Afstanden in Online Sociale Netwerken

Frank Takes (ftakes@liacs.nl)
Leiden Institute of Advanced Computer Science (LIACS)
Universiteit Leiden

Introductie

Online sociale netwerken zoals Facebook, Hyves, LinkedIn en Twitter zijn niet meer weg te denken uit ons dagelijks leven. Ook voor informatici zijn deze netwerken enorm interessant: de netwerkstructuur die schuil gaat achter een sociaal netwerk biedt allerlei uitdagingen op het gebied van data mining, algoritmen en complexiteit. Bovenal wordt de vriendschapsgraaf van sociale netwerken uitgebreid gemeten, gemodelleerd en bestudeerd. Dit artikel gaat specifiek in op het berekenen van kortste afstanden binnen deze vriendschapsgraaf; een taak die met miljoenen gebruikers en tientallen miljoenen vriendschappen niet geheel triviaal is. Deze kortste afstanden zullen we onder andere gebruiken om de zogenaamde Six Degrees of Separation binnen online sociale netwerken te verifiëren.

Vriendschapsgraaf

De *vriendschapsgraaf* is de structuur die schuilgaat achter gebruikers en hun onderlinge vriendschappen binnen een sociaal netwerk. Naast deze *expliciete* vriendschapsgraaf, zijn er ook allerlei *impliciete* grafen af te leiden uit sociale netwerken, bijvoorbeeld aan de hand van wie elkaar een bericht stuurt, wie elkaars profiel bezoekt of wie zich abonneert op wiens videostream.

Formeel kunnen we zeggen dat de vriendschapsgraaf $G = (V, E)$ van een sociaal netwerk bestaat uit een verzameling $V = \{1, 2, \ldots, n\}$ van $n \geq 1$ knopen (gebruikers) en een verzameling $E \subseteq V \times V$ van $m \geq 0$ takken (vriendschappen). We modelleren gebruikers dus als unieke getallen, en een vriendschap tussen twee gebruikers u en v noteren we met (u, v). Afhankelijk van welk sociaal netwerk we bekijken, kan een vriendschap al dan niet wederzijds zijn: op Facebook moet een vriendschap expliciet door beide kanten worden goedgekeurd (als $(u, v) \in E$, dan ook $(v, u) \in E$), terwijl men op bijvoorbeeld Twitter of YouTube zonder toestemming iemand anders kan "volgen" ($(u, v) \in E$, maar niet $(v, u) \in E$). Voor het gemak gaan we in de rest van dit artikel van het eerste (symmetrische) geval uit.

Figuur 1: Sociale Netwerken.

De afstand (distance) tussen twee gebruikers, $d(u, v)$, definiëren we als de lengte van een kortste pad tussen twee gebruikers u en v. Personen die vrienden zijn binnen een sociaal netwerk, liggen dus op afstand 1 van elkaar, vrienden van vrienden op afstand 2, enzovoort. Doorgaans vormt 99% van de gebruikers van een sociaal netwerk een *samenhangend component*, wat betekent dat de waarde van $d(u, v)$ eindig is voor alle u en v, en dat dus "iedereen met iedereen" is verbonden door middel van een serie van vriendschappen. Voor het gemak laten we daarom de mensen die niet tot het grootste samenhangende component behoren (bijvoorbeeld mensen met 0 vrienden, of een groepje volledig geïsoleerde vrienden) buiten beschouwing.

De graaf van een sociaal netwerk is over het algemeen *sparse*: het werkelijke aantal takken m is erg klein ten opzichte van het maximale aantal takken $n(n-1)$. We slaan de graaf daarom in het geheugen op als adjacency list (en niet als adjacency matrix). De kortste afstand tussen twee knopen in de graaf is doorgaans met behulp van het algoritme van Dijkstra [1] te vinden in $O(n^2)$ tijd, maar omdat onze graaf ongewogen en sparse is, kunnen we met behulp van Breadth First Search [2] vanuit de beginknoop in $O(m)$ tijd het kortste pad van A naar B vinden.

