

© 2019 Qinglin Chen

A LARGE-SCALE STUDY OF FASHION INFLUENCERS ON TWITTER

BY

QINGLIN CHEN

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Adviser:

Assistant Professor Ranjitha Kumar

ABSTRACT

The rise of social media has changed the nature of the fashion industry. Influence is no longer concentrated in the hands of an elite few: social networks distribute power across a broad set of tastemakers; trends are driven bottom-up and top-down; and designers, retailers, and consumers are regularly inundated with new styles and looks.

This thesis presents a large-scale study of fashion influencers on Twitter and proposes a fashion graph visualization dashboard to explore the social interactions between these Twitter accounts. Leveraging a dataset of 11.5k Twitter fashion accounts, a content-based classifier was trained to predict which accounts are fashion-centric. With the classifier, I identified more than 300k fashion-related accounts through a snowball crawling and then defined a stable group of 1000 influencers as the fashion core. I further human-labeled these influencers' Twitter accounts and mine their recent tweets. Finally, I built a fashion graph visualization dashboard that allows users to visualize the interactions and relationships between *individuals*, *brands*, and *media* influencers.

To my parents, for their love and support.

ACKNOWLEDGMENTS

First, I would like to express my sincere gratitude to my advisor, Professor Ranjitha Kumar for her patient guidance, enthusiastic encouragement and constructive critiques during the planning and development of this research work. She has been my research advisor for almost two years and has been guiding me on how to conduct research and come up with HCI story insights on fashion related topics. Without her trust and support, I wouldn't be able to make this thesis possible.

Second, I express my very great appreciation to Professor Hari Sundaram, Associate Professor at UIUC for his valuable and helpful guidance and suggestions on social network analysis. He guided me on how to build a system to understand collective behaviors in large Twitter networks. I would also like to thank Jinda Han, a Ph.D. student at UIUC, who led the research project and contributed to the PageRank results the thesis utilized to identify the insights. I was able to learn essential research skills and natural language processing skills from him. My grateful thanks are also extended to Xilun Jin, an undergraduate student at UIUC for his help in building the visualization dashboard interface.

In addition, I would like to thank the Computer Science Department at the University of Illinois at Urbana-Champaign for providing me great opportunities and resources during both undergraduate and graduate studies. I would like to extend my thanks to all the group members in the Data Driven Design Group led by Professor Ranjitha Kumar. They offered me incredibly helpful suggestions in the research group.

TABLE OF CONTENTS

CHAPTER 1	INTRODUCTION	1
CHAPTER 2	RELATED WORK	2
CHAPTER 3	CREATING THE FASHION GRAPH DATABASE	3
3.1	Categorizing Influencers	4
3.2	Mining Tweets, Mentions, and Retweets	5
CHAPTER 4	EXPLORING THE SOCIAL INTERACTIONS	7
4.1	Backend	8
4.2	Fashion Graph Visualization Dashboard	8
CHAPTER 5	ANALYZING TWITTER’S FASHION INFLUENCERS	11
5.1	Who Are They?	11
5.2	How Do They Rank?	14
5.3	How Do They Interact?	15
CHAPTER 6	DISCUSSION	23
6.1	The Size of the Fashion Subgraph	23
CHAPTER 7	FUTURE WORK	27
7.1	Mapping to Other Social Networks	27
7.2	Identifying New Accounts	27
7.3	Detecting the Fashion Trends or Events	28
REFERENCES	29
APPENDIX A: FASHION CATEGORIES	32

CHAPTER 1: INTRODUCTION

The rise of social media has affected the creation and diffusion of fashion trends. These new networks distribute influence across a broader set of tastemakers, begetting rapidly changing, bottom-up trends [1]. Fashion individuals, brands, media, and retailers realize it helpful to explore the relationships between them and find it challenging to predict emerging trends [2] and identify the trendsetters who have reach over their target audience. Knowing who these fashion influencers help people know who to reach out to if they want to expand their own fashion influence or reach new people in the industry [3].

This thesis presents a graph visualized system that helps practitioners identify and understand the fashion influencers in the social network. In the meantime, it also provides a way of visualizing and analyzing how fashion influencers are connected, how they interact, and who the important influencers are in a network. The system consists of a social graph analysis component and a network visualization dashboard.

I used a content-based classifier with 90% accuracy to predict whether the accounts are fashion-centric. I used this classifier to snowball sample more than 9.2 million Twitter accounts and discovered over 300,00 fashion-related accounts. By running PageRank over the follow-based graph, I identified a set of 1000 fashion. We further categorize these influencers and mine their recent tweets.

Finally, we build a graph visualization dashboard that allows users to visualize the interactions between PEOPLE, BRANDS, and MEDIA. The goal of the system is to help people in the fashion industry to identify fashion influencers and explore compelling stories on Twitter. By using our visualization dashboard, I am able to come up with some insights from the following aspects, *who are they*, *how do they rank*, and *how do they interact*. We find that media accounts such as magazines only mention brands and individuals; similarly, brand accounts only mention media and individuals in their tweets in Figure 5.6. The only time brands mention other brands is when there they have collaborations. We believe that people in the fashion industry can leverage these types of insights to do competitive analysis and identify future trends.


CHAPTER 2: RELATED WORK


Social networks have become more prevalent in the past decade. Researchers have studied a variety of online social communities such as Twitter [4, 5], Pinterest [6], Tumblr [7], Github [8], etc. and applied different methods to study trends and influence on fashion [9]. For example, Chang et al. [10] explored how gender and topics affect users’ following relationships on Pinterest. Gilbert et al. developed a statistical model that predicts tie strength based on social media data [11], and Ward analyzed employees’ social influence at his company by comparing their hierarchical position in the organization to their rank in the network [12]. The most recent paper [13] applied a convolution-deconvolution based object detector to identify fashion trends on social media platforms.

Popularity in social networks can be measured with just quantitative factors, such as the number of *followers* or *retweets*. However, to rank influence, we also need to consider quality. For example, according to PageRank [14] an account that is followed by many fashion influencers in the network, has a higher rank than a celebrity who has many followers, most of whom are from the ordinary population. PageRank has been extended in many areas and applied to different contexts [5, 15, 16]. In this thesis, I will use the basic formulation of PageRank results to evaluate the relative importance of fashion accounts.

Besides, Zhao and Min talked about how large-scale datasets can reshape the fashion industry [17], and used dynamic network visualizations of Paris Fashion Week to analyze the graph of Hashtags and Keywords. To estimate the size of the subgraph, researchers usually use sampling methods to estimate the size of a network before identifying influencers and trends. Gjoka et al. [18] obtained unbiased estimators using Metropolis-Hastings Random-Walk sampling (MHRW) and Re-Weighted Random Walk sampling (RWRW) to produce a representative sample of Facebook users. *@MarpolePlan* is one of Twitter accounts have been set up by the city government to promote civic engagement in various neighborhoods in Vancouver. Researchers created a social network diagram of anyone who has mentioned “Marpole” in a tweet and has visualized how those people are interrelated [3], in order to figure out how this kind of account can be used to improve reach and engagement.

CHAPTER 3: CREATING THE FASHION GRAPH DATABASE

#1 vogue magazine 

 Name:

Location: Following: 527

Follower: 13.55M

Description: The official twitter page of Vogue Magazine.

PEOPLE

BRANDS

RETAIL

MEDIA

OTHER

Figure 3.1: The figure shows the labeling interface that users accessed during when labeling.

After we identified the Fashion core accounts, we categorized each fashion account with a certain label such as *bloggers*, *celebrities*, *brands* and *magazines*. Then I created a document database using MongoDB that stores each account as a model document with detailed information such as follower/following relationship and also stores each tweet as a model document with the mention/retweet relationship based on the contents of tweets.

3.1 CATEGORIZING INFLUENCERS

First, to understand the category of each account, a labeling system was built in the forms of a labeling web interface in Figure 3.1 to facilitate the labeling process. Leveraging Human-in-the-loop, I decided to manually label until I reached 5000 valid fashion accounts. 22 preset labels was obtained from an editor in Vogue Magazine and were mapped into 5 main categories (*People*, *Brands*, *Retail,Media* and *Other*. Details see Appendix). *Other* labels are meant for invalid accounts, such as accounts that have been inactive for more than one year, nonfashion related, or not in the designed category lists. Users can pick one or multiple labels for an account, or input a new label in *Other* if they think none of the labels can match the account. To better monitor and analyze the fashion accounts labels, I also generated label statistics along with the labeling process. The label distribution is visualized in a d3pie [19] graph.

```
_id: 1092623343717621760
user_id: "136361303"
created_at: "2019-02-05 03:17:59"
text: "Watch what happened when we asked @giseleofficial #73Questions- joined..."
relation: "mention"
relation_node: Array
  0: "giseleofficial"
hashtags: Array
  0: "#73questions"
```

Figure 3.2: A snippet of the *tweet* model document stored in MongoDB.

There are two difficulties encountered during labeling. One problem that occurred while labeling was that people had different views on whether the account is fashion-related or not. Some were very strict about labeling the account as fashion-related. Thus, I had to relook at some accounts that were labeled as non-fashion and see if they were non-fashion. Another difficulty was that the profile text was sometimes misleading. Many accounts introduced themselves as a fashion account in their bios, but many of them were just comprehensive accounts that had minimal fashion-related content.

Then, I created a document database using MongoDB [20]. Each account is stored as a model document that with detailed information such as user profile information *follower/following* relationships and updated accurate categories as displayed in Figure 3.3.

3.2 MINING TWEETS, MENTIONS, AND RETWEETS

There are two types of social interaction through tweets. People can tweet by mentioning or retweeting others' tweets to deliver their social influence. We say if a user retweets a post, then he/she is influenced by the original author of the post. If a user mentions someone on a tweet, then this user is trying to expand his/her influence to the user mentioned.

To explore the interesting tweet relationships between the fashion accounts, I crawled the most recent 3200 tweets of each account in our top 5000 accounts list multiple times to get the tweets in over a year¹. There are some accounts, such as *Vogue Magazine* that tweet very frequently. To get complete tweets of the whole year, I kept crawling tweets of these accounts periodically. Then, I filtered the tweets into a timeframe from January 1st, 2018, to February 1st, 2019. Using an iterative Natural Language Processing approach, I identified the type of each tweet and the relating accounts in the tweet by detecting the keywords like MENTION, RETWEET and their location in the text. For example, a tweet will be recognized as a RETWEET if an "RT" appears at the beginning of the sentence and is followed by a Twitter screen name. A tweet will be recognized as a MENTION if it is not a RETWEET and there a @ symbol in the text followed by a Twitter screen name.

