การวิเคราะห์การตรวจจับชุมชน: กรณีศึกษาความเกี่ยวข้องกับเพจเฟซบุ๊คขนาดใหญ่ในช่วงเดือนพฤศจิกายน 2017

Community Detection: The Case Study of the Relevance of Large Facebook Page in November 2017

นายกริช ปรัชญาวาทิน (63199130106)
สาขาวิชาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
กรุงเทพฯ ประเทศไทย
E-mail: krich.prach@g.swu.ac.th

บทคัดย่อ (Abstract)

ในบทความนี้ เป็นการทดลองจากสิ่งที่ได้เรียนในวิชา Social Network Analysis โดยชุดข้อมูลที่ใช้จะเป็นข้อมูลเกี่ยวกับ Facebook page ขนาดใหญ่ที่มีความเกี่ยวข้องกันโดยมีการเก็บในช่วงเดือน พ.ย.2017 [1] โดยศึกษาจากสมาชิกเครือข่ายในชุมชน ความเกี่ยวข้องของเครือข่าย และความเข้ากันของประเภทของเพจ เพื่อวิเคราะห์ถึงความเป็นไปได้ที่แต่ละเพจของ Facebook

จากการสำรวจและทดสอบ พบว่า เพจ Facebook ส่วนใหญ่จะมีความเกี่ยวข้องกันในส่วนประเภท Category ของเพจ เช่น เพจบริษัทเกี่ยวกับเทคโนโลยี ก็จะมีความเกี่ยวข้องกับเพจของบริษัทที่มีความเกี่ยวข้องกับเทคโนโลยีเช่นกัน หรือเพจของนักการเมือง ก็จะมีความเชื่อมโยงกับเพจของนักการเมืองในประเทศนั้น ๆ ทำให้เห็นว่าความเชื่อมโยงส่วนใหญ่ของเพจใน Facebook เกี่ยวข้องกับ Category ของแต่ละเพจ

คำสำคัญ (keywords) – community detection, Facebook Page, ความสัมพันธ์ระหว่างเพจ, เพจเฟซบุ๊ค

บทนำ (Introduction)

เนื่องจากในปัจจุบัน เรามีการใช้เครือข่ายสังคมออนไลน์มากมาย ทั้ง Facebook, Instagram, Twitter และอื่น ๆ เครือข่ายเหล่านี้ไม่ได้อยู่แต่ในด้านเทคโนโลยี แต่มีความเกี่ยวข้องกับผู้คนในสังคม ในทุกกลุ่ม เนื่องจากหลายคนสามารถเข้าถึงอินเทอร์เน็ตได้ และมีการใช้งานกันเป็นชีวิตประจำวัน ชุมชนเป็นส่วนหนึ่งของเครือข่ายที่มีอยู่เป็นจำนวนมหาศาล ซึ่งเครือข่ายหนึ่งสามารถมีหลายชุมชน ที่โหนดภายในชุมชนมีการเชื่อมโยงกันอย่างแน่นหนา จึงทำให้หลายกลุ่มต้องการที่จะศึกษาในการตรวจจับชุมชน (Community Detection) เพื่อทำความเข้าใจ ศึกษาในความสัมพันธ์ของแต่ละเครือข่ายชุมชน เพื่อให้เราได้เข้าใจพฤติกรรมของแต่ละชุมชนและปรับตัวเข้ากับแต่ละเครือข่ายในชุมชนต่อไปได้

การตรวจจับชุมชน (Community Detection) คือ การทำความเข้าใจในโครงสร้างของเครือข่ายที่มีความซับซ้อน สามารถนำเอาข้อมูลส่วนที่มีประโยชน์เพื่อนำมาปรับใช้ต่อไป [2]

แนวคิดในการตรวจจับชุมชนคือ การค้นหากลุ่มที่ซับซ้อนที่ออกมาในรูปแบบกราฟ สามารถเห็นเครือข่ายย่อย ๆ ที่มีความเชื่อมโยงกันในระหว่างโหนดในกลุ่มเดียวกันมากกว่าโหนดในกลุ่มต่าง ๆ ทางสถิติ โดยมีศูนย์กลางของการตรวจจับชุมชนคือ แนวคิดในเชิงโมดูลาร์ซึ่งมีเมตริกที่สามารถตรวจจับความแตกต่างของชุมชนได้ [3]

งานที่เกี่ยวข้อง (Related Works)

