Minería de datos y análisis de redes sociales: malabarismos de una experiencia de investigación

César Augusto Rodríguez Cano

El análisis de redes sociales (ARS) es una tradición de investigación de larga data que ha destacado en el contexto de las llamadas tecnologías digitales, en particular para el estudio de las plataformas de redes sociales. Este renovado interés se explica en parte por el advenimiento de uno de los fenómenos culturales más importantes de la época, el *big data*, en el marco del giro computacional de la cultura (Berry, 2011) y la cuantificación del ser (Swan, 2013), pero también por una coincidencia clave entre esta metodología y los nuevos espacios de socialización digital: su dimensión reticular.

El ARS no se debe confundir con el estudio de plataformas en línea. El análisis de redes sociales o *Social Network Analysis* es un enfoque cuantitativo creado mucho antes del nacimiento de Internet en el plano de la sociometría o indagación de las relaciones sociales. Para los interesados en la historia de esta metodología, sugiero recurrir a L. Freeman (2004), quien expone cuatro elementos fundacionales de esta tradición de investigación: intuiciones estructurales, datos relacionales sistematizados, imágenes gráficas y modelos matemáticos.

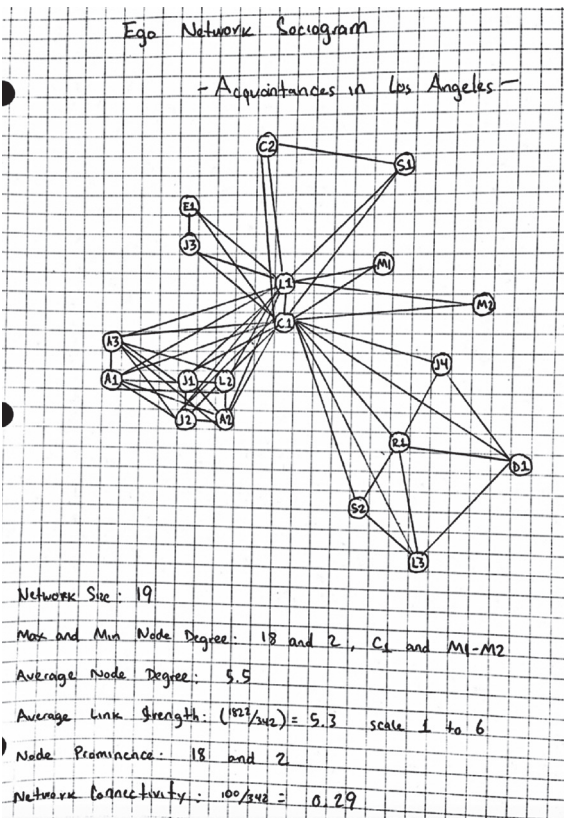
Mi primer encuentro con el ARS[1] fue en el seminario metodológico que tomé en la Escuela de Posgrado en Estudios sobre Educación e Información de la Universidad de California en Los Ángeles (UCLA), a finales de 2013, durante mi año de estancia como Investigador de Posgrado Visitante. No se trató tanto de un curso totalmente dedicado al uso de la metodología y la teoría de grafos, sino al repaso de métodos multivariados entre los cuales pude conocer los fundamentos del ARS para desarrollar mi primera red —dedicada a vínculos interpersonales en Los Ángeles (véase Imagen 1)— y asistir a un taller sobre el software Gephi, dirigido por Zoe Borovsky dentro del curso cuantitativo que impartió Leah Lievrouw, conocida entonces por su libro *Alternative and Activist New Media* (2011). Al estudiar Twitter como parte de mi tesis doctoral, esta novedad en mi formación se convirtió de inmediato en un aprendizaje alentador al ajustarse a uno de los objetivos de mi estancia de investigación: renovar el repertorio de técnicas y metodologías de investigación en los ambientes digitales. En ese entonces, ya era evidente el protagonismo de las plataformas de redes sociales en la transformación de la vida pública del país, por ejemplo, en términos de consignas ciudadanas con la movilización conocida como #YoSoy132 en el contexto de las elecciones presidenciales en México de 2012.