Then in MongoDB [20], each account is stored as a model document with detailed information such as extended tweets and MENTION/RETWEET relationship based on the extended tweets as displayed in Figure 3.2 .

¹3200 is the maximum number of tweets that we can crawl using Tweet API

```
> {
  "_id": "136361303"
  "twitter_user_id": "136361303"
  "twitter_screenname": "voguemagazine"
  "twitter_name": "Vogue Magazine"
  "ranking": 1
  > "ranking_history": Array
  > "labels": Object
  > "twitter_profile": Object
    "id_str": "136361303"
    "location": "New York, NY"
    "profile_location": null
    "description": "The official twitter page of Vogue Magazine."
    "url": "https://t.co/vaoj57vMYU"
    > "entities": Object
      "protected": false
      "followers_count": 13596800
      "friends_count": 526
      "listed_count": 23591
      "created_at": "Fri Apr 23 18:33:32 +0000 2010"
      "favourites_count": 5800
      "utc_offset": null
      "time_zone": null
      "geo_enabled": true
      "verified": true
      "statuses_count": 124240
      "lang": "en"
    > "followinglist": Array
      "follow_ranking": 1
      "mention_ranking": 1
      "retweet_ranking": 1
```

Figure 3.3: A snippet of the *user* model document stored in MongoDB.

CHAPTER 4: EXPLORING THE SOCIAL INTERACTIONS

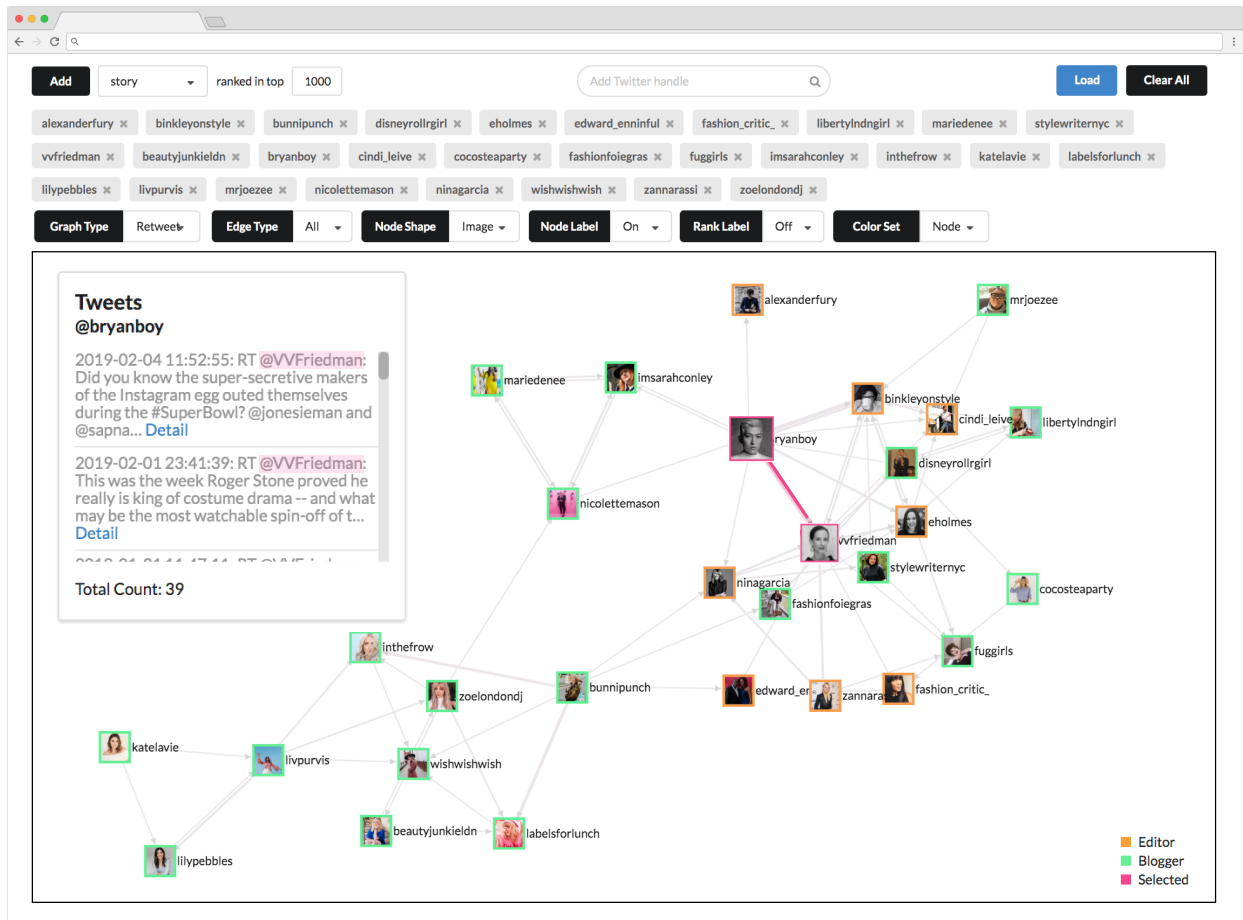


Figure 4.1: The figure shows the overall design of the fashion graph visualization dashboard. The graph inside the dashboard is the retweet relation of some top-ranked fashion bloggers and editors. Some bloggers are intimately connected with editors while others bridge the connection in the retweet graph between editors and bloggers. The tweet snippet shows an example of *Bryan Grey Yambao* (@bryanboy) often retweeting *Vanessa Friedman* (@VVFriedman).

To better understand the top 1000 fashion core sub-graph, we built a fashion graph web interface to facilitate the visualization of the FOLLOW, MENTION and RETWEET among the accounts. This web app is available for public use at <http://www.fashioninfluencerexplorer.com>. The social interactions mainly focus on three categories which are *brands*, *individuals*, and *media*. *Individuals* includes a subset of *models*, *celebrities*, *bloggers*, *editors*, *designers* and *photographers*. *Media* consists a subset of *magazine*, *e-zine* and *blogs*.

4.1 BACKEND

The backend of the Fashion Influencers application consists of a Flask server and a MongoDB database [20]. Using a python script, I iterated through the entirety of the raw data of the twitter accounts and tweets dataset, and then inserted the data into MongoDB. We indexed certain fields, such as *twitter_id* and *follow_ranking*, to allow fast searching over the huge dataset. I implemented some custom endpoints for some queries. For example, the endpoints allow user to flexibly set the rank limit to get the top ranked accounts in each category in terms of the FOLLOW PageRank.

4.2 FASHION GRAPH VISUALIZATION DASHBOARD

The frontend of the Fashion Influencers application is a single page application (SPA) written in React.js [21]. The interface works as a visualization dashboard that needs the user to input a single twitter account or a set of twitter accounts. The interface allows the user to search the accounts in our top 1000 database and then add the accounts. Some default sets of the accounts grouped by different categories are also provided. After accounts are selected, the system will generate a network graph of these accounts. This network graph is constructed from the force-directed graph [22] with weighted nodes and edges added. The network graph is focused on three types of relations including *follow*, *mention* and *retweet* among the twitter accounts. In addition, users are also able to adjust the appearance and information volume of the graph in their needs. Users can turn on or off the ranking and screen name labels of the nodes, and choose to use curved or straight edges between the nodes as well. Figure dashboard provides an example of how dashboard shows the result when certain settings are put. In Figure 4.3, I provide the detailed descriptions for the UI components designed in the interface.

Fashion on Twitter

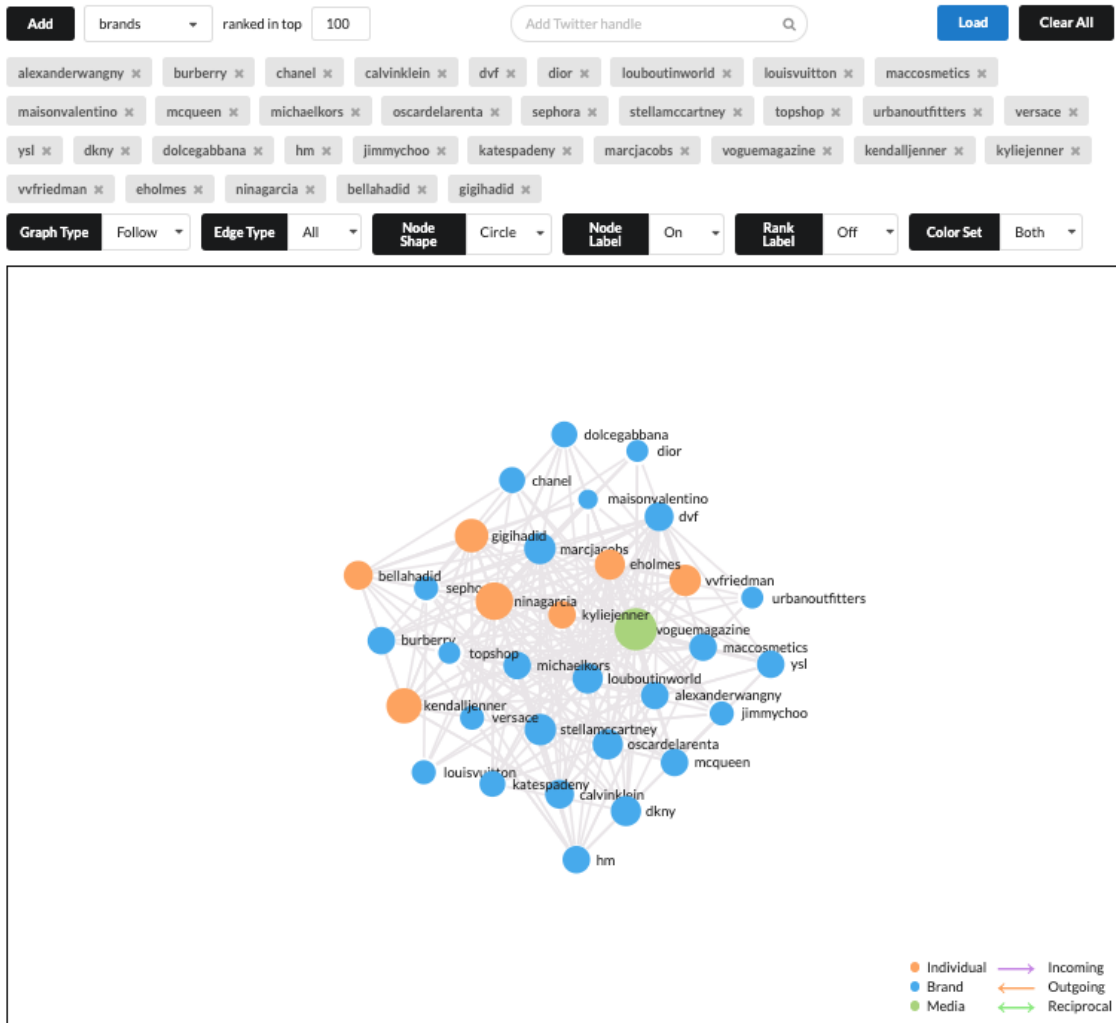


Figure 4.2: The figure shows an example of the fashion graph visualization dashboard when the *Node Label* is on, only *Node Label* has a *Color Set*, and *Graph Type* is FOLLOW

