Caracterização e Influência do Uso de Tips e Dones no Foursquare

Marisa Vasconcelos[1], **Saulo Ricci**[1], **Jussara Almeida**[1], **Fabricio Benevenuto**[2], **Virgílio Almeida**[1]

[1] Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)
31.270-010 - Belo Horizonte - Brasil

[2]Departamento de Ciência da Computação
Universidade Federal de Ouro Preto (UFOP)
35.400-000 - Ouro Preto - Brasil

{marisav,saulomrr,jussara,virgilio}@dcc.ufmg.br, benevenuto@gmail.com

Abstract. *Online Location Based Social Network (LBSNs) allow users to share their location with friends and followers through check ins or tips, textual impressions about a certain venue. A tip can be marked as done, if a user agrees with the content of the tip. Unlike user check ins, tips and dones are visible to everybody and they promote the interaction among venue visitors and venue owners besides being used to recommend new venues. This work presents a user behavior characterization regarding tips and dones in Foursquare, the most popular LBSN, performed over a dataset crawled from the same network. Through this analysis, we have observed the presence of very active or influent users who post many tips and received a large number of dones. However, measure influence based on the number of dones might be subject to spammers. Other factors such as the number of posted tips and dones given by other influent users should be considered. To better identify who are the influential users, we have propose a PageRank-like algorithm which considers not only the done network but the number of posted tips.*

Resumo. *As redes sociais baseadas em geolocalização (LBSN) são redes que permitem que seus usuários compartilhem sua localização com seus amigos e seguidores através de check ins e tips que são impressões sobre determinado local. Uma tip pode ser marcada como "done" se um usuário concorda com o conteúdo da tip. Ao contrário dos check ins, as tips e dones são visíveis a todos os usuários do sistema e são importantes para promover a interação entre os visitantes e proprietários do local, além de serem utilizadas na recomendação de novos locais. Esse trabalho apresenta uma caracterização de como os usuários interagem entre si utilizando tips e dones, através da coleta de seus perfis do Foursquare, a maior das redes LBSN. Através dessa análise, observou-se a presença de usuários muito ativos ou influentes que postam muitas tips e recebem muitos dones. Entretanto, medir influência baseado no número de dones recebido pelas tips de um usuário é sujeito a spammers. Outros fatores como*

o número de tips postadas e dones dados por outros usuários influentes devem ser levados em consideração para avaliação destes. Para melhor se identificar quais são esses usuários foi proposto um algoritmo baseado no PageRank que leva em conta não só a rede de dones como também o número de tips postadas.

1. Introdução

Redes sociais baseadas em geolocalização (LBSNs) são um novo paradigma nas redes sociais e vêem atraindo cada vez mais novos usuários. Redes como Foursquare, Brightkite, Gowalla e outras permitem que o usuário compartilhe a sua localização geográfica com sua rede social (check-in) através de algum tipo de smartphone dotado de GPS. Nesse trabalho iremos focar no Foursquare que é a maior rede desse tipo com aproximadamente 10 milhões de usuários [News 2011].

No Foursquare, os check-ins podem ser realizados em uma variedade de locais (venues) e podem ser acumulados na forma de pontos que permitem que os usuários ganhem medalhas (badges), prefeituras (mayorships) e recebam ofertas do local. Outro aspecto do Foursquare são as dicas (tips) postadas pelos usuários nos venues, que refletem suas impressões positivas ou negativas sobre a visita a esse local. As tips podem ser marcadas por outros usuários como "done" se o usuário concorda com ela ou adicionada a uma lista de to-dos para ser realizada posteriormente. Quando um usuário posta uma tip, ele está interagindo com um negócio no mundo real e expondo fraquezas ou qualidades sobre este negócio para todos, incluindo a concorrência. Assim, tips e dones são ferramentas valiosas para os negócios do mundo real porque além de serem fontes da opinião de seus clientes, elas podem proporcionar oportunidades de melhorias e pode afetar muito as visitas futuras e finalmente os lucros.