Así como asistí al seminario cuantitativo, más adelante pude tomar un curso cualitativo con Chris Kelty, quien acababa de publicar el libro *Two Bits: The Cultural significance of free software* (2008), sobre el enfoque antropológico y las

1 En este libro, opté por un enfoque testimonial con tintes autoetnográficos sobre mi experiencia en la aplicación del análisis de redes sociales, en lugar de una especie de manual sobre sus principales elementos y aplicaciones. Por lo mismo, he decidido orientar el sentido de lo escrito al lector con conocimientos mínimos de esta metodología. Para una explicación a detalle, recomiendo los textos de Lozares (1996), Wasserman y Faust (1994) y Molina (2001).

humanidades digitales, lo que complementó mi visión de investigador en el área de estudios de la comunicación y de muchas maneras marcó mi perspectiva de la metodología como un modelo para armar aprendizajes abiertos e incesantes en que lo cuantitativo y lo cualitativo constantemente se traslapan.

Imagen 1
Visualización de mi red social de conocidos en Los Ángeles, 2013



Fuente: Elaboración propia.

Todo esto estuvo enmarcado en la coasesoría de mi proyecto de investigación sobre el uso de Twitter en las elecciones presidenciales de 2012 en México, por parte de Ramesh Srinivasan —junto con Raúl Trejo Delarbre como tutor principal—, quien en ese entonces estaba preparando su libro *Whose Global Village? Rethinking How Technology Shapes Our World* (2017) y recién había publicado un artículo que me parecía clave para entender las movilizaciones digitales englobadas en la llamada primavera árabe, mucho antes de que la dimensión conectiva se apoderara de la discusión: *Bridges between Cultural and Digital Worlds in Revolutionary Egypt* (2013). En ese entonces había pocos artículos publicados por investigadores que habían trabajado directamente en estas geografías, que junto con las del movimiento de los indignados en España y el Occupy Wall Street en Estados Unidos, eran parte de la irrupción global de las multitudes conectadas, como teóricamente se les comenzó a llamar a partir de la noción de tecnopolítica (Toret, 2013). A la par, como complemento del ambiente de insurgencia, pero también de la teoría que se respiraba en esos años en Los Ángeles, me escapaba regularmente para escuchar a Manuel Castells en la Annenberg School of Communication de la Universidad del Sur de California, quien acababa de presentar su influyente libro *Redes de indignación y esperanza* (2012), producto de una mirada de las movilizaciones sociales desde el llamado giro afectivo que había explorado ya desde Comunicación y poder (2009) con perspectivas en red.

Al mismo tiempo, tuve la posibilidad de asistir a un taller de innovación en la investigación en el desbordante MIT Media Lab en Boston, con el equipo de investigación de Ethan Zuckerman, director del Center For Civic Media, en el marco de una gira que me llevó a conocer los espacios, bibliotecas y proyectos del Berkman Center for Internet and

Society, de la Universidad de Harvard, la UC Berkeley School of Information (para escuchar a Danah Boyd), y el Center for Ethnography de la Universidad de California en Irvine (en un coloquio con Tom Boellstorff), así como las universidades de Stanford, CalTech, UC Riverside y Santa Barbara, Columbia, la New York University y la New School for Social Research. Asimismo, en un lado de la moneda que me interesaba para reconfigurar la mitología que rodeaba Internet, también dediqué tiempo para recorrer Sillicon Valley, en particular las oficinas de Google, Amazon y Facebook —esta última irónicamente ubicada en Hacker Way—, así como las de Twitter en San Francisco. Un cúmulo de experiencias que me parece importante recordar porque la formación metodológica en cualquier área científica debe considerarse más un aprendizaje holístico relacionado con nuestra historia de vida, que solamente con un curso académico y el desarrollo de un tema de investigación.

Esta experiencia, valga mencionarlo, fue la clave de la propuesta metodológica que realicé en mi tesis doctoral, además de la columna vertebral de los cursos sobre Métodos Digitales Cuantitativos y Métodos Digitales Cualitativos que desde 2017 pude impartir en el Programa de Posgrado en Ciencias Políticas y Sociales de la Universidad Nacional Autónoma de México (UNAM), en donde redoblé esfuerzos para fortalecer el debate de la cuestión metodológica en los estudios sobre Internet, con el reto de colaborar en el diseño de proyectos de investigación relacionados.