Zhige Xin, Chun-ming Lai, Jon William Chapman (2019) ได้ทำการศึกษา Facebook Page ของ CBS News และ The New York Times พบว่า เครือข่ายมีความกระจัดกระจาย และโครงสร้างของชุมชนที่หนาแน่น สามารถพบได้ในช่วงระหว่างการเข้าใช้ในหน้าของ New York Times [4]

Pranita Jain และ Deepak Singh Tomar (2019) ได้ทำการปรับขนาดและคุณภาพของเครือข่ายชุมชนขนาดใหญ่ตามพารามิเตอร์ เมื่อทำการประเมิน พบว่า การตรวจจับชุมชน (Community Detection) ให้ผลดีกับการปรับขนาดของเครือข่ายชุมชน [5]

Victor Stany Rozario, AZM Ehtesham Chowdhury, Muhammad Sarwar Jahan Morshed (2019) ได้นำเสนอ Interlinked Spatial Clustering Model (ILSCM) ซึ่งสามารถทำการตรวจจับชุมชน (Community Detection) ตามเนื้อหาและ edges มีการเชื่อมโยงลิงก์ระหว่างผู้ใช้กับโหนดโดยรอบ ซึ่งมีการเชื่อมโยงที่ดี มีการจัดเรียงหัวข้อ และกราฟแสดงผลที่เหมาะสม [6]

เบื้องหลังทางเทคนิค (Technical Background)

เนื่องจากในช่วงหลัง ได้มีการเรียน Social Network Analysis โดยใช้ Neo4j ซึ่งสามารถวิเคราะห์เครือข่ายสังคมได้ จึงได้ทำการทดลองการเรียกดูข้อมูลจากชุดข้อมูลที่มีอยู่ตามขั้นตอน จนได้มาพบว่า ไม่สามารถทำการเชื่อมข้อมูลเข้าหากันได้ เนื่องจากชุดข้อมูลเป็นแบบ Undirected Graph ที่เป็นกราฟไม่มีทิศทางที่ชัดเจน ซึ่งในส่วนของ Neo4j สามารถทำได้เพียงข้อมูลแบบ Directed Graph ที่มีทิศทางของกราฟที่ชัดเจนเท่านั้น จึงได้ลองเปลี่ยนวิธีการโดยการไปทดลองทำโดยใช้ Python ในส่วนของ NetworkX แทน เพื่อให้สามารถทำการสำรวจข้อมูลต่อไปได้

โดยในการทำครั้งนี้เป็นการสำรวจตามความเข้าใจ ไม่เน้นไปทางทฤษฎีที่จะต้องมีสูตรคำนวณ แต่จะเน้นไปในภาพรวมที่สามารถเข้าใจได้ง่าย

ขั้นตอนการทำงาน (Methods)

ขั้นตอนในการทำงาน มีการอ้างอิงแนวทาง ขั้นตอนการทำงานมาจาก [7] มีขั้นตอนการทำ ดังนี้

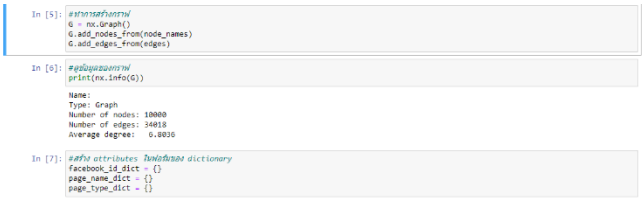
1. ทำการค้นหาชุดข้อมูลที่เหมาะสมในการทำการทดลอง จนได้ค้นพบกับชุดข้อมูล Facebook Large Page-Page Network จึงได้นำข้อมูลชุดนี้มาศึกษา
2. ทำการทดลองกับชุดข้อมูลในครั้งแรกกับ Neo4j ผลออกมาคือ มีชุดข้อมูลมากเกินไป ทำให้เสียเวลาในการประมวลผลของข้อมูล
3. ทำการลดจำนวนของข้อมูลลง เหลือในส่วนชุดข้อมูลของเพจจำนวน 10,000 เพจ ลดการเชื่อมโยงของข้อมูลที่เกี่ยวข้องกับข้อมูลที่อยู่ในลำดับที่ 10,001 เป็นต้นไป ทั้งหมดออกในทุกไฟล์ของชุดข้อมูล และตัดสินใจใช้เพียงแค่ไฟล์ fb_target_edge.csv สำหรับไฟล์ node และไฟล์ fb_edges_edit.csv สำหรับไฟล์ edges
4. ทำการทดลองกับชุดข้อมูลที่ทำการลดจำนวนของข้อมูลแล้ว สามารถทำงานกับชุดข้อมูลได้ดี แต่ก็พบปัญหาการไม่เชื่อมของโหนด จนมาพบว่า Neo4j ไม่เหมาะกับการทำงานกับชุดข้อมูลนี้ จนได้ปรึกษากับอาจารย์จนได้ข้อสรุปว่า ให้ลองกลับไปใช้กับ Python ที่สามารถทำงานร่วมกันได้กับ Undirected Graph
5. นำเข้าข้อมูลลงใน Python โดยในการทดลองนี้ ใช้วิธีนำเข้าลงใน Jupyter notebook