Alguns trabalhos recentes analisaram propriedades das redes LBSNs e características de seus usuários, inclusive para o Foursquare. No entanto, a maioria desses estudos focou na dinâmica dos *check ins*, em propriedades do grafo social e nas informações geográficas relacionadas a essas métricas [Scellato et al. 2010, Noulas et al. 2011]. Em um trabalho anterior [Vasconcelos et al. 2012], o único a respeito de tips e dones em LBSNs, foram identificados perfis de usuários através de algoritmo de clusterização, onde dois tipos principais de usuários se destacaram: os influentes que recebiam um grande número de dones em suas tips e os spammers cujas tips continham links e eram postadas em vários venues.

Recentemente, o Foursquare disponibilizou através de sua API, novos dados a respeito do usuário, como a sua rede social e seus dones dados. Assim, foi feita uma nova coleta para a análise desses novos aspectos e da estrutura da rede de dones que possibilitou observar aspectos de interação entre os usuários, que não puderam ser capturados pela primeira análise. Foi observado um grande número de componentes fortemente conectadas, o que sugere a ausência de comunidades e uma baixa influência da rede social no número de dones recebidos pelo usuário. Esses resultados juntamente com diversos comportamentos dos usuários nas diversas categorias de venues podem estar relacionados com a influência de certos usuários.

A influência dos usuários nas redes sociais é vem sendo muito discutido na literatura e vários serviços na web (Klout[1], PeerIndex[2] entre outros) apresentam rankings compostos por usuários pertencentes a várias redes sociais. Existem serviços como o Venue Machine[3] que enviam alertas ao proprietário do venue quando usuários influentes definidos pelo Klout realizam *check ins* nesses locais. O objetivo é que os proprietários possam estar alerta para atender àquele usuário da melhor forma possível. Atualmente, o único índice de influência disponível para o Foursquare é o Klout. A desvantagem do Klout é que o índice de influência apresentado é um mix entre as várias redes sociais e não há um índice separado para cada rede. Isso é um problema já que, por exemplo, um usuário pode ser influente no Twitter, mas pode não ser influente no Foursquare.

Outro problema desse índice é que a influência é medida em termos do número agregado de checkins e dones recebidos pelo usuário. O próprio Foursquare também utiliza essas métricas para o venue como entrada do seu sistema de recomendação, além de dar destaque às tips como maior número de dones na página do venue. No entanto, em redes como o Foursquare, a ocorrência de abusos (por exemplo, spammers) é facilitada já que não há a participação ativa de moderadores. Sem uma indicação sobre a reputação/influência desses usuários recomendadores, o sistema fica vulnerável a todo tipo de ataque [Post 2011]. Assim para o sistema, a identificação de usuários influentes poderá servir com um índice de credibilidade para as tips postadas e para os usuários isso serviria como um filtro de tips.

Existem diversas abordagens para identificação de influentes e aquelas que são o estado da arte fazem uso da estrutura da rede social utilizando algoritmos como PageRank. Outro aspecto que deve ser levado em consideração para medir a influência de um usuário é o aproveitamento das tips postadas, ou seja, a fração das tips do usuário que receberam dones. Nossa proposta o PageRank Normalizado leva em consideração esse aproveitamento para o ranqueamento do usuários influentes.

2. Trabalhos Relacionados

A área de redes sociais de geolocalização têm tido poucos estudos publicados principalmente por ser uma área recente. Uma das primeiras publicações correlacionou propriedades sociais com geográficas em várias redes sociais incluindo as redes LBSNs e mostrou-se que como nas redes sociais tradicionais, há a prevalência de conexões de amizade em distâncias curtas [Scellato et al. 2010]. Em [Cho et al. 2011] padrões de mobilidade foram investigados e modelados nas redes do Gowalla, Brightkite e traces de telefones celulares e em [Cheng et al. 2011] foi feita uma análise quantitativa sobre os padrões de mobilidade humana levando em consideração propriedades espaciais, temporais, sociais e textuais dos *check ins* em diversas redes LBSNs.

