgram 1967	cartas físicas, cadenas completadas	~5 intermediarios
Facebook 2012	721 M usuarios, 69 mil M aristas	3.74 intermediarios

El resultado no dice que cualquier par de personas del mundo esté a cuatro pasos: dice que dentro de esa plataforma, entre cuentas activas, la distancia media es esa. Confundir la red medida con la población general es un error de interpretación, no de cálculo.

Fuentes: Watts y Strogatz, *Collective Dynamics of Small-World Networks*, Nature 393, 1998. <https://www.nature.com/articles/30918> Backstrom et al., *Four Degrees of Separation*, ACM Web Science 2012. <https://arxiv.org/abs/1111.4570>

Homofilia, Tríadas y Lazos Débiles

Homofilia

Los nodos similares tienden a conectarse. Es de las regularidades más robustas en redes sociales, y trae un problema causal: si dos amigos se parecen, puede ser porque se influyeron o porque se hicieron amigos por parecerse. Los datos observacionales rara vez distinguen ambas cosas, y esa ambigüedad es la crítica

estándar a los estudios de contagio social.

Cierre triádico

Si A conoce a B y B conoce a C, es probable que A y C terminen conectados. De aquí sale el coeficiente de agrupamiento alto de las redes reales, y también la lógica de las recomendaciones de amistad: sugerir a quienes comparten muchos vecinos.

Lazos fuertes y débiles

Granovetter argumentó que los lazos débiles (conocidos, no amigos cercanos) son los que traen información nueva, porque los lazos fuertes se mueven en el mismo círculo cerrado y comparten la misma información. La evidencia original era observacional: personas que consiguieron trabajo por contactos con quienes tenían poco trato.

Un lazo débil es, en términos de grafo, una arista puente: aquella que al quitarla alarga mucho la distancia entre dos regiones. Es la misma idea que mide la intermediación, vista desde la arista en lugar del nodo.

Fuente: Granovetter, *The Strength of Weak Ties*, American Journal of Sociology 78, 1973. <https://www.jstor.org/stable/2776392>

Detección de Comunidades

Una comunidad es un conjunto de nodos con muchas conexiones entre sí y pocas hacia afuera. Detectarlas es agrupamiento, pero sobre estructura en lugar de sobre vectores de atributos.

Modularidad

La medida estándar de calidad de una partición compara las aristas dentro de los grupos contra las que habría en una red aleatoria con la misma distribución de grados.

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \Delta(c_i, c_j)$$

Valores típicos útiles van de 0.3 a 0.7. Un valor alto no garantiza que las comunidades signifiquen algo: la modularidad tiene un límite de resolución que le impide ver comunidades pequeñas dentro de redes grandes, y encuentra particiones con modularidad decente incluso en grafos aleatorios.

Algoritmos

Girvan-Newman	quita iterativamente las aristas de mayor intermediación intuitivo, jerárquico, costoso: no escala
Louvain	optimiza modularidad de forma voraz y jerárquica muy rápido, el más usado en la práctica
Leiden	corrige un defecto de Louvain: comunidades mal conectadas recomendado cuando la partición se va a interpretar
Label propagation	cada nodo toma la etiqueta mayoritaria de sus vecinos casi lineal, resultado inestable entre corridas
Etiquetas	no hay verdad de referencia: la validación es cualitativa

Louvain fue probado por sus autores sobre una red de telefonía móvil belga de 2 millones de clientes, donde recuperó la división lingüística del país, y sobre un grafo web de 118 millones de nodos.

Ejemplo documentado: el club de karate de Zachary

Zachary observó durante tres años un club universitario de karate de 34 miembros, registrando qué pares interactuaban fuera de las actividades del club. Durante el estudio surgió un conflicto entre el instructor y el administrador que terminó partiendo el club en dos.

Usando solo la estructura de la red y un modelo de flujo de información, Zachary predijo correctamente a qué facción se fue cada miembro, con una sola excepción: el miembro 9, que se fue con el instructor a pesar de que el modelo lo asignaba al otro grupo. La explicación estaba fuera del grafo: le faltaban semanas para obtener el cinturón negro con ese instructor.

El caso enseña las dos cosas al mismo tiempo. La estructura predice el comportamiento con precisión notable, y siempre queda un residuo que solo se explica con contexto que la red no contiene.