Six Degrees of Separation

In 1967 voerde de socioloog Stanley Milgram een experiment uit met een brief, die hij vanuit Omaha (westen Verenigde Staten) naar een adres in Boston (oosten Verenigde Staten) wilde sturen. Basisinformatie zoals de voor- en achternaam van de geadresseerde werden meegegeven, maar bijvoorbeeld niet het precieze adres. Milgram stuurde de brief naar 300 willekeurige personen, met het verzoek de brief door te sturen naar — of in de richting van — de geadresseerde, en Milgram op de hoogte te stellen van hun actie.

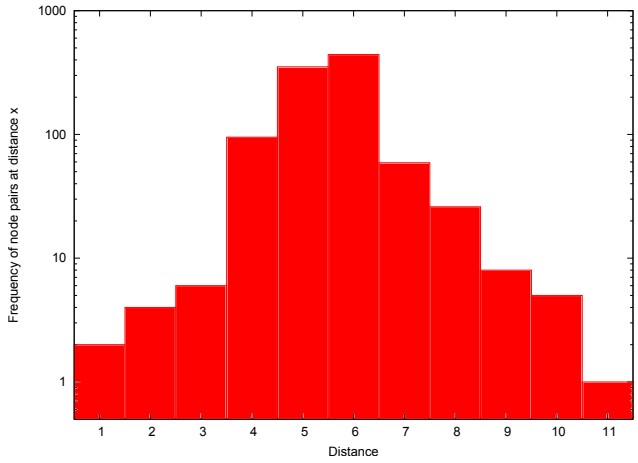
Uiteindelijk bleek dat de brief de geadresseerde niet alleen meervoudig bereikte, maar ook dat de brief daar gemiddeld 6 (met enkele uitschieters naar boven en beneden) stappen voor nodig had. Hoewel Milgram dit zelf nog niet aanduidde als *Six Degrees of Separation* [3], was dit onderzoek wel aanleiding voor veel wetenschappers om dit zogenaamde *Small World Phenomenon* te bestuderen. Blijkbaar was de wereld relatief klein, en was het, ondanks miljoenen inwoners, gemiddeld gezien gemakkelijk om binnen een klein aantal stappen een bericht van A naar B te verspreiden. Formeel gezegd voldoet een graaf aan de *small-world* eigenschap als de gemiddelde afstand tussen twee willekeurige knopen zich logaritmisch verhoudt tot het totale aantal knopen. De oorzaak van deze korte afstanden werd later onder andere toegewezen aan het bestaan van zogenaamde *hubs*: knopen in de graaf met een hoge graad die fungeren als "bruggen" voor het realiseren van een korte afstand tussen de andere knopen.

Er bleken in de jaren daarna nog veel meer netwerken te zijn die aan de small-world eigenschap voldeden. Zo bleek deze eigenschap niet alleen voor de "echte" wereld te gelden, maar bijvoorbeeld ook voor het routing netwerk van het *internet* en voor de linkstructuur van *Wikipedia*. Ook zogenaamde *gene networks*, netwerken die de interactie tussen verschillende genen in ons DNA beschrijven voldoen aan de small-world eigenschap. Het ligt voor de hand dat de Six Degrees of Separation ook binnen de vriendschapsgraaf van een online sociaal netwerk gelden: deze netwerken zijn immers een grove afspiegeling van de "echte" samenleving.