UI COMPONENT	FEATURE	DESCRIPTION
Search Bar/Add	Input	Use can input a set of twitter accounts from the Add button or a single account by searching its twitter handle
Node	Size	The size of the node represents the Log of sum of the inward degree and a default size. The default size is set for the node with 0 inward degree
	Color	The color of the node represents its category
	Shape	The shape of the node can either be a circle or the twitter profiles image of the accounts
Edge	Inward Edge	If node A has a inward edge from B, then A is followed/mentioned/retweeted by B
	Outward Edge	If node A has a outward edge from B, then A follows/mentions/retweets B
	Double-sided Edge	If there is a Double-sided edge between A and B, then A and B have mutual relationship of follow/mention/retweet
	Color	The color of the edge represents the type of the edge. The color of the edge has a default light grey when the node or the edge is unselected
	Thickness	The thickness of the edge represents the number of the tweets related to that type of edge
	Select the edge	When the edge is selected, the number and text of tweets related in the edge will show in the Tweet Info box under the graph. Each tweet has a link to its original post on Twitter

Figure 4.3: This table explains a complete list of the interactions and the features the interface supported

CHAPTER 5: ANALYZING TWITTER'S FASHION INFLUENCERS

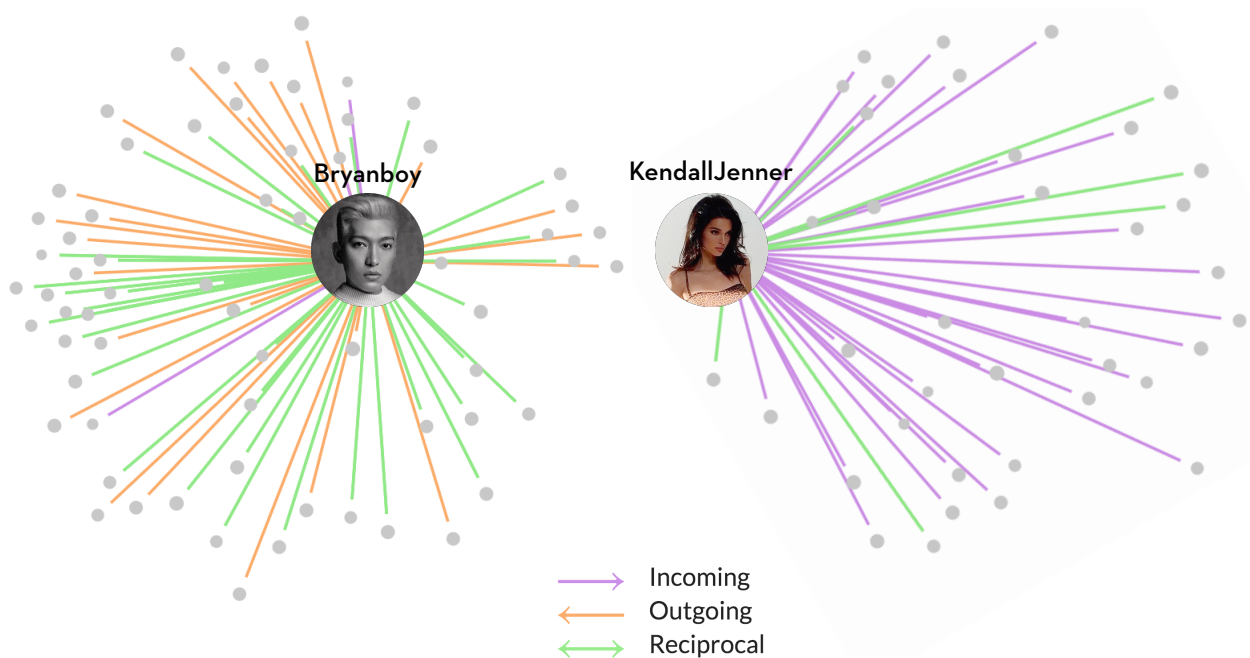


Figure 5.1: The example is illustrating FOLLOW patterns among bloggers and models. In general, bloggers tend to have more reciprocity FOLLOW relation, while celebrities like models tend only to have one-way relation mostly, are being followed.

To explore the potential stories behind the fashion graph constructed through the interface, I selected *brands*, *individuals* and *media* accounts ranked in the top 1K according to their FOLLOW rankings. Through analyzing the FOLLOW, MENTION and RETWEET relations among these accounts, I identified some interesting stories that are difficult to parse from the raw data.

5.1 WHO ARE THEY?

According to Figure 5.3, there are 376 (37.6%) media accounts, 225 (22.5%) brands accounts and 226 (22.6%) individual accounts ranked in the top 1000. There are 1,132 (22.6%) media accounts and 740 (14.8%) brands listed in the top 5K.

The influence of an account was measured by computing the total number of incoming FOLLOW links (also known as *indegree centrality*) divided by the total number of nodes in the graph. Together, the top 10 accounts under PageRank have 49.7% influence, which means that nearly half of the fashion subgraph I identified (301K accounts) follow these ten

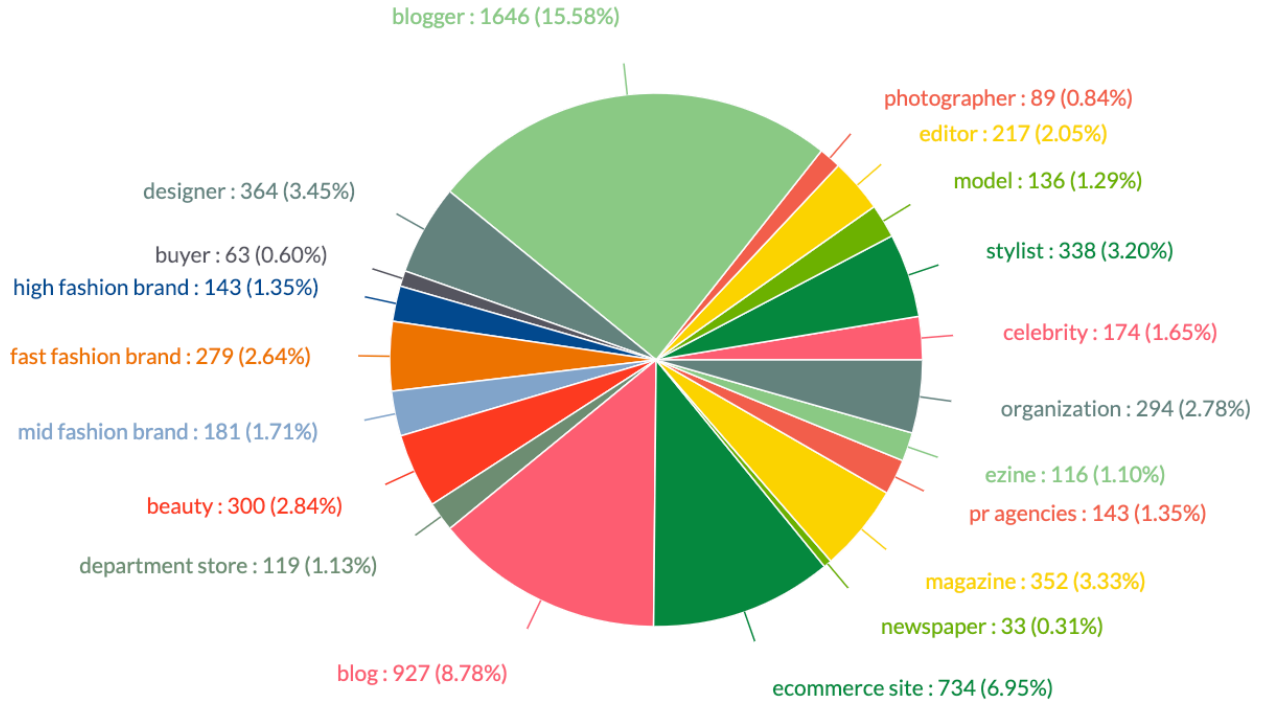


Figure 5.2: The example is illustrating the top 5000 accounts stats after the human-in-the-loop labeling.

accounts; the top 100 accounts have 70.8% influence and top 1K accounts which I identified as the fashion core, have 90.9% influence.

Similarly, the influence of the top 10 accounts per category was also measured ???. The giant *vogue* magazine have 101.8K (33.8%) followers in our 301K Fashion graph, and follows by *wwd* with 70.5k (23.4%), *ELLE* magazine with 85.1k (28.3%), *British Vogue* with 78.2K (26.0%), and Fashionista.com with 76.8K (25.5%). The top 10 magazine and newspaper accounts (i.e., print media) have the most influence (56.4%), followed by the fashion brands (51.0%), models and celebrities (46.0%), blogs and e-zines (45.5%), and bloggers (27.7%). While there are more blogs and e-zines represented in the top 1K accounts, the top 10 magazines and newspapers who still produce print media command the most influence in the fashion graph: *Vogue Magazine* (@voguemagazine), *ELLE Magazine* (@ElLEmagazine), *British Vogue* (@BritishVogue), *InStyle*, *VogueRunway* (@VogueRunway), *W magazine* (@wmag), *The Cut* (@The Cut), *Harper's Bazaar* (@harpersbazaarus), *Times Fashion* (@TimesFashion) and *Marie Claire* (@marieclaire). For Models and Celebrities, their influence is somehow higher than Bloggers, because most of them have much more social effect in their domain. Such as *victoriabeckham* (Ranked 23) is not only the famous singer, successful business-women, but also the wife of sports star David Beckham's wife. Also, *KimKardashian*

(Ranked 35), *KendallJenner* (Ranked 113), and *KylieJenner* (Ranked 167) are all American media personality and model from same family, which all create high influence.