Fuente: Zachary, *An Information Flow Model for Conflict and Fission in Small Groups*, Journal of Anthropological Research 33, 1977.
<https://www.journals.uchicago.edu/doi/10.1086/jar.33.4.3629752>

Difusión y Contagio

La estructura importa porque las cosas se mueven sobre ella: información, comportamientos, enfermedades, rumores.

Modelos básicos

Cascada independiente	cada nodo activo intenta activar a cada vecino una sola vez, con probabilidad p
Umbral lineal	un nodo se activa cuando la fracción de vecinos activos supera su umbral θ

SIR

susceptible -> infectado -> recuperado
prestado de la epidemiología

La diferencia práctica: en cascada independiente basta un contacto (contagio simple, típico de información), mientras que en umbral lineal se necesita refuerzo de varios contactos (contagio complejo, más típico de adopción de comportamientos).

Maximización de influencia

El problema de elegir k nodos iniciales que maximicen el alcance final es NP-difícil, pero la función de propagación es submodular, lo que garantiza que un algoritmo voraz alcance al menos el 63% del óptimo. Es la base formal del marketing de influencers, y explica por qué la selección óptima rara vez son los k nodos de mayor grado: esos suelen tener audiencias solapadas.

Ejemplo documentado: la propagación de noticias falsas

Vosoughi, Roy y Aral analizaron en Science todas las cascadas verificadas de noticias verdaderas y falsas en Twitter entre 2006 y 2017: unas 126,000 historias difundidas por cerca de 3 millones de personas más de 4.5 millones de veces. La verificación vino de seis organizaciones independientes de fact-checking, con acuerdo entre ellas del 95 al 98%.

Las noticias falsas se difundieron más lejos, más rápido, más profundo y más ampliamente que las verdaderas en todas las categorías, y el efecto fue más marcado en política. El 1% de cascadas falsas más grandes alcanzó entre mil y cien mil personas, mientras que las verdaderas rara vez pasaron de mil.

El hallazgo que más se cita es el que descarta la explicación fácil: los bots aceleraban por igual lo verdadero y lo falso, así que la diferencia la produjeron personas. La hipótesis de los autores es la novedad: lo falso era más novedoso y

la gente comparte más lo que sorprende.

Este análisis es puramente estructural. Las métricas usadas fueron profundidad de cascada, tamaño, amplitud máxima y tiempo hasta alcanzar cada nivel, no propiedades del texto.

Fuente: Vosoughi, Roy y Aral, *The Spread of True and False News Online*, Science 359, 2018. <https://www.science.org/doi/10.1126/science.aap9559>

Detección de Coordinación y Cuentas Automatizadas

Las redes también sirven para detectar comportamiento inauténtico, y aquí la estructura delata cosas que el contenido esconde.

Señales estructurales y temporales

- Co-retuit: cuentas que retuitean repetidamente el mismo contenido en ventanas de tiempo muy cortas forman conglomerados anormalmente densos.
- Sincronía temporal: publicar en el mismo minuto de forma sistemática no ocurre por azar.
- Contenido idéntico: reutilización literal del mismo texto entre cuentas sin relación aparente.
- Cuentas puente: perfiles cuya única función es conectar el núcleo coordinado con audiencias reales.
- Ciclo de vida: creación masiva de cuentas en la misma fecha, o reactivación simultánea de cuentas dormidas.

La técnica general es construir una red de similitud entre cuentas (co-retuit, co-hashtag o co-timing), y luego aplicar detección de comunidades sobre ella. Los grupos coordinados aparecen como componentes densos y separados del resto.

Ejemplo documentado: netcenters en Guatemala

En 2019 la CICIG publicó el informe *Bots, netcenters y combate a la impunidad: el caso de Guatemala*, que documentó redes de cuentas dedicadas a atacar a operadores de justicia y periodistas. El informe definió netcenter como un conjunto de cuentas conectadas de forma directa o indirecta, que puede mezclar perfiles falsos, cuentas de personas reales y cuentas institucionales, y usó geolocalización por IP para ubicar concentraciones de actividad en Ciudad de Guatemala, Quetzaltenango y Cobán.

En 2022, Agencia Ocote junto al Centro Latinoamericano de Investigación Periodística documentó cómo esas estructuras mutaron. El análisis siguió 42 cuentas de X durante doce meses, entre agosto de 2021 y agosto de 2022, que produjeron unos 14,000 tuits, 121,000 retuits, 52,000 respuestas y 28,000 citas.