Degree of Separation in een Online Sociaal Netwerk

Om te kijken of de Six Degrees of Separation ook gelden in een online sociaal netwerk, moeten we de afstand van iedere knoop naar iedere andere knoop berekenen. Als we deze afstanden optellen en delen door het aantal paren van knopen, hebben we de gemiddelde "degree of separation", in formulevorm: $\frac{\sum_{u \neq v \in V} d(u,v)}{n(n-1)}$. Een interessante en bruikbare observatie is, dat als we met het algoritme van Dijkstra het kortste pad tussen u en v berekenen, we behalve deze afstand ook de afstand $d(u, w)$ tot alle andere knopen $w \in V$ krijgen. Om de kortste afstand tussen ieder paar knopen te verkrijgen, moeten we dus n keer het algoritme van Dijkstra uitvoeren om de afstand tussen alle $n(n-1)$ paren van knopen te verkrijgen. De complexiteit wordt dus uiteindelijk $O(mn)$ (n kortstepadberekeningen, elk met tijdscomplexiteit $O(m)$). Voor een graaf van 5 miljoen knopen (en 100 miljoen takken) duurt 1 iteratie van Dijkstra al snel 1 seconde, wat neer zou komen op 58 dagen rekenen om alle kortste afstanden te berekenen.

Een betere oplossing is *sampling*: we berekenen de kortste afstand tussen zeg 1000 willekeurige knopen, en rekenen vervolgens de gemiddelde gevonden afstand uit. Het resultaat van de distributie van deze berekeningen voor een (groot deel van) het sociale netwerk LiveJournal (`http://www.livejournal.com`) met 5 miljoen gebruikers en 100 miljoen vriendschappen is weergegeven in Figuur 2. We kijken dus naar het *aantal* paren van gebruikers (verticale as) dat op respectievelijk op afstand $1, 2, 3, \ldots$ (horizontale as) van elkaar ligt. Er is een duidelijke piek rond 5 en 6 (merk op dat de verticale schaal logaritmisch is), wat erop duidt dat de gemiddelde afstand ergens tussen deze twee waarden ligt. En inderdaad, het blijkt dat de gemiddelde afstand 5.6 is, wat dus betekent dat persoon A gemiddeld in 5.6 stappen persoon B kent.



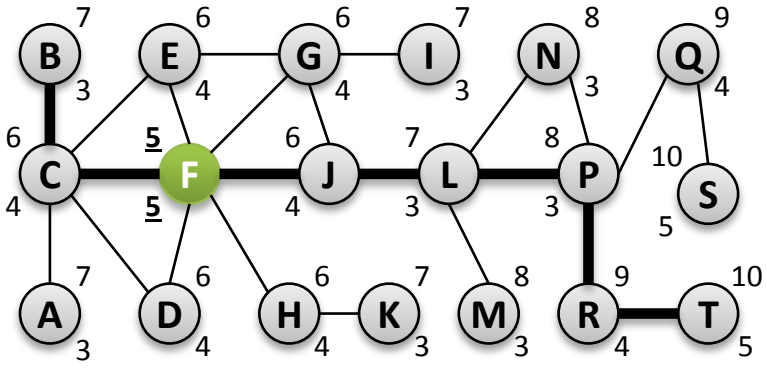
Figuur 2: Distributie van de gevonden afstand tussen een sample willekeurige knopen A en B.

"Maximale" Degree of Separation

We hebben met ons eenvoudige sampling-experiment laten zien dat de small world-eigenschap zeker ook geldt binnen een online sociaal netwerk, en dat de "degree of separation" zelfs iets minder dan 6 is. Soms willen we echter ook iets zeggen over de *langste* kortste afstand, ookwel bekend als de *diameter* van de graaf (formeel: $\max_{u,v \in V} d(u,v)$). De diameter zou wellicht iets kunnen zeggen over hoe snel een bericht zich in het ergste geval verspreidt via een sociaal netwerk. Ook in andere netwerken kan de diameter interessant zijn: in het routing netwerk van het internet zegt de diameter bijvoorbeeld iets over de worst-case response-time tussen twee machines.