รูปที่ 1 import CSV และ NetworkX และนำเข้าข้อมูล CSV ลงใน Jupyter

6. ทำการสำรวจข้อมูล ดูว่ามีกี่โหนด กี่จุดเชื่อม

รูปที่ 2 สำรวจโหนดและจุดเชื่อมของกราฟ

7. ทำการสร้างกราฟ ดูข้อมูลโดยรวมของกราฟ และสร้าง attribute ของข้อมูล ในฟอร์มของ dictionary



รูปที่ 3 สร้างกราฟ ดูข้อมูลของกราฟ สร้าง attribure ข้อมูล

8. ทำการตั้งค่า Loop ในลิสต์ของโหนด จากนั้นก็ทำการตั้งค่า node attributes

รูปที่ 4 ตั้งค่า loop ในลิสต์โหนด และตั้งค่า node attributes

9. สำรวจดูว่า สามารถดูข้อมูลของโหนดได้หรือไม่

รูปที่ 5 สำรวจดูข้อมูลในโหนดในที่นี้ดูในส่วนชื่อของเพจ

10. ดูและตั้งค่า Network density และตั้งค่า degree

รูปที่ 6 ดูค่า Network density และตั้งค่า degree

11. สำรวจข้อมูลแบบเป็นรายโหนด เพื่อดูข้อมูลในโหนดว่ามีอะไร เป็นอย่างไรบ้าง จากนั้นก็ดู sort ของ degree

รูปที่ 7 สำรวจดูข้อมูลเป็นรายโหนด เพื่อดูข้อมูลว่ามีอะไรบ้าง

รูปที่ 8 สำรวจดูข้อมูลเป็นรายโหนด เพื่อดูข้อมูลว่ามีอะไรบ้าง

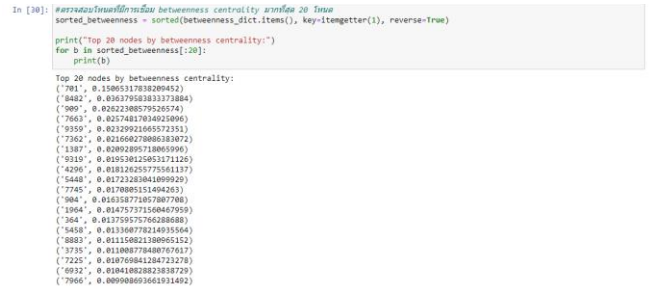
12. ทำการสำรวจว่าโหนดไหนมีการเชื่อมของ degree มากที่สุด 20 อันดับ จากนั้นก็สำรวจดโหนดนั้นว่าเป็นอย่างไร

รูปที่ 9 สำรวจดูโหนดที่มีการเชื่อม degree มากที่สุด 20 อันดับ พร้อมดูโหนดนั้นว่าเป็นอย่างไรบ้าง

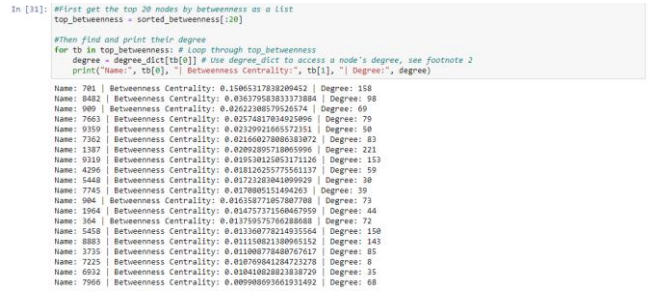
13. สร้างการเชื่อมต่อ betweenness และ eigenvector