El patrón encontrado es estructural y jerárquico: cuentas anónimas publicaban el ataque inicial, figuras públicas identificables lo amplificaban, y decenas de cuentas de personas reales lo expandían. Los picos de actividad coincidieron con acciones judiciales concretas, lo que descarta actividad orgánica.

El caso vale como ejemplo local de que la analítica de redes no es solo un tema de plataformas extranjeras, y de que las mismas herramientas sirven para el periodismo de investigación.

Fuentes: CICIG, *Bots, netcenters y combate a la impunidad*, 2019.

[https://www.cicig.org/wp-](https://www.cicig.org/wp-content/uploads/2018/05/Informe_bots_y_netcenters_2019.pdf)

[content/uploads/2018/05/Informe_bots_y_netcenters_2019.pdf](https://www.cicig.org/wp-content/uploads/2018/05/Informe_bots_y_netcenters_2019.pdf) Agencia Ocote y

CLIP, *Cómo mutaron los netcenters anti justicia y derechos humanos en Guatemala*, 2022. <https://www.agenciaocote.com/blog/2022/11/30/mutaron-netcenters-anti-justicia-yderechos-humanos-en-guatemala/>

Obtención de Datos

Esta es la etapa que más proyectos mata, y su dificultad cambió radicalmente en los últimos años.

Vías de acceso

API oficial	estable y legal, con límites de tasa y costo
Scraping	frágil, suele violar términos de servicio
Datos donados	los usuarios exportan y ceden sus propios datos
Conjuntos públicos	SNAP, Netzscheider, Kaggle, repositorios de papers
Acceso regulado	DSA art. 40 en la Unión Europea, para investigadores acreditados

El cierre del acceso

Dos hechos concretos cambiaron el panorama y conviene conocerlos antes de planificar un proyecto.

- En febrero de 2023, X eliminó el acceso gratuito a su API, incluido el track académico que durante años había permitido descargar el archivo histórico de tuits. Los reemplazos de pago dejaron a cientos de proyectos de investigación sin fuente de datos.
- El 14 de agosto de 2024, Meta cerró CrowdTangle, la herramienta con la que investigadores y periodistas seguían la circulación de contenido en Facebook e Instagram, y la sustituyó por Meta Content Library, con acceso más restringido.

La consecuencia metodológica es directa: mucha de la literatura clásica del área se construyó sobre datos que hoy no se pueden recolectar, y replicar esos estudios en las mismas plataformas ya no es posible. Los proyectos actuales se apoyan más en plataformas abiertas (Mastodon, Bluesky, Reddit, Telegram), en datos donados y en conjuntos públicos.

Fuentes: Center for an Informed Public, University of Washington, sobre el cierre de la API académica de Twitter, 2023. <https://www.cip.uw.edu/2023/02/02/twitters-api-access-changes-academic-research/> Columbia Journalism Review, sobre el cierre de CrowdTangle, 2024. https://www.cjr.org/tow_center/meta-is-getting-rid-of-crowdtangle.php

Muestreo y su sesgo

Casi nunca se tiene la red completa, y la forma de recolectar el pedazo determina el resultado.

Por hashtag o palabra	solo ve la conversación explícita; pierde a quien participa sin usar el término
Bola de nieve	partir de semillas y expandir por vecinos; sobrerrepresenta nodos de grado alto
Recorrido BFS	captura bien la componente cercana; nunca alcanza las regiones desconectadas
Ventana temporal	una semana de datos no revela estructura estable

El sesgo más peligroso es el que no se ve: una red muestreada tiene métricas globales distorsionadas. La densidad y la longitud media de camino calculadas sobre un fragmento no son estimadores de las de la red completa, así que reportar densidad de una muestra como si fuera la de la red es un error frecuente.

Construcción del grafo

Antes de medir nada hay decisiones que cambian los resultados.

- Desduplicar entidades: la misma persona con varias cuentas debe unificarse o no, y hay que decidirlo explícitamente.
- Umbral de arista: ¿una sola interacción crea vínculo, o hacen falta tres?
- Ventana temporal: qué período define la red, y si se analiza como estática o como secuencia de instantáneas.
- Autoenlaces y multiaristas: qué hacer con retuits propios o con interacciones repetidas.
- Filtrado por k-core: quitar nodos periféricos para que la estructura del núcleo sea visible.

Scraping

El web scraping consiste en extraer datos de un sitio leyendo las páginas que el servidor entrega a un navegador, en lugar de pedirlos por una interfaz diseñada para programas.