Magazines/ Newspapers	Rank	Fashion Brands	Rank	Models/ Celebrities	Rank	Blogs/Ezines	Rank	Bloggers	Rank
vogue magazine	1	Topshop	26	victoriabeckham	23	wwd	2	RachelZoe	20
ELLE magazine	3	Burberry	27	KimKardashian	35	Fashionista.com	5	susiebubble	37
BritishVogue	4	YSL	28	ladygaga	41	WhoWhatWear	9	bryanboy	70
InStyle	6	CHANEL	30	rihanna	51	Refinery29	10	mrjoezee	73
VogueRunway	7	marcjacobs	33	ninagarcia	69	BoF	15	garancedore	99
wmag	8	StellaMcCartney	38	taylorswift13	79	NYTFashion	18	JacquelineRLine	108
TheCut	11	MichaelKors	42	DerekBlasberg	101	ManRepeller	50	LaurenConrad	122
harpersbazaarus	12	hm	45	KendallJenner	113	marieclaireuk	75	byEmily	150
TimesFashion	13	Dior	47	LaurenConrad	122	POPSUGARFashion	76	fashionioiegras	153
marieclaire	14	LouboutinWorld	52	OliviaPalermo	123	AOLLifestyle	77	chictopia	158
56.4% Influence		51.0% Influence		46.0% Influence		45.5% Influence		27.7% Influence	

Figure 5.3: The example is illustrating the top 5000 accounts stats after the human-in-the-loop labeling.

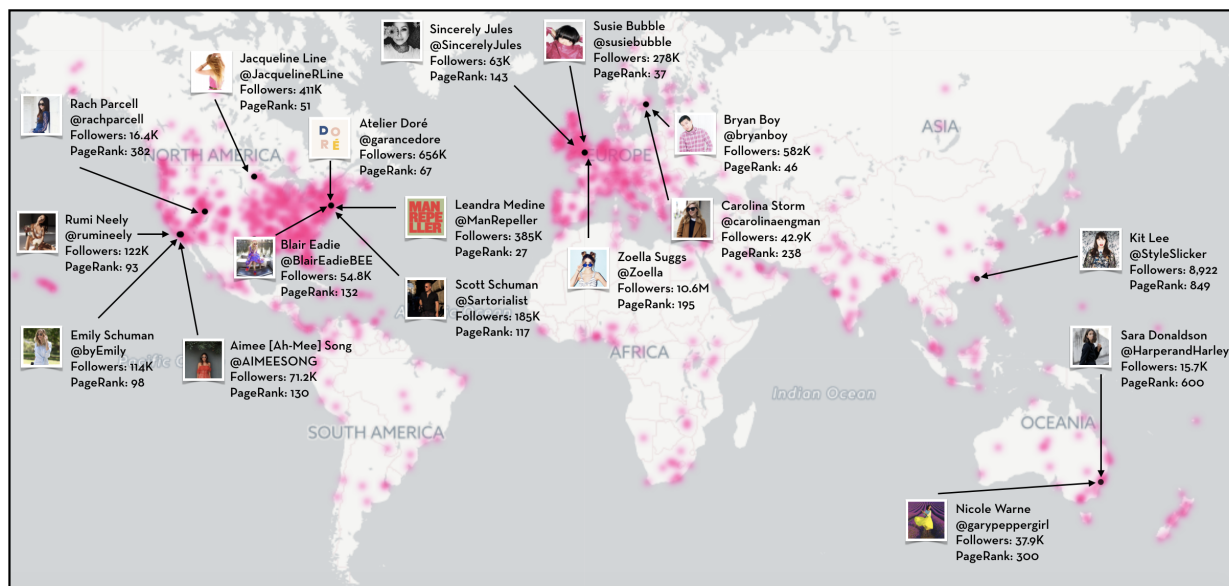


Figure 5.4: The figure is showing the FOLLOW ranking and influences of top rated accounts.

To study the geographical diffusion of fashion bloggers, the Geocoder library [23] and Google geocoding service [24] were used to retrieve the coordinates of the accounts using the location text shown on the accounts' profiles. Based on Figure 5.4, while a large portion of bloggers is concentrated in fashion hubs like New York, London, and Paris, others are also widely distributed: some bloggers with extremely high influence live in states such as

North Dakota *Jacqueline Line* and Utah *Rach Parcell*, which are traditionally not considered fashion-centric in the US.

5.2 HOW DO THEY RANK?

From the top 30 ranking, I found out that the PageRank results for FOLLOW and RETWEET are similar. The media accounts such as magazines, e-zines, and blogs have relatively higher FOLLOW PageRank.

In the RETWEET ranking, I observe that the posts from magazines and e-zines are retweeted the most, mainly by brands and individuals. In MENTION ranking, *brand* accounts are ranked relatively higher.

By comparing the node sizes in the MENTION graph of the accounts in the top 1K, the most mentioned accounts are celebrities such as *Lady Gaga (ladygaga)*, *Kim Kardashian West (@Kimkardashian)* and *Rihanna(@rihanna)*.

Besides, I also compared MENTION and RETWEET rankings of individuals such as editors and models. Editors usually have higher RETWEET rankings while models have higher MENTION rankings. For example, *Nina Garcia(@ninagarcia)*, editor-in-chief of *ElleMagazine*, ranks 31 in the RETWEET ranking and 1,237 in the MENTION ranking; *VVFriedman*, Fashion Director and Chief Fashion Critic at *The New York Times*, ranks 30 in the RETWEET ranking and 1,237 in the MENTION ranking; American models *Kendall Jenner (@KendallJenner)* and *Bella Hadid (@bellahadid)* respectively rank 997 and 3,291 in RETWEET ranking but 167 and 169 in MENTION ranking. The difference of RETWEET and MENTION rankings can be explained to the fact that editors have provided original fashion content and have been retweeted much more than famous models. For example, *VVFriedman* likes expressing her fashion insights in the form of articles and then sharing her publications in *nytimes.com* on Twitter.

In comparison, fashion models prefer to tweet about their daily lives as celebrities. They share the products they use, and companies take their endorsement as an advertisement. They also show their friendships with other stars. Although people do not retweet celebrities' posts as much as editors' food for thought, the public loves paying attention to and mentioning these superstars.

Besides, we came up another story that the RETWEET and MENTION rankings of some accounts are significantly lower than the follow rankings. For example, *Eva Chen (@evachen212)*, the director of fashion partnerships at *Instagram*, ranks 95 in FOLLOW, 889 in MENTION and 2,578 in RETWEET. From the Figure 5.5, the FOLLOW graph has the most density MENTION graph is much sparser and there is no edges associated with the node in the RETWEET graph.

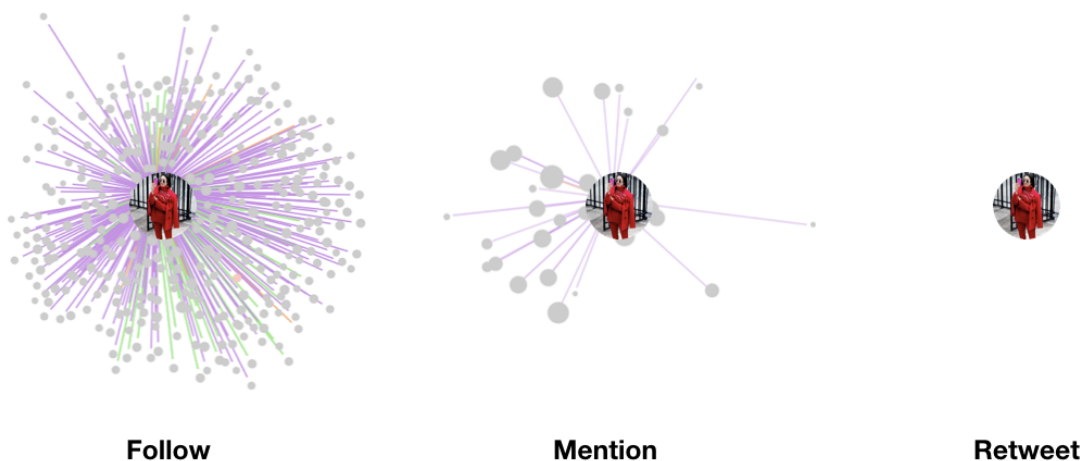


Figure 5.5: An example is illustrating that the decrease of density from the FOLLOW to MENTION and RETWEET relation of *evachen212*.

Since she is famous and socially active, she has established an influencing FOLLOW relationship on Twitter. However, she switched her main social media platform to Instagram. In the recent years, she mainly uses Instagram and has been inactive on Twitter, impacting her data interactions and consequentially giving her lower RETWEET and MENTION rankings.

5.3 HOW DO THEY INTERACT?

By comparing *follow*, *mention* and *retweet* graphs among the accounts ranked in the top 1k. We dug out some interesting patterns on how accounts in each category interact with others.

5.3.1 How they Follow

To understand how accounts in the fashion subgraph connect, reciprocity was measured in the *follow* relation network (i.e., how many connections in the graph are two-way edges). Only 7.4% of the 22,644,014 edges in the graph are reciprocal. However, reciprocity is more prevalent among top-ranked accounts. Reciprocity within the top 1k fashion core is 17.13%, and 46.19% within the top 100 accounts. This phenomenon is related to homophily and has been observed before: top influencers in the community follow each other and form social clusters.

Most media accounts, shown in 5.6, follow everyone to be in the know. Most of these

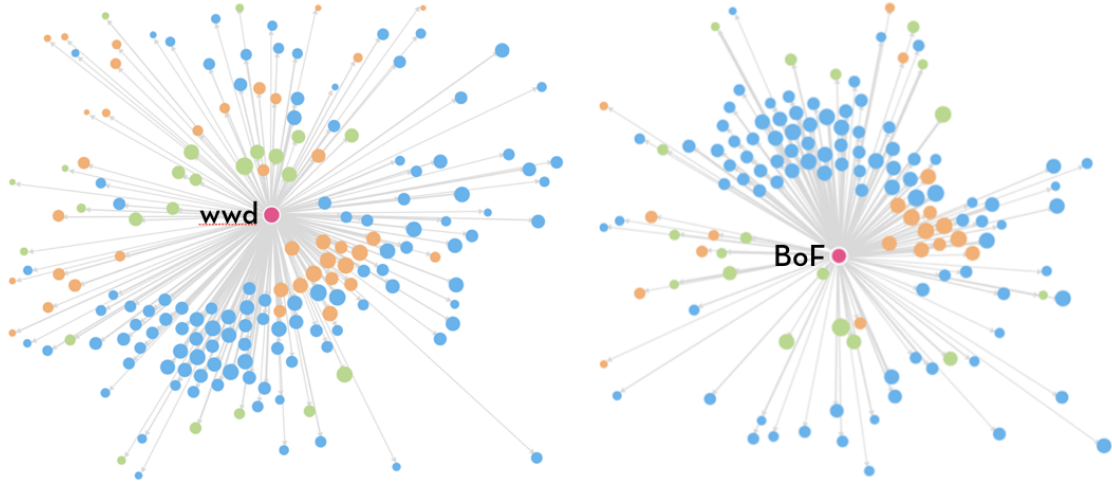


Figure 5.6: The example is illustrating unlike other fashion MEDIA accounts, fashion trade journals and business intelligence companies actually mentions other fashion MEDIA accounts.

media accounts will have mutual *follow* relations with other media accounts. However, there are some exceptional cases. For example, *voguemagainze* follows and mentions brands and people, but not other media accounts. We observe that *voguemagainze* mostly has reciprocity with accounts of brands and individuals and only follows seven media accounts among which two accounts, *Voguerunway* and *Britishvogue* are related to *voguemagainze* itself.