Een naieve manier om de diameter te berekenen is om voor ieder van de n knopen (in $O(m)$ tijd) zijn langste kortste afstand (de zogenaamde *eccentriciteit* van die knoop) te berekenen, en uiteindelijk de hoogste gevonden waarde te retourneren alszijnde de diameter. De complexiteit van deze naieve aanpak is dus wederom $O(mn)$: veel te complex gezien de aantallen knopen en takken in de vriendschapsgraaf van een sociaal netwerk. In dit geval kunnen we echter geen sampling gebruiken om deze complexiteitsexplosie op te lossen: we zijn namelijk niet op zoek naar een gemiddelde kortste afstand, maar naar de *maximale* kortste afstand.

Een efficiëntere manier om de diameter van een sociaal netwerk exact te berekenen, is met behulp van onder- en bovengrenzen. We zullen een voorbeeld van dergelijke grenzen schetsen. Stel dat we weten dat persoon u eccentriciteit 5 heeft en iedereen dus kan bereiken in maximaal 5 stappen. Dan weten we ook dat een directe vriend van u, zeg v, iedereen in het slechtste geval kan bereiken in maximaal 6 stappen, namelijk door dit altijd via u te moeten doen. In het beste geval gebruikt u altijd v om alle andere knopen te bereiken, en zou v precies 4 stappen nodig hebben om overal te kunnen komen. We weten nu dus dat de eccentriciteit van v in ligt tussen 4 en 6, alleen op basis van de eccentriciteit van 5 van zijn vriend u. Algemener kunnen we zeggen dat de eccentriciteit van alle knopen



Figuur 3: Graaf met 19 knopen en 23 takken. Het pad van B naar T realiseert de diameter. Getallen onder en boven de knoop geven de onder- en bovengrenzen voor eccentriciteit aan, nadat de eccentriciteit van F excact is berekend.

v op afstand $d(u, v) = k$ van knoop u, een eccentriciteit hebben die ligt tussen de eccentriciteit van u minus k (of indien groter: k zelf) en de eccentriciteit van u plus k. Omdat de diameter gelijk is aan de grootste eccentriciteit, kunnen we zeggen dat de diameter ligt tussen het maximum van alle bekende ondergrenzen voor de eccentriciteit en het maximum van alle bekende bovengrenzen van de eccentriciteit. Daarnaast weten we ook dat de diameter nooit groter kan zijn dan 2 keer de kleinst bekende eccentriciteitsbovengrens over alle knopen in het netwerk. Met behulp van deze grenzen is de diameter van de graaf in Figuur 3 met drie eccentriciteitsberekeningen (knoop F, T en L) te achterhalen.

Uit experimenten blijkt dat het gebruik van dergelijke boven- en ondergrenzen in het geval van sociale netwerken (en andere small-world netwerken) heel erg goed werkt. Vaak kunnen we door het herhaaldelijk toepassen van deze grenzen met enkele tientallen (in plaats van n) eccentriciteitsberekeningen de diameter van deze netwerken exact berekenen. Dit komt doordat de hubs (populaire gebruikers) in een sociaal netwerk over het algemeen een lage eccentriciteit hebben die de bovengrens van de diameter verlagen, en reguliere (minder actieve) gebruikers juist een wat hogere eccentriciteit hebben die daarmee weer de ondergrens van de diameter verhogen.

Voor het eerder genoemde sociale netwerk LiveJournal van 5 miljoen knopen, bleek de diameter 23 te zijn: ondanks dat de gemiddelde afstand tussen twee mensen dus erg klein is, zijn er toch mensen die op afstand 23 van elkaar liggen. Deze afstanden zijn echter een hoge uitzondering: twee willekeurig geselecteerde personen kennen elkaar *gemiddeld* nog altijd in slechts 5 of 6 stappen.

Referenties

[1] E. W. Dijkstra, *A note on two problems in connexion with graphs*, Numerische Mathematik, 1 (1959), pp. 269–271.

[2] A. Levitin, *Introduction to the design & analysis of algorithms*, Addison-Wesley, 2003.

[3] D. Watts, *Six degrees: The science of a connected age*, WW Norton & Company, 2004.

Deze tekst is verschenen in Eureka! Magazine jaargang 9, nummer 33, pp. 8–11, 2011.