รูปที่ 10 เชื่อม betweenness และ eigenvector

14. สำรวจดูว่า 20 อันดับของโหนดที่มีการเชื่อมbetweenness เป็นอย่างไรบ้าง

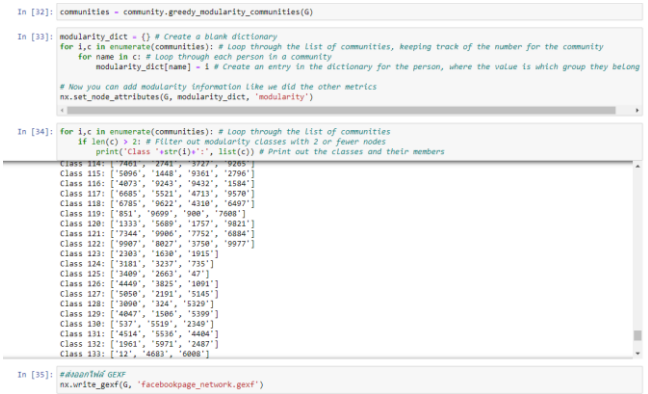


รูปที่ 11 สำรวจ 20 อันดับของโหนดที่เชื่อมแบบ betweenness มากที่สุด



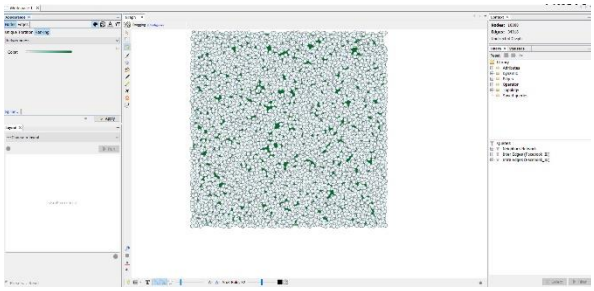
รูปที่ 12 สำรวจ 20 อันดับของโหนดที่เชื่อมแบบ betweenness มากที่สุด โดยดูเพิ่มในส่วนของข้อมูล

15. สำรวจในส่วน community detection และ modularity และส่งออกไฟล์เป็น .gexf



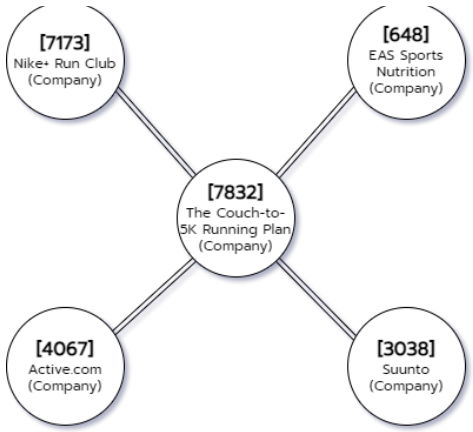
รูปที่ 13 สำรวจข้อมูลส่วน community detection และ modularity จากนั้นก็ส่งออกไฟล์เป็น gexf

16. ลองเปิดข้อมูลสำรวจใน Gephi โดยใช้ไฟล์ gexf ที่ได้จากการส่งออกข้อมูลใน Jupyter มาเปิดเพื่อให้ได้เห็นภาพของ network ได้มากขึ้น แต่ก็ไม่สามารถดูข้อมูลในแต่ละโหนดได้เนื่องจากมีข้อมูลที่มากเกินไป แต่สามารถดูข้อมูลที่มีการประมวลผลได้ ทำได้เพียงเห็นภาพว่า ส่วนที่มีการเชื่อมโยงกันมาก มีจำนวนน้อยกว่าจำนวนปกติ



รูปที่ 14 ภาพข้อมูลจากการเปิดใน Gephi

17. ทำการสร้างเป็น diagram เพื่อให้เห็นรูปแบบของการเชื่อมต่อที่น่าสนใจ



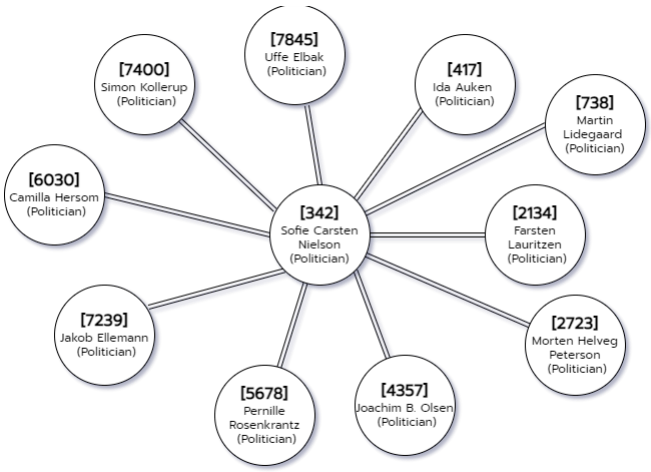
รูปที่ 15 ตัวอย่าง diagram การเชื่อมต่อของโหนด 7832 หมวด Company