Brands mostly follow media and have reciprocity with many media and some individuals. Brands barely have any mutual relations with other brands. However, we also found some exceptions. Some brands are very arrogant when they come to follow. For instance, *Chanel* has 0 public followings. Some high fashion brands do not often follow other accounts but tend to follow their related branches in other countries. For example, *LouisVuitton* has only 13 followings including *LouisVuitton France*, *Louis Vuitton DE* and *FondationLV*.

Besides, individuals such as bloggers follow everyone to be in the know as well. For example, in figure 5.1 we see some bloggers such as *Rachel Zoe (@rachelzoe)* and *bryanboy* have reciprocity with most brands and media but follow most of the celebrities without being followed back. In comparison, celebrities such as *KendallJenner* mostly are being followed instead of following other accounts. From Figure 5.8, we can see that models like *Jenner* and *Hadid* sisters have a solid mutual relationships and forms a cluster. The elder sisters have a mutual follow relation with *voguemagazines*. Editors such as *VVFriedman*, *ninagaricia* forms a cluster and are connected to model clusters through *ninagaricia* as she follows *kendallJenner*



Figure 5.7: An example is illustrating that editors and celebrities may form their own clusters. And magazine and some certain editors and celebrities plays a role like a connector.

5.3.2 How They Mention and Retweet

In 5.9, we can see that the MEDIA mentions BRANDS and PEOPLE. Usually, media tweet about the stories they are writing about. However, in 5.7 fashion trade journals and business intelligence companies such as *The Business of Fashion (@BoF)* and *Women’s Wear Daily (@wwd)* mention all types of accounts because they aim to help fashion accounts connect.

Compared the FOLLOW and MENTION graphs of *vogue magazine* figure 5.9 and figure 5.6, it is very obvious to spot that *vogue magazine* has a denser relationship with top 1000 accounts in FOLLOW relation compared to MENTION relation. For the FOLLOW relation, *vogue magazine* mostly follows accounts in BRANDS and INDIVIDUALS and only follows 7 media accounts among which two accounts *Voguerunway* and *Britishvogue* are actually related to vogue magazine itself.

Brands usually mention or retweet media as an act of self-promotion to showcase an article that was written about them, featuring their new design features and collections. The only time brands mention other brands is when they have collaborations. For example, *swarovski* in figure 5.11, is one of the accounts that frequently mention other brands and is in mention, because other fashion clothing brands widely use the *swarovski* crystal as a type of material.

In contrast to the clothing brands, the mentioned pattern for self-promotion among beauty brands is more evident. The brands that are mentioned most by other brands tend to be beauty brands. In general, beauty brands are mentioned by many people including other beauty brands. For example, *Sephora* mentions some high fashion brands such as *CHANEL*

and *Dior* and is mentioned by many other cosmetics brands such as *MAC cosmetics* (@MACcosmetics). As we know, *Sephora* not only has its beauty brand, it also carries most of the beauty brands such as *MACcosmetics* and *Marc Jacobs* (@marcjacobs) in its websites and stores. Therefore, it is natural for these beauty brands to promote their influence by mentioning *Sephora* when they have restocking or new products in markets.

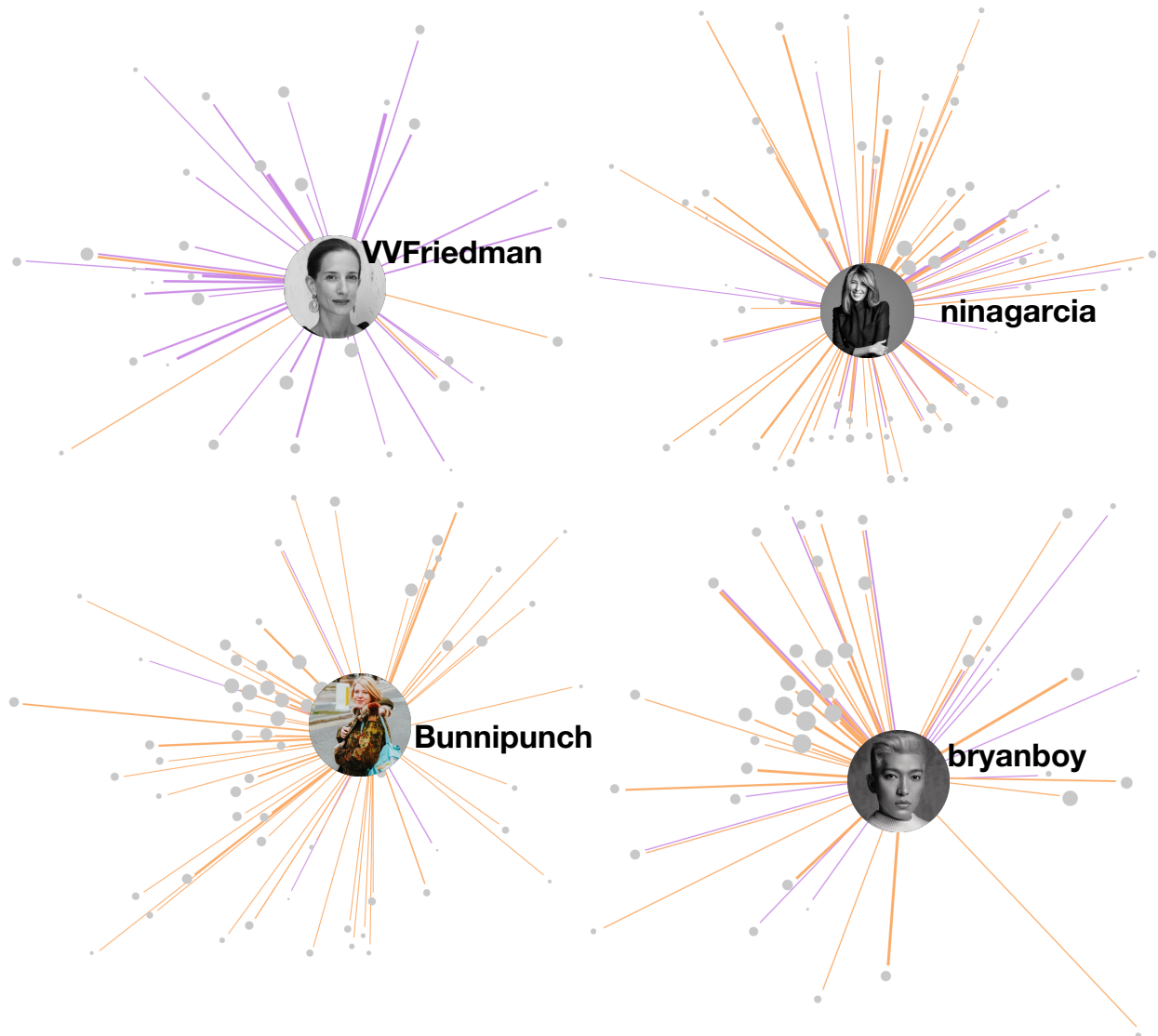


Figure 5.8: The example is illustrating RETWEET patterns among editors and bloggers. Purple edges represent being retweeted, and orange edges represent retweeting. In general, two editors *VVFriedman* and *ninagarcia* tend to be retweeted more while bloggers *textitLois Spencer-Tracey* (@Bunnipunch) and *bryanboy* tend to retweet more.

When brands are mentioning brands, sometimes it also prompts brands to retweet other brands. For example in figure 5.12, *swarovski* once mentioned *JASON WU* (@JasonWu) that

“Swarovski ambassador @Karlieklos wears a custom @JasonWu gown and #AtelierSwarovski jewelry as she walks the #CFDAAwards red carpet.” [25] *Jason Wu* also retweeted this, which established a win-win situation for them to mutually promote the products.

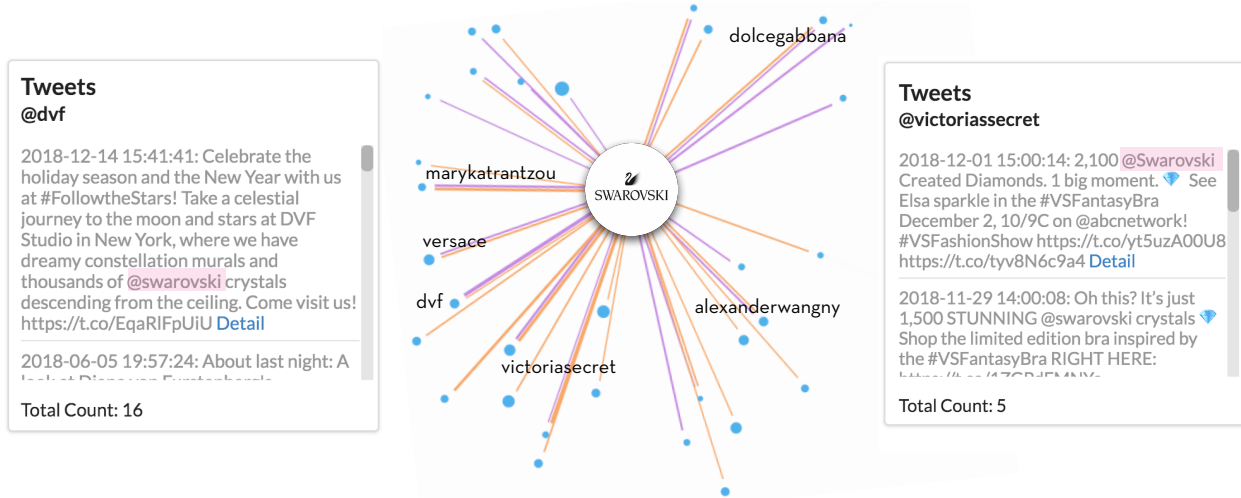


Figure 5.9: An example is illustrating that when brands mention other brands is when they have a collaboration. For example, *DVF* mentioned *Swarovski* (*@swarovski*) crystals used in its store and *Victoria’s Secret* (*@VictoriasSecret*) mentioned about sparkling *swarovski* crystals in its fantasy bras.