Três estudos sobre caracterização dessas redes se destacam. Em [Li and Chen 2009], os autores utilizaram duas técnicas de clusterização para identificar padrões de comportamento de usuários na rede do Brightkite. Uma das abordagens classificou os usuários quanto à mobilidade dos padrões de atualização (*check ins*, fotos e comentários) e outra agrupou esses mesmos usuários levando em conta também

[1]http://www.klout.com
[2]www.peerindex.com
[3]http://www.venuemachine.com

os aspectos sociais. O segundo estudo [Noulas et al. 2011] utilizou um algoritmo de clusterização espectral para agrupar os usuários baseado nos padrões de *check ins*, por categoria a fim de identificar comunidades e caracterizar o tipo de atividade em cada região de uma cidade. Finalmente, o estudo feito por [Vasconcelos et al. 2012] iniciou as análises sobre os padrões de comportamento na interação dos usuários no Foursquare utilizando tips e dones e continuada nesse trabalho. Nesse estudo foi realizada uma caracterização dos venues e dos usuários quanto ao número de tips postadas, dones recebidos, venues recomendados e porcentagem de tips que continham links. Um algoritmo de clusterização foi utilizado para agrupar os usuários em quatro grupos com diferentes comportamentos. Dois desses grupos correspondiam a usuários regulares que se diferenciavam apenas em termos do nível de postagem de tips. Outro grupo consistia em usuários influentes, engajados em postar tips em uma grande variedade de venues e o último grupo era composto por spammers que foram validados através de uma inspeção manual de suas tips.

Na área de influência em redes sociais ainda não foram feitos trabalhos sobre LBSNs, mas três trabalhos em outras redes foram utilizados com base para esse. Em [Zhang et al. 2007] os autores analisam diversas abordagens de ranqueamento para identificação de especialistas em um fórum sobre Java, utilizando métodos de redes sociais como PageRank e HITS. Em [Adamic et al. 2008] foram feitas análises sobre os usuários do Yahoo Answers considerando também propriedades estruturais como grau do vértice e número de arestas mútuas. Finalmente, em [Weng et al. 2010] foi proposto um algoritmo baseado no PageRank para identificação de influentes no Twitter que levava em consideração o tópico do *tweet*. Embora todos esses estudos ofereçam idéias para as análises sobre as interações dos usuários nas redes LBSNs, nenhum deles é focado nos aspectos da interação por tips e dones como meio de recomendação de locais ou serviços.

3. Base de Dados do Foursquare

O Foursquare foi criado em 2009 e atualmente é a maior rede social baseada em geolocalização que permite que seus usuários compartilhem sua localização com amigos/seguidores através de *check ins*. Similarmente ao vídeo para o YouTube, images para o Flickr, os venues são os objetos principais no Foursquare. Um venue representa uma local no mundo real como, por exemplo, uma loja, um restaurante, um aeroporto ou um monumento, onde o usuário pode dar *check in* se estiver fisicamente perto do local e munido de um algum celular com GPS. Os *check ins* podem ser acumulados na forma de medalhas (badges) ou prefeituras (mayorship) se o usuário realizar mais *check ins* que qualquer outro usuário no mesmo venue.

Os venues são criados pelos próprios usuários do Foursquare, mas podem ser demandados pelos proprietários reais caso apresentem algum tipo de comprovação. Se aprovada, o venue muda seu status para verificado e os proprietários podem oferecer promoções para os usuários que frequentemente dão *check ins*. Cada venue no Foursquare é classificado em uma das nove categorias pré-definidas: *Arts* & *Entertaiment*, *Colleges* & *Universities*, *Food*, *Great Outdoors*, *Nightlife Spots*, *Travel Spots*, *Residences*, *Professional* & *Other Places* e *Shops* & *Services*.

Além do *check in* , o usuário pode postar tips nos venues para que futuros visitantes possam ter algum tipo de referência sobre o local. As tips são pequenos textos que

contém recomendações positivas ou negativas sobre um venue, como por exemplo, o melhor prato de um restaurante ou uma reclamação sobre o atendimento recebido no local. Como os *check ins*, elas são compartilhadas com os amigos e seguidores assim que são postadas e têm o diferencial de estarem acessíveis quando se realiza um busca por algum local nas redondezas ou se acessa o venue. Após ler uma tip, o usuário pode adicioná-la em sua lista privada de *to-do* ou marcá-la como *done*. O conteúdo de uma lista de to-dos é uma informação privada ao usuário e sua rede social. O número de vezes que uma tip foi marcada como done é uma informação disponível no Foursquare e serve com estimativa da quantidade de *feedback* vinda de outros usuários que leram a tip, além de ser uma mecanismo de identificar boas recomendações a serem seguidas.