ผลการทดลอง (Experimental Results)

จากการทดลอง สำรวจข้อมูลทั้งหมด พบได้ว่า จำนวนของข้อมูลมีผลกับการสำรวจเป็นอย่างมาก เนื่องจากในการประมวลผลจะต้องมีการใช้เวลาของเครื่องคอมพิวเตอร์ในการประมวลผลอยู่พอสมควร

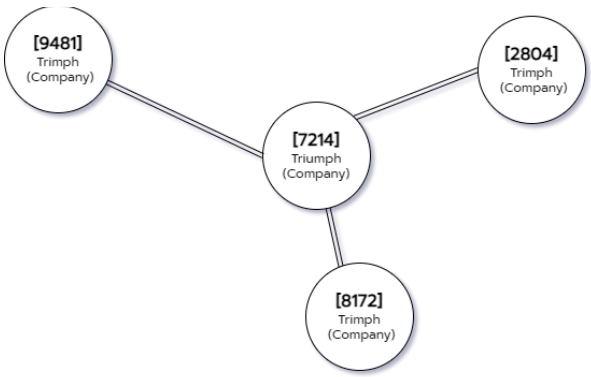
เมื่อได้ทำการสำรวจในหลายส่วนของข้อมูลชุดนี้ ได้ทำการปรับปรุงข้อมูลแล้วแล้ว พบว่า

1. โหนดที่มีการเชื่อมต่อกันมากที่สุดคือโหนด 1387 เพจ “Honolulu District, U.S. Army Corps of Engineers” หมวด Government มีการเชื่อม degree ถึง 221 โหนด เลยทีเดียว ทำให้เห็นว่า ในหมวด Government มีการพบปะ รู้จักกันและกันได้มากที่สุด
2. หมวดที่มีการเชื่อม Betweenness Centrality มากที่สุด คือ โหนด 701 “Facebook” ซึ่งอยู่ในหมวด Company มีค่าในส่วนนี้สูงถึง 0.15065317838209452
3. จากการสำรวจในหมวดอื่น พบว่า ส่วนใหญ่การเชื่อมโยงถึงกัน มักจะอยู่ในหมวดเดียวกัน เช่น หมวดนักการเมือง (Politician) ดังภาพจะเห็นว่า Sofie Carsten Neilson นักการเมืองจากเดนมาร์ก มีการเชื่อมโยงกับนักการเมืองในชาติเดียวกัน แสดงให้เห็นว่านักการเมืองแต่ละท่านมีการติดต่อสื่อสาร ทำความรู้จักกันในประเทศของตนมาก



รูปที่ 16 diagram การเชื่อมโยงของ Sofie Carsten Neilson นักการเมืองชาวเดนมาร์ก

4. ในแต่ละบริษัท ก็จะมีเพจของตัวเอง ทำให้มีการเชื่อมโยงถึงกันของในเพจในต่างพื้นที่หรือต่างสาขากันขึ้น ดังภาพจะเห็นได้จาก Triumph มีมากกว่า 1 เพจ



รูปที่ 17 diagram การเชื่อมโยงของบริษัท Triumph ที่มีเพจชื่อเดียวกัน 3 เพจ

5. โหนดที่น่าสนใจโหนดหนึ่งที่ค้นพบคือ โหนด 986 รายการ series ทางโทรทัศน์เรื่อง Marvel's DareDevil มีการเชื่อมโยงเพียงโหนดเดียว แต่ต่างหมวดกัน ซึ่งอีกโหนดอยู่ในหมวด Company คิดว่าน่าจะเพราะเป็น Sponsor ให้หรืออาจจะมีการรู้จักกันในระหว่าง 2 เพจ ดังจะเห็นได้จากภาพ

รูปที่ 18 diagram การเชื่อมโยงรายการ Marvel's DareDevil ที่มีการเชื่อมโยงกับเพจที่อยู่คนละหมวดกัน

บทสรุป (Conclusions)