In addition, in general, people do not retweet much. When I look at the RETWEET graph among the top 1k accounts, it is much sparser compared to the FOLLOW and MENTION graphs. We selected around 30 fashion editors, and bloggers ranked in the top 1k and demonstrated a clear RETWEET relation among them in Figure 4.1. The RETWEET relation of editors and bloggers falls into two distinct clusters. Editors mostly retweet other editors. Some bloggers such as *bryanboy* have a very close and frequent RETWEET relation with editors, while others help bridge the RETWEET connection between bloggers and editors. These connecting bloggers such as *Bunnipunch* and *Victoria Magrath* (*@inthefrow*) are retweeting both editors and bloggers.

According to Figure 5.10, I discovered that editors are more likely to be retweeted in comparison to bloggers. In general, bloggers are retweeting more. Some editors such as *VVFriedman* are mostly being retweeted while others such as *Ninagarcia* retweet a lot. Similarly, some bloggers such as *Bunnipunch* are mostly retweeting while others such as *bryanboy* tend to be retweeted more in comparison.

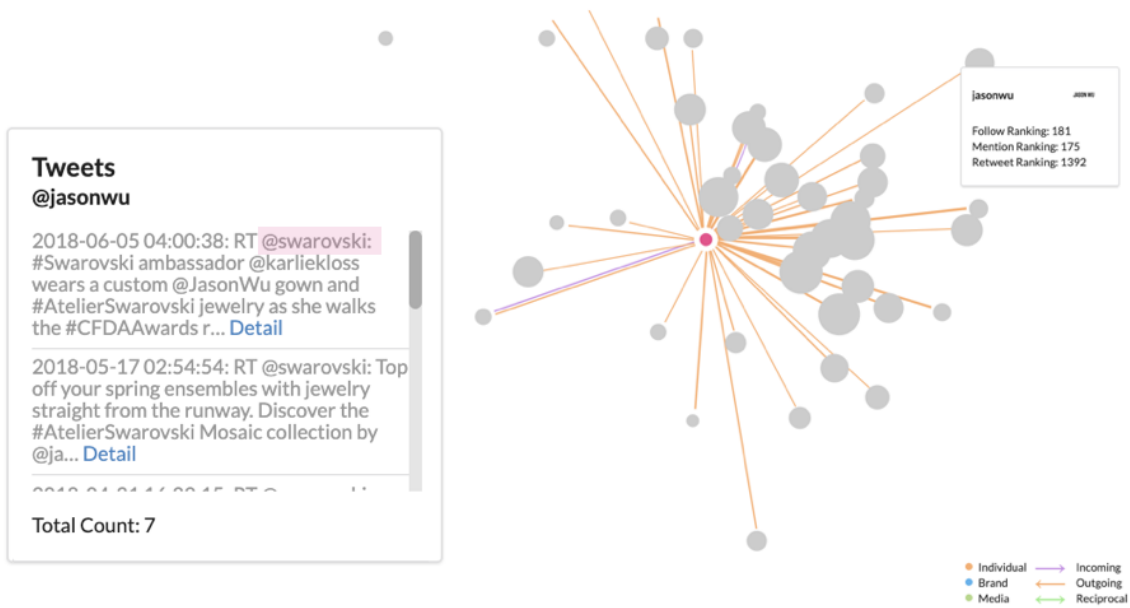


Figure 5.10: An example is illustrating *JasonWu* retweeted the tweet that *swarovski* mentioned about it.

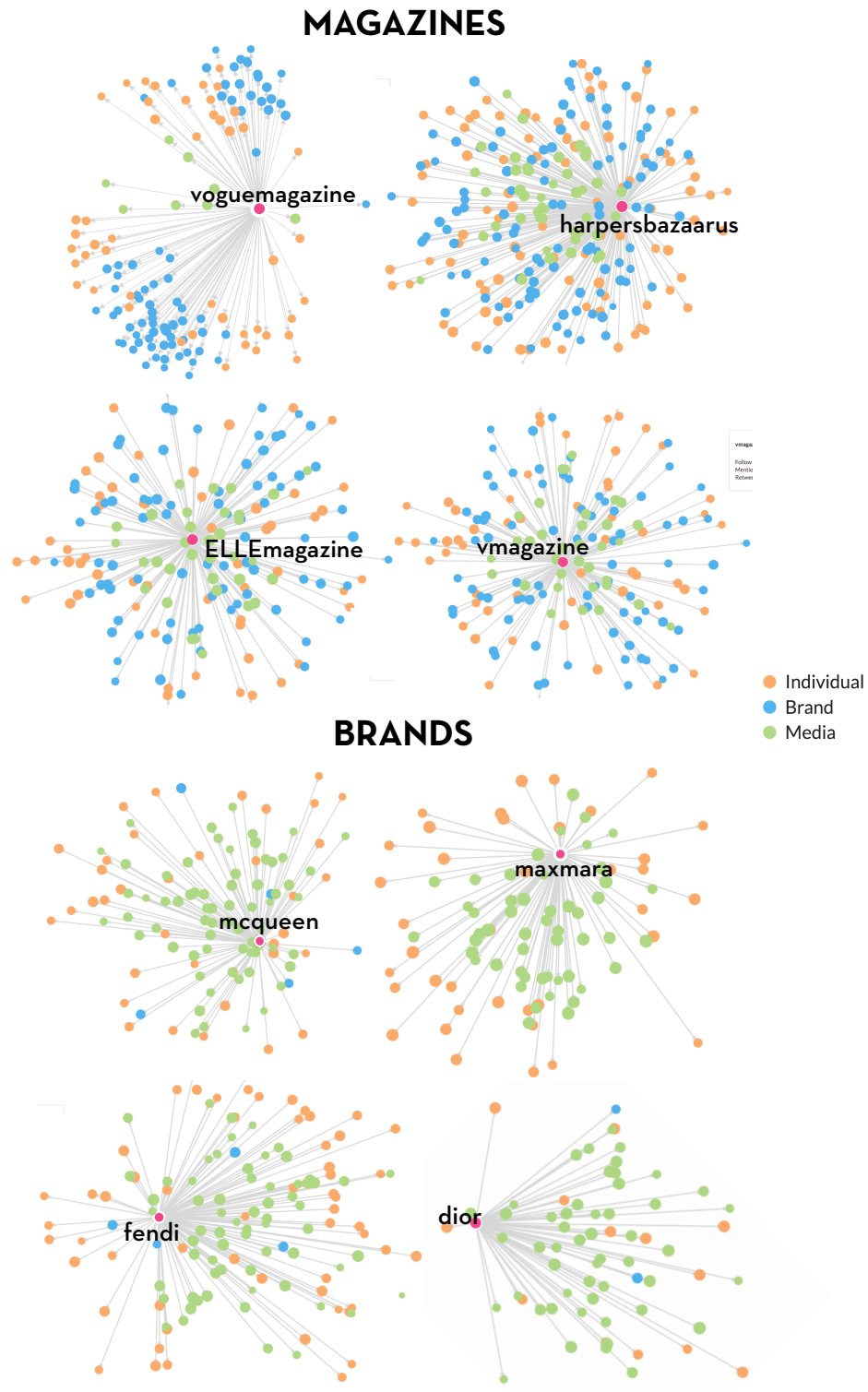


Figure 5.11: The example is illustrating FOLLOW patterns among magazines and brands. In general, magazines follow everyone in the know except *voguemagazine* while brands only follow media and people.

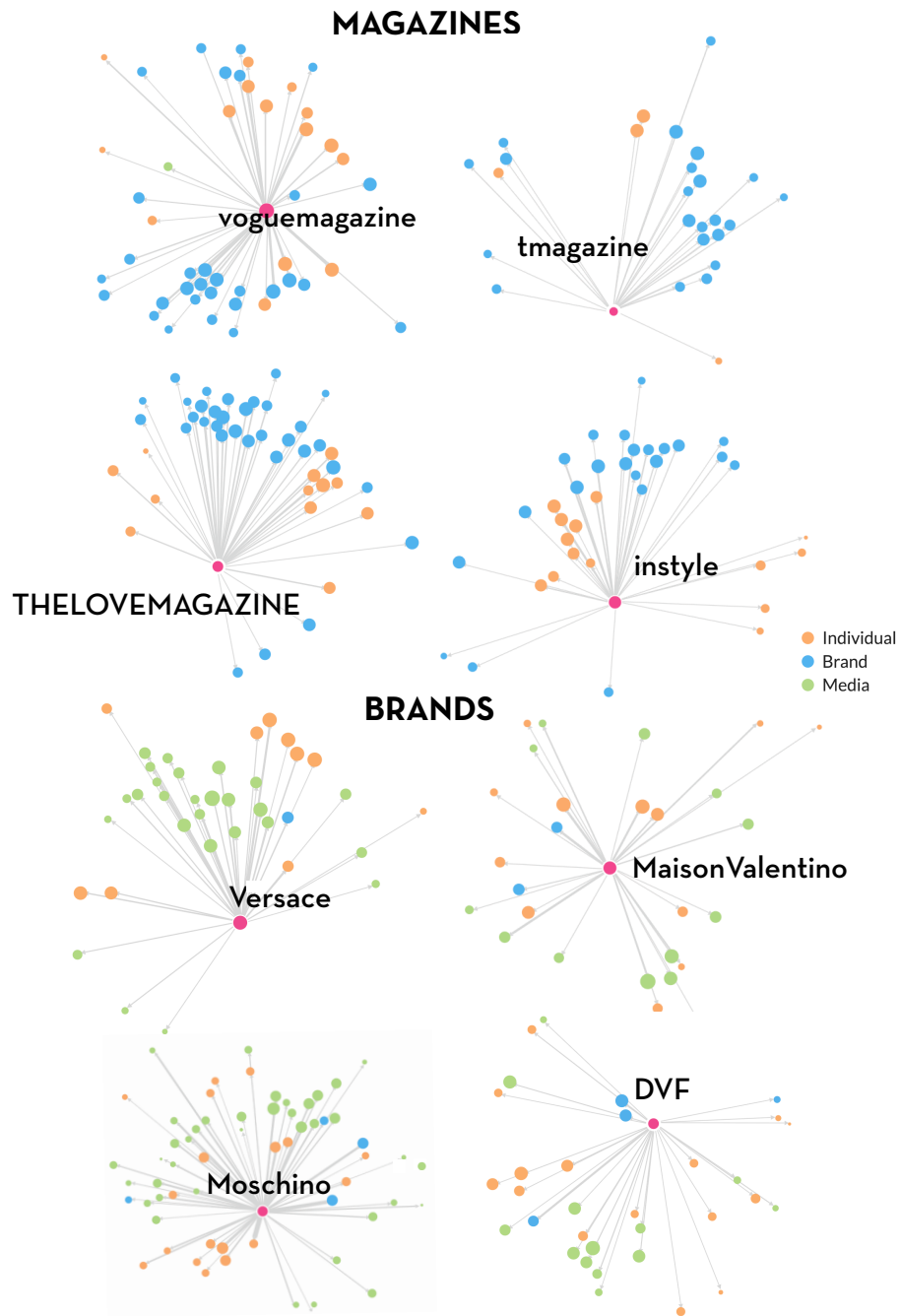


Figure 5.12: The example is illustrating MENTION patterns among magazines and brands. In general, MEDIA mentions BRANDS and PEOPLE, but not other MEDIA accounts while BRANDS mostly mention MEDIA and PEOPLE but not other BRANDS.