Cabe mencionar que, además del ARS, durante este trayecto como profesor, profundicé mi conocimiento sobre otras técnicas, autores, escuelas y tradiciones relevantes que sugiero conocer, tales como la etnografía digital y el destacado trabajo de Édgar Gómez Cruz, Elisenda Ardèvol, Adolfo Estalella, Tom Boellstorff, Sarah Pink y Christine Hine, entre otros; la escuela

de métodos digitales, de la Universidad de Amsterdam, con Richard Rogers a la cabeza; la analítica cultural, propuesta por Lev Manovich para trabajar con grandes cantidades de datos; técnicas emergentes de las que resalto el análisis de sentimientos, el análisis crosesférico y la tecnografía, entre otras tradiciones como el análisis crítico del discurso aplicado a las narrativas en línea.

Respecto al uso de análisis de redes sociales, una de las consideraciones emanadas del seminario en la UNAM y de los incontables ejercicios realizados en clase, fue la propuesta de integrar a manera de taller este análisis con dos posibilidades en la obtención de datos: de arriba hacia abajo (*top-down)* y de abajo hacia arriba (*bottom-up*). La primera a partir de la descarga masiva en bases de datos relacionales, un fenómeno conocido como minería de datos, en las cuales las variables no son creadas por el investigador y están a expensas de los alcances y anaqueles del mecanismo de extracción —una situación que ha despertado reflexiones sobre el postempirismo y los riesgos de desarticular la indagación —. No obstante, ha sido reivindicado en la tradición de técnicas de investigación en el que la replicabilidad y validez no necesariamente recaen en categorías preestablecidas (Rose, 2016).

La minería de datos, también conocida como *data mining*, es la actividad mediante la cual se extraen datos y metadatos de la actividad social en los ambientes digitales, con el objetivo de analizarlos para realizar inferencias con sentido en direcciones que van desde la publicidad y la mercadotecnia hasta el diseño de políticas públicas y la investigación académica. Rogers ha considerado esta información un tipo de datos posdemográficos (2013), mientras que han surgido enfoques más críticos para definir este fenómeno en el marco de la cultura que describe, a partir de nociones como sociedad de la transparencia (Han, 2014), justicia de datos (Dencik,

Hintz, Redden y Treré, 2019) o capitalismo de la vigilancia (Zuboff, 2019).

La segunda posibilidad para obtener datos corresponde al modo tradicional y se enfoca en la transición de la creación de matrices a la búsqueda de datos a conveniencia teórica del analista, con la limitante de que el número de campos es mínimo y no puede ser integrado dentro del fenómeno del *big data*, aunque sí en el marco de las sociedades datificadas como veremos a continuación.

El domador de datos

El análisis de las redes sociodigitales me enfrentó a la noción conceptual de *big data*, aspecto nodal para entender la cultura de la hiperconectividad y que se refiere a la explotación de datos en cantidades masivas, gracias a las posibilidades tecnológicas derivadas del rastreo, extracción y almacenamiento de las actividades sociales mediante dispositivos computacionales, un aspecto que posibilita observar las microinteracciones por primera vez en la historia de las ciencias sociales, las trazas digitales (Venturini y Latour, 2010). De acuerdo con uno de los artículos más citados sobre este tema en el área social, se trata de un fenómeno cultural con una cierta carga mitológica de trascender la versión estadística de la muestra para contener la totalidad del universo de estudio (Boyd y Crawford, 2012), algo que en mi experiencia como investigador ha sido imposible por una de sus características: la masividad. En las extracciones que he logrado hacer a lo largo de los años, sobre todo en Twitter, una de las dudas más acuciantes es cómo funciona el mecanismo de selección que mina los datos, pues las descargas son limitadas y representan solo una pequeña parte de lo existente. Por supuesto que en fenómenos con una interacción menor, es probable la obtención de la población total, pero ya no es *big data*.

Es decir, el imperativo de la obtención de datos nos hace quedar entrampados en la paradoja de que mientras menor sea la población también es limitada la posibilidad de considerarse *big data*, pero mayor la de estudiarse en su totalidad. Por el contrario, cuando es mayor la cantidad de datos extraídos puede llegar a considerarse *big data,* pero es común que no represente la totalidad del universo, por lo que es necesario mayor rigor en la justificación del muestreo. En este sentido, la selección del corpus en términos de grandes datos exige una cuidadosa explicación sobre su matiz representativo.