O Foursquare também classifica seus usuários em três tipos: *users*, *celebrities* e *brands*. A principal diferença entre eles é o tipo de relacionamento social que eles podem ter. No caso do *user* somente relações do tipo amizade (mútua) ou eles podem também seguir usuários dos outros tipos. O usuário do tipo *celebrity* possui os dois tipos de relacionamento: amizade e seguidor/seguido. Os usuários do tipo *brand* são usuários encarregados de postar conteúdo na forma de tips, fotos e comentários e só podem se relacionar com seguidores.

3.1. Coleta dos dados

Foi obtida uma amostra dos perfis de usuários, através de coleta utilizando a API do Foursquare. Para que a coleta ficasse minimamente tendenciosa foi realizada uma coleta aleatória dos usuários. Essa estratégia de coleta foi baseada no fato que Foursquare assinala identificadores numéricos únicos e sequenciais (ID) para cada um de seus usuários. Estimando-se que o maior identificador assinalado a um usuário seja M, o coletor escolhe um ID aleatoriamente entre 0 e M de acordo com uma distribuição uniforme. De cada um desses IDs escolhidos foram requisitadas à API as informações sobre o usuário, suas tips postadas, sua lista de dones, seus amigos e seguidores. Para cada tip coletada também foram extraídas informações sobre o venue como categoria e coordenadas geográficas.

Para se estimar o valor de M, foram feitas requisições HTTP às páginas dos usuários identificados por seus IDs. Foram feitas tentativas para se aumentar o valor dos IDs, começando por zero. O maior ID para o qual se obteve uma resposta de página válida foi 10 milhões e muitos IDs com valores maiores que esse obtiveram resposta do tipo "Página não encontrada". Assim, acreditamos que durante o período da coleta, esse era o valor do maior ID de usuário e esse valor foi utilizado como entrada do coletor. O coletor foi executado de 18 de Outubro a 10 de Novembro de 2011, obtendo-se dados de quase 9 milhões de usuários. A Tabela 1 sumariza a base de dados coletada. Associado aos usuários coletados, foi possível identificar mais de 5 milhões de tips e mais de 2 milhões de venues que receberam pelo menos uma tip em diversos países.

4. Caracterização do Uso de Tips

Nessa seção é apresentada uma caracterização dos três tipos de usuários (*user*,*celebrity* e *brands*) em termos do número de tips postadas, do número de venues recomendados, do número de dones marcados ou recebidos e do número de amigos e seguidores para os usuários que postaram tips. Em seguida, apresentamos uma análise das tips e dones por categoria de venue.