Aparece en esta lección porque, después del cierre de las APIs, para muchas preguntas de redes sociales es la única vía que queda. También es la más frágil, la más lenta y la que más problemas legales trae, así que se usa sabiendo lo que se está aceptando.

API contra scraping

	API oficial	Scraping
Formato	JSON estable	HTML que cambia sin aviso

Permiso	explícito, con clave	ausente o prohibido
Límites	documentados	desconocidos, se descubren al chocar
Campos	los que exponga	los que se vean en pantalla
Mantenimiento	bajo	alto: se rompe solo
Costo	dinero	tiempo de ingeniería

La regla de decisión es simple: si existe una API que responde la pregunta, se usa la API. El scraping es el recurso cuando no hay alternativa, no un atajo para evitar leer documentación.

El proceso

1. Cola de URLs: la lista de páginas por visitar, alimentada por las semillas y por los enlaces que se van descubriendo.
2. Solicitud HTTP: pedir la página, con espera entre solicitudes.
3. Guardar el HTML crudo: archivar la respuesta tal como llegó, antes de extraer nada.
4. Parsear y extraer: aplicar los selectores para sacar los campos, y encolar las URLs nuevas.
5. Normalizar entidades: unificar identificadores de usuario, fechas y nombres.
6. Construir el grafo: convertir las interacciones extraídas en nodos y aristas.

Guardar el crudo primero es lo que vuelve reproducible el trabajo. Cuando el parser resulte incorrecto —y va a resultarlo— se vuelve a extraer del archivo local en vez de volver a golpear el sitio.

Estático contra dinámico

Página estática	el HTML ya trae el contenido requests + BeautifulSoup, o Scrapy rápido y barato
-----------------	---

Página dinámica	el HTML llega vacío y JavaScript lo llena después navegador sin interfaz: Playwright o Selenium lento, pesado, se rompe con cada rediseño
Endpoint interno	la propia página llama a una API no documentada se copia esa llamada desde las herramientas del navegador lo más eficiente cuando funciona, lo más inestable

Para saber cuál es, se pide la página sin ejecutar JavaScript y se busca el dato en la respuesta. Si aparece, es estática. Casi todas las redes sociales modernas son dinámicas, que es la razón de que scrapearlas cueste tanto.

Lo que rompe un scraper

- Rediseños: un cambio de clase CSS invalida todos los selectores a la vez.
- Límites de tasa: respuestas 429 y bloqueos por IP al pedir demasiado rápido.
- Defensas: CAPTCHA, huella del navegador, muros de inicio de sesión.
- Paginación infinita: el contenido solo aparece al desplazarse, y el final no está definido.
- Personalización: dos sesiones ven contenido distinto, así que lo recolectado no es una muestra limpia.
- Bloqueo silencioso: el sitio responde 200 con datos incompletos y el scraper no se entera. Es el más peligroso porque no falla, solo miente.

Buenas prácticas

1. Leer robots.txt y los términos de servicio antes de escribir código, no después del primer bloqueo.
2. Limitar el ritmo: una solicitud por segundo como máximo, y menos en sitios pequeños. Reintentos con espera exponencial y algo de azar, nunca en bucle cerrado.

3. Identificarse: User-Agent con nombre del proyecto y forma de contacto.
Suplantar un navegador para evadir defensas cambia el problema de técnico a ético.
4. Recolectar solo lo necesario, de forma incremental. Menos datos personales guardados es menos riesgo que administrar.

Legalidad y ética

En hiQ Labs contra LinkedIn, la corte del Noveno Circuito de Estados Unidos sostuvo que raspar datos de páginas públicamente accesibles no constituye acceso no autorizado bajo la ley federal de fraude informático (CFAA).

Lo que ese fallo no dice es igual de importante. La misma empresa perdió después por incumplir el contrato de uso del sitio: no violar una ley penal no es lo mismo que tener permiso. Y ninguna de esas dos cosas resuelve la protección de datos personales ni la jurisdicción aplicable en Guatemala.

Fuente: Noveno Circuito, abril de 2022, sobre hiQ Labs contra LinkedIn y la CFAA.
<https://www.tradesecretslaw.com/2022/05/articles/computer-fraud-and-abuse-act/ninth-circuit-reaffirms-that-data-scraping-from-public-websites-does-not-violate-the-computer-fraud-and-abuse-act/>

El scraping después del cierre de las APIs

- Las bibliotecas que raspaban X sin autenticación dejaron de funcionar en 2023, cuando la plataforma cerró el acceso anónimo del que dependían.
- Lo que queda exige cuentas autenticadas, lo que expone esas cuentas a suspensión y añade un problema ético propio.
- Las plataformas abiertas (Mastodon, Bluesky, Reddit) publican APIs utilizables, y ahí el scraping deja de ser necesario.