El Gran Archivero, como he propuesto traducir el fenómeno de los datos masivos (Rodríguez Cano, 2020), es un canon económico, político, social y, de nuestro interés, analítico. Las famosas tres V que lo definen: velocidad, variedad y volumen resultan en sendos desafíos para la investigación social. La velocidad exige una instantaneidad de captura inusitada; la variedad un proceso de limpieza y estructuración que es difícil realizar rudimentariamente y el volumen un procesamiento y almacenamiento computacional monumental. Nuevamente, en mi experiencia de investigación, he tenido que recurrir a diferentes diseños para plantear caminos más o menos estables en el trabajo de análisis.

Como señala Meneses Rocha (2018), los grandes datos son un gran desafío para las ciencias sociales, en parte por las dificultades en cada uno de los elementos necesarios en su cadena de valor: generación, recolección, almacenamiento, procesamiento, distribución y análisis. De acuerdo con este punto de partida, el *big data* exige un volumen casi ilimitado, velocidad rápida y continua, y una variedad amplia. Contrario a lo que Rogers (2013) denomina *small data*, que precisa un volumen limitado, velocidad lenta y una variedad también limitada. Con esta distinción, el trabajo que he realizado a lo largo de estos años con extracción de información y

minería de datos ha sido más desde una perspectiva de los pequeños datos.

Sin embargo, aquí es importante hacer una aclaración. El *big data* no solo es un fenómeno tecnológico, por ende detonador de desafíos técnicos, sino que representa en mayor medida un fenómeno social cuya principal característica es la datificación de la cultura, una de cuyas fuentes es la sociedad de las plataformas (Van Dijck, Poell y De Waal, 2018). Desde esta perspectiva, al hacer minería de datos, incluso en cantidades menores, nos ubicamos en el marco de la datificación y por lo menos de inicio a la sombra de los grandes datos como paradigma de la época.

Como señala la misma Meneses, los desafíos respecto al *big data* en las ciencias sociales son de varios órdenes: en primer lugar, la importancia de tomar distancia de los discursos técnicos y mercadológicos que buscan objetivos más concretos sin fundamento, más que la propia utilidad de los datos, una tendencia explicada por los flujos cuantificables de cascadas mercantiles que caracterizan a la cultura digital. En segundo lugar, señala la autora, existe el reto de no ser avasallados por el saber computacional y, por el contrario, poner a la tecnología al servicio del conocimiento de lo social. Esto, quiero argumentar, supone no reivindicar la dimensión mitológica del *big data* al cuestionar investigaciones que tienen una menor cantidad de datos, muchas veces obtenidas al límite de las capacidades técnicas, siempre y cuando se encuentren en la lógica de la investigación científica.

De hecho, Brooker, Barnett, Cribbin y Sharma (2016) señalan que a pesar de los límites del *data mining,* podemos hacer análisis significativos, mientras se tome en cuenta la comprensión profunda de cómo se ha construido el conjunto de datos con la intención de diseñar aproximaciones analíticas apropiadas con las cuales lidiar, esto es pensar en procesos

de obtención de los datos, no en los datos exclusivamente, entendiendo que la traducción de lo social convertido en datos envuelve un proceso de abstracción que impone ciertos compromisos en la forma en que los datos son generados, recolectados, seleccionados y analizados (Schäfer y Van Es, 2017).

En este sentido, en la discusión sobre grandes o pequeños datos, fue bastante conocida la mirada antropológica que aportó Wang (2013), con el término datos densos. Como imaginará el lector, en una analogía con la descripción densa que propuso Clifford Geertz (1973), por lo tanto, un enfoque desde la dimensión interpretativa de la cultura, esta autora cuestiona la reverencia al *big data* al reiterar la importancia de trabajar los datos en el sentido inverso, a conveniencia dela investigación y con estrictas precauciones: cercanía, precisión, descubrimiento, interpretación. Contrario a las bases que puedes descargar en las modernas aplicaciones, el acercamiento desde los datos densos permite elegir el corpus de investigación de la manera tradicional, en concreto las variables que conforman la matriz de información, con compleja comprehensión del fenómeno estudiado y sin imperativos cuantitativos inmanentes. En el sentido de los datos densos, como veremos, mi estrategia ha sido buscar la obtención de datos de abajo hacia arriba.

Equilibrista del software

El uso de Gephi fue parte de mi primer experiencia con el análisis de redes sociales, como he comentado. No así con software con enfoque cuantitativo, pues ya había trabajado con estadística descriptiva, inferencial y diseño factorial en el Paquete Estadístico para Ciencias Sociales (SPSS, por sus siglas en inglés) durante el procesamiento de bases de datos y generación de gráficas derivadas de cuestionarios para