CHAPTER 6: DISCUSSION

This work is a first step towards studying and monitoring the diffusion of fashion influence on social media. Sampling Twitter is challenging, not only is it large, but also it is dynamic. Such as new accounts are being formed, accounts become inactive, and the accounts that a person chooses to change follow relationship over time.

Using the recovery rate, I obtained a converged fashion sub-graph, however, to have a better sense of the real size of the fashion sub-graph, I need to compute an unbiased sample of the Fashion sub-graph, where the probability that a node is labeled ‘Fashion’ is the same as in the underlying original Twitter graph. While there are several un-biased samplers, including Re-Weighted Random Walk (RWRW) and the Metropolis-Hastings Random Walk (MHRW)[18], I chose to work with RWRW since as [18] points out, RWRW has better convergence properties. While RWRW gives an unbiased sample, given the large size, and the fact that we must necessarily perform a finite crawl, it is possible that we miss some accounts.

Analyzing Twitter streaming API is an alternative approach. The Streaming API collects the most recent tweets from active Twitter users, which allows us to study the active user base. The main challenge with using the streaming API is that the sample is not unbiased, and susceptible to daily variation.

6.1 THE SIZE OF THE FASHION SUBGRAPH

I used a converged snowball sampling crawling strategy to crawl about 300k fashion-related accounts. However, even the dataset shows converge in the top 1k, the ranking still not fully converge. As such, I need to expand the graph size to stabilize the ranking of our graph. To start with an estimate the size of the fashion sub-graph, I compute an unbiased sample of the fashion sub-graph. By an unbiased sample, I mean a sub-graph where the probability that a node is labeled *fashion* is the same as in the underlying original Twitter graph.

While there are several un-biased samplers, including Re-Weighted Random Walk (RWRW) and the Metropolis-Hastings Random Walk (MHRW) [18], I chose to work with RWRW since as [18] points out, RWRW has better convergence properties. In Figure 6.1, I randomly picked 5 seeds for Random Walks, including 2 fashion accounts (*blackbirdlondon* and *The-Source*) from 11.5k fashion accounts I have identified and 3 non-fashion accounts (*LamWill*, *EhhYoWIL*, and *ShowoffMadeThis*) from the non-fashion accounts of our original database.

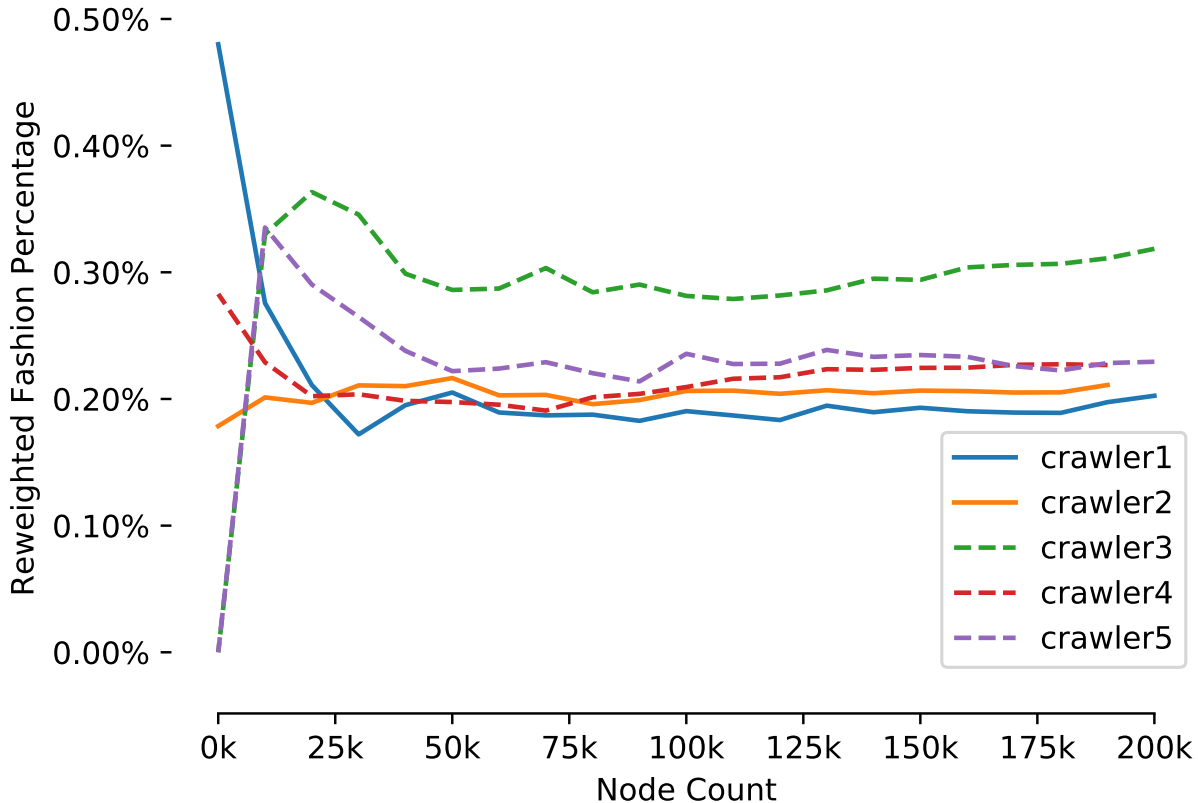


Figure 6.1: The figure shows the fashion ratio of 5 Re-Weighted Random Walks in 5 months(1M total nodes). Crawlers 1 and 2 started the random walk from a fashion account while crawlers 3, 4, and 5 started from a nonfashion account

Given a seed, first, we use `tweepy.API` [26] to crawl this vertex’s followings and followers on Twitter as its neighbors. Consider vertex v_i , the probability that a vertex moves to its neighboring vertex v_j is $1/d(v_i)$ if (v_i, v_j) is an edge in G , otherwise the probability is 0. Second, we randomly selected a new vertex connected to the previous vertex. We repeat these two steps for crawling. We crawled 1 million nodes to determine our estimate.

An ordinary random walk on a graph is biased towards the high degree nodes: the probability that the random walk visits a node v is proportional to its degree d_v i.e. $p(v) \propto d$. To de-bias the walk, we need to re-weight the visited nodes [18] by inverse degree. Thus:

$$f_a = \frac{\sum_{x \in F_a} \frac{1}{d_x}}{\sum_{x \in F} \frac{1}{d_x} + \sum_{y \in \bar{F}_a} \frac{1}{d_y}} \quad (6.1)$$

Where f_a is the fraction of active¹ English speaking fashion accounts and where F_a and

¹We define an active account as one that has activity in the past month; at least two tweets, and where at least two tweet were posted six hours apart. We find that six hours separates social media activity well.

\bar{F}_a refer to the set of active English speaking accounts labeled as Fashion and Non-Fashion respectively.

Analyzing Twitter streaming API is an alternative approach. The Streaming API collects the most recent tweets from active Twitter users, which allows us to study the active user base. The primary challenge of using the streaming API is that the sample is not unbiased, and susceptible to daily variation. I used the Twitter Streaming API to collect over 1 million Twitter accounts over 10 days. I am able to reach a convergent estimate of the expected fraction of Twitter users who are actively tweeting about fashion.

	Streaming		Random Walk	
	Fashion	NonFashion	Fashion	NonFashion
Active accounts Percentage	95.53%	93.65%	59.84%	54.89%
Average Time between tweets	0.52 days	0.85 days	2.26 days	3.44 days

Figure 6.2: The statistics of tweets comparison of Steaming and Random Walk

To summarize our crawling results, the ratio of active English fashion accounts was 0.986% in Streaming and 0.231% in Random Walk. When I examine the average time between tweets, I see that the difference between the fashion percentage in Streaming and Random Walk is ~ 4.264 (0.986%/0.231%). To explain the difference in the fashion percentage, in Figure 6.2 and Figure 6.3 I also calculate the active accounts percentage and the average time between tweets for fashion and non-fashion accounts identified from Streaming and Random Walk. I find that the average time between tweets of fashion accounts from Random Walk is 4.346 (2.26% / 0.52%) times higher than accounts in Streaming. So the difference in tweet arrival rate for accounts labeled as *fashion*, in the two methods. What is clear is that the population present in the streaming data is significantly missing from the random walk. Thus, people who tweet about fashion are much more likely to be also active, partially explaining why I see a higher percentage in streaming data. The other factor is explaining the difference: the Streaming API does not provide an unbiased estimate and may be biased towards more active accounts.

That is, an account is active if it has been used twice over two separate occasions over the past month.

Besides, I want to estimate the size of fashion networks (graph) on Twitter. Based on the 2018 Q2 monthly report, I have about 335 million monthly active Twitter accounts[27] and only 34% of them are English accounts based on the latest data I can find in 2013[28], so I have roughly 113.9 Million English Twitter accounts. When I add the active account criteria, English fashion fraction for Re-Weighted Random Walk (0.231%) and Streaming (0.986%) sampling methods are respectively 263k to 1.123M. So far, I have identified 300,991 active fashion-related accounts among more than 72 million accounts. Based on the estimated total active fashion Twitter accounts, the 300k fashion accounts are roughly covers between 26.8% (1.123M) to 114.4% (263k) of the total active fashion networks. The results indicate that I have enough data to cover the conservative number of fashion-related Twitter accounts.