การสำรวจข้อมูล ศึกษาชุดข้อมูลนี้ ทำให้เราได้เรียนรู้ถึงแนวทางและวิธีการในการสำรวจข้อมูลการเชื่อมโยงของเครือข่ายในหลายรูปแบบ ทั้งจาก Neo4j ทั้งจาก Python ทั้งจาก Gephi หรือจากการสำรวจข้อมูลภายในไฟล์ของชุดข้อมูลว่ามีการเชื่อมต่ออะไรน่าสนใจ โดยเลือกจากการสุ่มมาหลาย ๆ category

เมื่อเราได้ลองสำรวจ ลองทำดู จึงทำให้เห็นว่า

1. Facebook page ส่วนใหญ่ที่มีความสัมพันธ์กัน มักจะเป็นเพจที่มีลักษณะที่ใกล้เคียงกัน โดยเฉพาะในส่วนของความสนใจ หากเพจหนึ่งเป็นเพจเกี่ยวกับนักการเมือง ก็จะมีความสัมพันธ์กับเพจของนักการเมืองเพจอื่น ๆ หรือถ้าหากเป็นเพจบริษัทเทคโนโลยี ก็จะมีความสัมพันธ์กับเพจที่มีลักษณะไปในทางเทคโนโลยีเหมือนกัน
2. Facebook page ที่เป็นบริษัทที่มีหลายสาขา มักจะมีเพจย่อยของแต่ละสาขา ตัวอย่างที่เห็นคือ บริษัท Triump ที่มีชื่อเพจที่เหมือนกันหลายเพจ ทำให้เห็นว่า ถ้าเป็นพวกบริษัทที่มีเพจ แล้วเป็นบริษัทใหญ่ ก็จะมีเพจย่อย ๆ ไว้กระจายไปตามสายงานของบริษัทหรือส่วนนั้น ๆ
3. Facebook page อาจจะไม่จำเป็นต้องมี category ที่เหมือนกัน แต่อาจเกี่ยวข้องกันในการทำงาน การเป็นผู้ร่วมงานกันได้ จะเห็นได้จากตัวอย่างของเพจ series เรื่อง DareDevil ที่อยู่ในหมวด TV Show แต่มีความเกี่ยวข้องกับเพจ Wizard World ที่เกี่ยวกับวัฒนธรรม Pop Culture ที่เกี่ยวข้องกับ Comics ภาพยนตร์ Sci-fi และเกม ที่สามารถเกี่ยวข้องกับรายการโทรทัศน์เรื่องนี้ได้

เมื่อเป็นดังนี้แล้ว การจัดตั้งเพจของ Facebook เราจึงควรมีปฏิสัมพันธ์กับเพจอื่นที่มีความใกล้เคียง หรือเกี่ยวข้อง เพื่อให้เราจะได้สามารถร่วมงานกัน เพื่อให้งานออกมามีความน่าสนใจ และทำให้มียอดผู้คนเข้ามากันได้อย่างแน่นอน

เอกสารอ้างอิง (References)

[1] B. Rozemberczki, C. Allen, and R. Sarkar. "Facebook Large Page-Page Network." Accessed April 4, 2021. https://snap.stanford.edu/data/facebook-large-page-page-network.html.

[2] Sobolevsky, Stanislav, and Riccardo Campari. "General Optimization Technique for High-Quality Community Detection in Complex Networks," January 1, 2014. http://senseable.mit.edu/papers/pdf/20141001_Sobolevsky_etal_GeneralOptimization_PhysicalReview.pdf. K. Elissa, "Title of paper if known," unpublished.

[3] Allman, Andrew. "Community Detection." Accessed April 28, 2021. https://www.sciencedirect.com/topics/computer-science/community-detection.Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[4] Xin, Zhige, Chun-ming Lai, and Jon William Chapman. "Multi-View Community Detection in Facebook Public Pages," January 7, 2019. https://www.researchgate.net/publication/334382861_Multi-View_Community_Detection_in_Facebook_Public_Pages.

[5] Jain, Pranita, and Deepak Singh Tomar. "Review of Community Detection over Social Media: Graph Prospective," n.d., 600–601.

[6] Rozario, Victor Stany, AZM Ehtesham Chowdhury, and Muhammad Sarwar Jahan Morshed. "Community Detection in Social Network Using Temporal Data," January 4, 2019. https://arxiv.org/ftp/arxiv/papers/1904/1904.05291.pdf.

[7] Ladd, John R., Jessica Otis, Christopher N. Warren, and Scott Weingart. "Exploring and Analyzing Network Data with Python." Accessed April 29, 2021. https://programminghistorian.org/en/lessons/exploring-and-analyzing-network-data-with-python.