- Los conjuntos ya recolectados y publicados son a menudo la opción más honesta y la más reproducible.

Para un proyecto de curso, la pregunta se elige en función de los datos que se pueden obtener de forma sostenible, no al revés.

Visualización

La visualización es la forma habitual de comunicar el resultado, y también la más fácil de convertir en decoración.

- Layouts dirigidos por fuerzas: Fruchterman-Reingold, ForceAtlas2. Simulan repulsión entre nodos y atracción por aristas; los grupos densos quedan juntos.
- Codificación visual: tamaño del nodo por centralidad, color por comunidad, grosor de arista por peso.
- El problema del ovillo: pasados unos miles de nodos el dibujo se vuelve una mancha sin información. Las soluciones son filtrar por k-core, agregar nodos por comunidad, o mostrar solo el subgrafo relevante.
- Una posición en el dibujo no es un dato: los layouts por fuerzas son estocásticos, y dos corridas producen imágenes distintas de la misma red. Lo que se puede leer es la agrupación, no las coordenadas.

Analítica de Plataforma

La otra acepción del término. Se mide desempeño de contenido, no estructura, y es lo que se encuentra en el trabajo diario de marketing y comunicación.

Alcance	cuentas únicas que vieron el contenido
Impresiones	veces que se mostró, con repeticiones
Engagement	interacciones: reacciones, comentarios, compartidos, guardados
Tasa de engagement	interacciones / alcance (o / seguidores; hay que declarar cuál)
CTR	clics / impresiones
Share of voice	menciones de la marca sobre el total de la categoría
Conversión	acciones de negocio atribuibles al contenido

Los puntos donde estas métricas engañan:

- Las definiciones cambian por plataforma y por año. Comparar la tasa de engagement de dos redes distintas casi nunca es válido.
- Alcance e impresiones no son lo mismo, y confundirlos infla los resultados.
- Un pico de menciones no dice si son elogios o linchamiento; ahí es donde entra el análisis de sentimiento visto en la lección anterior.
- Casi todas son métricas de vanidad si no se atan a un objetivo: sirven para comparar contra el propio historial, no como cifras absolutas.

La combinación útil es medir el contenido con analítica de plataforma y explicar por qué se difundió con analítica de redes.

Consideraciones Éticas

Analizar redes es analizar relaciones entre personas que no pidieron ser analizadas, y eso trae obligaciones concretas.

- Consentimiento de terceros: al recolectar la red de una persona se recolectan datos de sus contactos, que nunca aceptaron nada.
- Reidentificación: anonimizar quitando nombres no protege. La estructura de vecindario de un nodo suele ser suficiente para identificarlo cruzando con otra

fuentes.

- Inferencia de atributos sensibles: la homofilia permite predecir orientación, religión o postura política de una persona a partir de sus conexiones, aunque ella nunca los haya declarado. Es la técnica detrás del caso Cambridge Analytica.
- Experimentación sin consentimiento informado: manipular el feed de usuarios para medir su reacción es trivial técnicamente y no es un experimento aprobado por nadie.
- Usos de vigilancia: las mismas técnicas de detección de coordinación sirven para mapear organizaciones civiles o identificar a quienes se conectan con activistas. El método es neutro; su aplicación no.
- Daño por agregación: cada dato es público por separado, y el perfil resultante de juntarlos no lo es.

La pregunta operativa es siempre la misma: qué le pasa a la persona menos poderosa del grafo si el análisis se filtra.

Herramientas del Ecosistema

| Herramienta | Para qué sirve | |---|---| | NetworkX | Grafos en Python; la opción didáctica y la más común | | igraph | Núcleo en C con interfaces en Python y R; mucho más rápido | | graph-tool | Rendimiento alto y modelos estadísticos de bloques | | Gephi | Visualización interactiva y exploración; formato GEXF | | Cytoscape | Visualización, con fuerte adopción en redes biológicas | | Neo4j | Base de datos de grafos con consultas en Cypher | | SNAP | Biblioteca en C++ y repositorio de conjuntos de datos grandes | | PyTorch Geometric | Redes neuronales sobre grafos (GNN) | | Botometer | Puntuación de probabilidad de que una cuenta sea bot | | Scrapy | Marco de scraping con cola, reintentos y límite de ritmo | | Playwright | Navegador sin interfaz para páginas dinámicas | |

Regla práctica de escala: hasta unas decenas de miles de nodos NetworkX es suficiente; para millones se pasa a igraph o graph-tool, y para consultas interactivas sobre grafos persistentes, a una base de datos de grafos.