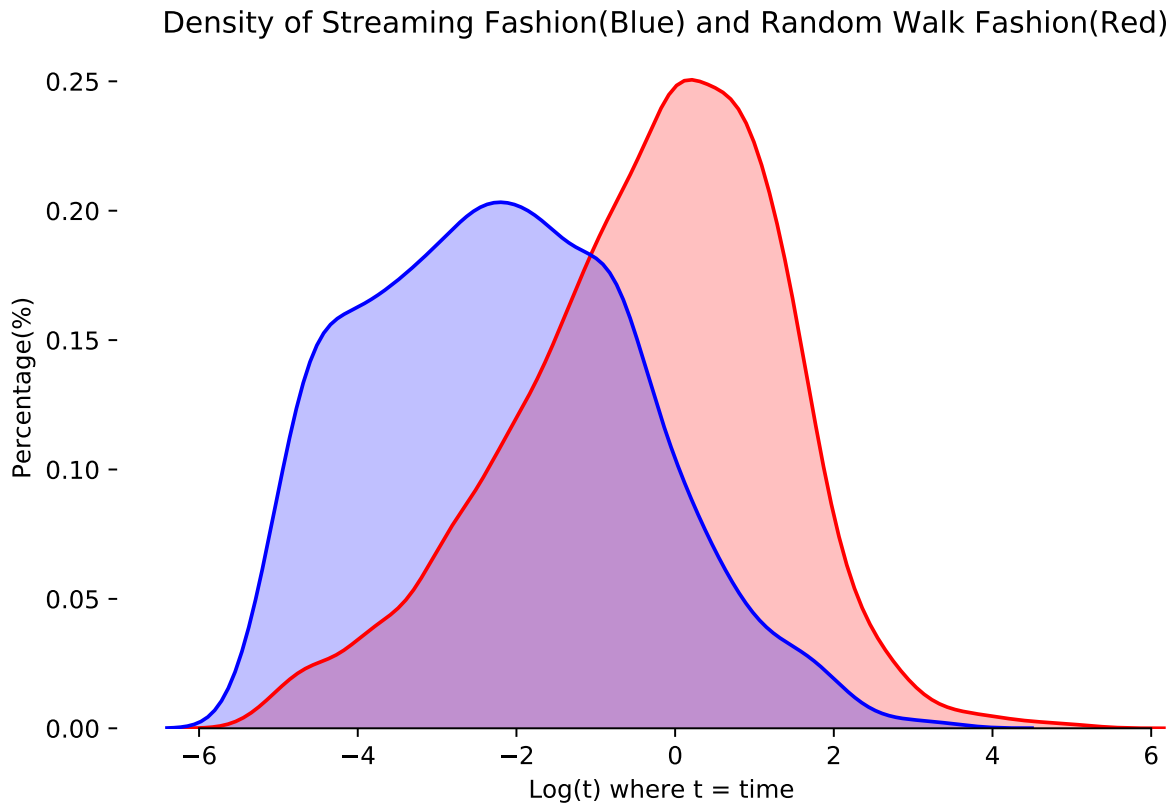


Figure 6.3: The Density Comparison of Streaming Fashion(Blue) and Random Walk Fashion(Red)

CHAPTER 7: FUTURE WORK

There is still a lot more to discover in the fashion graph on social media. Future work should examine how similar fashion accounts can be mined in other popular Social Media platforms such as Instagram [29]. Many accounts that we crawled cross-post on these two platforms. Prior work has studied the evolution of trends [30, 31]; this dataset allows researchers to capture the trends as they happen, and possibly even predict future ones.

7.1 MAPPING TO OTHER SOCIAL NETWORKS

As we all know, not all fashion influencers have Twitter accounts. More importantly, some fashion influencers start to shift their leading social media platforms from Twitter to Instagram where they have more space and attention to share photos of the products of brands or daily lives of individuals. For example, *Off White (@OffWhit)* hasn't posted any tweets since 23 Nov 2018 but has been very active on Instagram. As we compared in figure 5.5, she has been inactive on Twitter but instead has caught lots of attention on Instagram actively. Therefore, only targeting one who uses Twitter as a daily communication tool is a significant limitation. Thus, future work should target on expanding the influencers to similar fashion social network platforms, such as Instagram [29]. I can not only map our existing Twitter accounts to Instagram but also discover new fashion-related accounts to expand our influencer database. Also, unlike users on Twitter, most users on Instagram are willing to provide their geolocation on their Instagram posts, which is much easier for me to study their geometric distribution.

7.2 IDENTIFYING NEW ACCOUNTS

Based on the dynamic graph of the fashion network on Twitter, I cannot cover the complete graph, as such, new accounts are being formed, and accounts become inactive. Besides, we are still curious about the total size of the fashion subgraph and expand our size to dig out and predict the potential future influential account. For example, in 9 years ago, no one knows who is Zoe Sugg (*zoella*) on the Internet until she posted a video to YouTube titled *60 Things In My Bedroom* [32]. Now she becomes a super influential blogger and Internet celebrity. In addition, I restrict the analysis to English accounts only. Besides, I should have an automatic system that can constantly monitor and update the accounts, so none of them will be out of date.

7.3 DETECTING THE FASHION TRENDS OR EVENTS

In addition to detecting if a tweet is RETWEET or MENTION. Tweets serve as a huge text resource, and I believe we can detect event volumes and trends from the text. Besides, cleaning the text, I separated the hashtags from the main text of the tweet. For each tweet, I removed the non-ASCII characters in the text such as emojis, single characters, punctuations, and the stopword. Since the stopwords provided by nltk is not comprehensive, I also added some of the extra stop words. After cleaning the text, I used the Lancaster stemmer to stem each of the words. With these cleaned text data, I can generate spike on some bag of words and word pairs to detecting some potential trend information.

REFERENCES

- [1] B. L. English, *A cultural history of fashion in the 20th century: from the catwalk to the sidewalk*. Berg Publishers, 2007.
- [2] A. Trufelman, “The Trend Forecast,” <http://99percentinvisible.org/episode/the-trend-forecast/>, 2016.
- [3] “Social network analysis of twitter,” May 2015, <http://www.mediative.com/social-network-analysis-twitter/>.
- [4] H. Kwak, C. Lee, H. Park, and S. Moon, “What is twitter, a social network or a news media?” in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 591–600.
- [5] P. Gupta, A. Goel, J. Lin, A. Sharma, D. Wang, and R. Zadeh, “Wtf: The who to follow service at twitter,” in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 505–514.
- [6] E. Gilbert, S. Bakhshi, S. Chang, and L. Terveen, “I need to try this?: a statistical overview of pinterest,” in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2013, pp. 2427–2436.
- [7] Y. Chang, L. Tang, Y. Inagaki, and Y. Liu, “What is tumblr: A statistical overview and comparison,” *ACM SIGKDD explorations newsletter*, vol. 16, no. 1, pp. 21–29, 2014.
- [8] L. Dabbish, C. Stuart, J. Tsay, and J. Herbsleb, “Social coding in github: transparency and collaboration in an open software repository,” in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 2012, pp. 1277–1286.
- [9] D. Easley, J. Kleinberg et al., “Networks, crowds, and markets: Reasoning about a highly connected world,” *Significance*, vol. 9, pp. 43–44, 2012.
- [10] S. Chang, V. Kumar, E. Gilbert, and L. G. Terveen, “Specialization, homophily, and gender in a social curation site: findings from pinterest,” in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 2014, pp. 674–686.
- [11] E. Gilbert and K. Karahalios, “Predicting tie strength with social media,” in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2009, pp. 211–220.
- [12] H. Ward, “The shadow org chart,” Sep 2017, <https://medium.com/@henryward/the-shadow-org-chart-cfcdd644575f>.
- [13] V. Gabale and A. P. Subramanian, “How to extract fashion trends from social media? a robust object detector with support for unsupervised learning,” August 2018, <https://arxiv.org/pdf/1806.10787.pdf>.

- [14] L. Page, S. Brin, R. Motwani, and T. Winograd, “The pagerank citation ranking: Bringing order to the web.” Stanford InfoLab, Tech. Rep., 1999.
- [15] J. Weng, E.-P. Lim, J. Jiang, and Q. He, “Twiterrank: Finding topic-sensitive influential twitterers,” in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, ser. WSDM ’10. New York, NY, USA: ACM, 2010, pp. 261–270.
- [16] Y. Ding, E. Yan, A. Frazho, and J. Caverlee, “Pagerank for ranking authors in co-citation networks,” *Journal of the Association for Information Science and Technology*, vol. 60, no. 11, pp. 2229–2243, 2009.
- [17] L. Zhao and C. Min, “The rise of fashion informatics: A case of data-mining-based social network analysis in fashion,” *Clothing and Textiles Research Journal*, p. 0887302X18821187, 2018.
- [18] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, “Walking in facebook: A case study of unbiased sampling of osns,” *Infocom, 2010 Proceedings IEEE. Ieee, 2010.*, 2010.
- [19] “d3pie,” January 2019, <http://d3pie.org/>.
- [20] “Mongodb,” April 2019, <https://www.mongodb.com/>.
- [21] “React – a javascript library for building user interfaces,” April 2019, <http://www.reactjs.com/>.
- [22] “Force-directed graph,” November 2017, <https://observablehq.com/@d3/force-directed-graph>.
- [23] D. Carriere, “geocoder 1.8.0,” 2017, <https://pypi.python.org/pypi/geocoder/1.8.0>.
- [24] G. M. APIs, “Geocodingapi,” 2017, <https://developers.google.com/maps/documentation/geocoding/s>
- [25] “Swarovski’s tweets,” December 2018, <https://twitter.com/swarovski/status/1039216664678739968>.
- [26] Tweepy, “An easy-to-use python library for accessing the twitter api,” 2017, <http://www.tweepy.org/>.
- [27] “Number of monthly active twitter users worldwide from 1st quarter 2010 to 2nd quarter 2018 (in millions),” 2018, <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>.
- [28] “Only 34% of all tweets are in english,” 2013, <https://www.statista.com/chart/1726/languages-used-on-twitter/>.
- [29] L. Bauerova, “Fashion week on social media: Who took the lead?” *socialbakers.com*, March 2017.
- [30] S. Vittayakorn, A. C. Berg, and T. L. Berg, “When was that made?” *arXiv*, 2016.

- [31] R. He and J. McAuley, “Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering,” *Www*, pp. 507–517, 2016.
- [32] “How zoe sugg became zoella,” February 2018, <https://medium.com/digital-society/how-zoe-sugg-became-zoella-8b1f2a73fe7d>.

APPENDIX A: FASHION CATEGORIES

The following options were provided from the professionals in the fashion industry for categorizing Twitter accounts:

Individual	celebrity, model, stylist, blogger, photographer, editor, designer, buyer
Brand	high fashion brand, mid fashion brand, fast fashion brand, beauty
Retailer	department store, e-commerce site
Media	blog, magazine, newspaper, pr agency, organization, e-zine
Others	inactive, nonfashion, not in the list then user input