Resumen

- La analítica de redes sociales modela relaciones como un grafo y estudia la estructura; la minería de texto estudia el contenido. Se complementan, no se reemplazan.
- El término abarca dos prácticas: el análisis estructural con grafos y la analítica de desempeño de contenido en plataformas.
- Modelar es decidir qué es un nodo y qué es una arista, y con los mismos datos crudos se construyen redes distintas que responden preguntas distintas.
- La centralidad no es una sola cosa: el grado mide popularidad y la intermediación mide control del flujo, y sus rankings no coinciden.
- Las distribuciones de grado siguen leyes de potencias, así que el promedio deja de ser un resumen útil.
- Unos pocos atajos convierten una red agrupada en un mundo pequeño; en Facebook la distancia media medida fue de 3.74 intermediarios.
- Un lazo débil es una arista puente: la misma idea que mide la intermediación, vista desde la arista.
- La detección de comunidades optimiza modularidad; Louvain y Leiden son el estándar práctico, y el club de karate de Zachary muestra tanto su poder como su límite.

- Una cascada se describe por tamaño, profundidad, amplitud y velocidad; lo falso se propagó más lejos y más rápido, y fueron personas y no bots quienes lo hicieron.
- La coordinación inauténtica se detecta proyectando la actividad en una red de similitud entre cuentas, con casos documentados en Guatemala.
- El acceso a datos se cerró: la API académica de X en 2023 y CrowdTangle en 2024, lo que condiciona qué proyectos son viables hoy.
- El scraping es el recurso cuando no hay API: frágil, lento y sujeto a términos de servicio, con límite de ritmo y archivo del HTML crudo como mínimos no negociables.
- Que raspar datos públicos no viole una ley penal no significa tener permiso.
- El muestreo distorsiona las métricas globales, así que la densidad de una muestra no estima la de la red.
- Anonimizar quitando nombres no protege: la estructura reidentifica.

Referencias

- Zachary, W. W. (1977). *An Information Flow Model for Conflict and Fission in Small Groups*. Journal of Anthropological Research 33(4).
<https://www.journals.uchicago.edu/doi/10.1086/jar.33.4.3629752>
- Granovetter, M. (1973). *The Strength of Weak Ties*. American Journal of Sociology 78(6). <https://www.jstor.org/stable/2776392>
- Watts, D. y Strogatz, S. (1998). *Collective Dynamics of Small-World Networks*. Nature 393. <https://www.nature.com/articles/30918>
- Barabási, A.-L. y Albert, R. (1999). *Emergence of Scaling in Random Networks*. Science 286. <https://www.science.org/doi/10.1126/science.286.5439.509>

- Blondel, V. et al. (2008). *Fast Unfolding of Communities in Large Networks*. J. Stat. Mech. <https://arxiv.org/abs/0803.0476>
- Backstrom, L. et al. (2012). *Four Degrees of Separation*. ACM Web Science. <https://arxiv.org/abs/1111.4570>
- Vosoughi, S., Roy, D. y Aral, S. (2018). *The Spread of True and False News Online*. Science 359. <https://www.science.org/doi/10.1126/science.aap9559>
- CICIG (2019). *Bots, netcenters y combate a la impunidad: el caso de Guatemala*. https://www.cicig.org/wp-content/uploads/2018/05/Informe_bots_y_netcenters_2019.pdf
- Agencia Ocote y CLIP (2022). *Cómo mutaron los netcenters anti justicia y derechos humanos en Guatemala*. <https://www.agenciaocote.com/blog/2022/11/30/mutaron-netcenters-anti-justicia-yderechos-humanos-en-guatemala/>
- Noveno Circuito de EE. UU. (2022). *hiQ Labs contra LinkedIn: el scraping de datos públicos y la CFAA*. <https://www.tradeseecretslaw.com/2022/05/articles/computer-fraud-and-abuse-act/ninth-circuit-reaffirms-that-data-scraping-from-public-websites-does-not-violate-the-computer-fraud-and-abuse-act/>