Tabela 1. Sumário do dataset do Foursquare

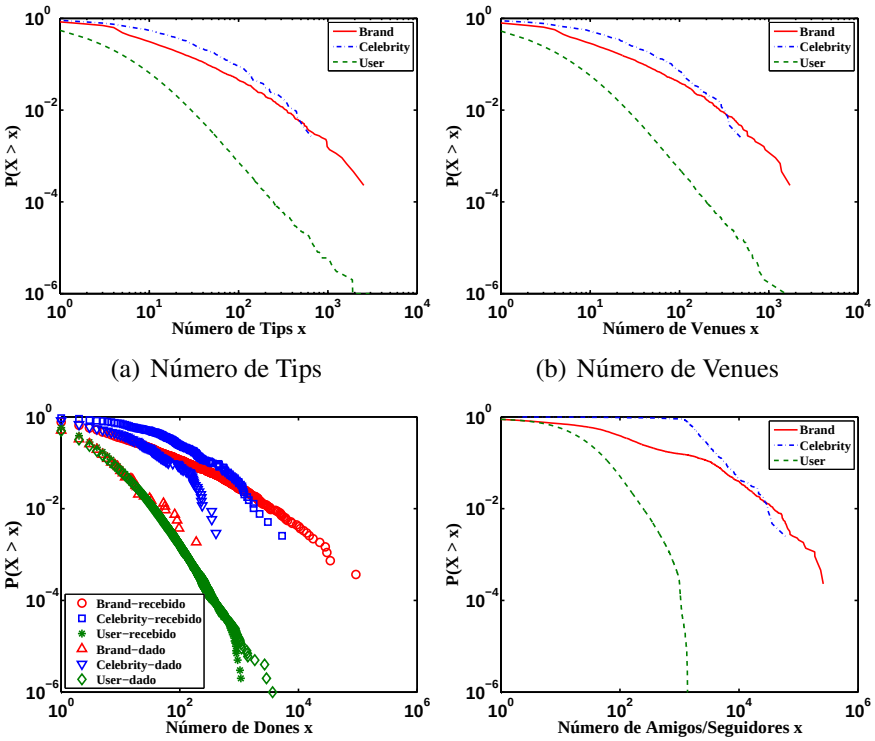
Users	
Número de usuários	8.997.265
Número de usuários que postaram pelo menos 1 tip	1.437.832
Número de usuários do tipo *brand*[a]	4367
Número de usuários do tipo *celebrity*[a]	423
Número de usuários do tipo *user*[a]	1.433.042
Número de usuários que receberam pelo menos 1 done	604.282
Número de usuários que deram pelo menos 1 done	850.344
Venues	
Número de venues com pelo menos 1 tip	2.683.709
Tips	
Número de tips	5.501.000
Número total de dones em todas as tips	3.318.353

[a] Usuários que postaram mais de uma tip

4.1. Participação dos usuários na recomendação de venues

Podemos observar que a participação de todos os tipos de usuário é desbalanceada, similarmente as distribuições de participação em outros tipos de redes sociais como fóruns, sistemas de *tagging* colaborativos, etc. A Figura 1 apresenta a função de distribuição acumulada complementar (CCDF) de cada uma das métricas. Desses gráficos podemos observar que os usuários diferem no nível de atividades (tips ou dones) no Foursquare. A Figura 1(a-b) mostra que para todos os tipos de usuários a maioria dos usuários postam poucas tips em poucos venues e que alguns usuários mais ativos postam muitas tips em vários venues. A maior diferença entre o número de dones dados e recebidos foi entre os usuários do tipo *brand* (Figura 1(c)). Essa diferença pode ser explicada pelo fato que os *brands* se ocupam principalmente por postar tips e não de avaliar tips de outros usuários. Comparando-se as Figuras 1(a) e 1(c) podemos observar que mesmo havendo usuários do tipo *user* que postam muitas tips, os usuários mais ativos que recebem dones são do tipo *brand*. Analisando-se a distribuição do número de amigos e seguidores para cada tipo de usuário (Figura 1(d)) podemos notar que a rede social é esparsa para os usuários que postaram pelo menos uma tip onde, somente 5% dos usuários do tipo *user* possuem mais de 100 amigos, 15% das *brands* possuem mais de 1000 seguidores, o que reforça a popularidade dos *brands* não só no número de dones mas também no número de seguidores.

Para cada uma dessas métricas foi observada também a correlação entre elas. Foi encontrada uma correlação forte positiva (coeficiente de Spearman [Kendall and Gibbons 1990] em torno de 0.98 para todas as classes de usuários) entre o número de tips e o número de venues recomendados por cada usuário. O que sugere que usuários que postam muitas tips tendem a espalhá-las em diversos venues. Para o número de tips e o número de dones recebidos foi encontrada uma correlação baixa para os *brands* (0.34) e *users* (0.43) e moderada para os *celebrities* (0.51) para o top 10% dos usuários com mais tips de cada tipo. Isso implica que os usuários que postam muitas tips não necessariamente recebem mais dones. Se observarmos o top 10% dos usuários com mais dones de cada tipo a correlação com o número de tips é de baixa a moderada (0.27-0.46). O que significa que usuários que recebem muitos dones não necessariamente postam muitas tips.



(a) Número de Tips (b) Número de Venues

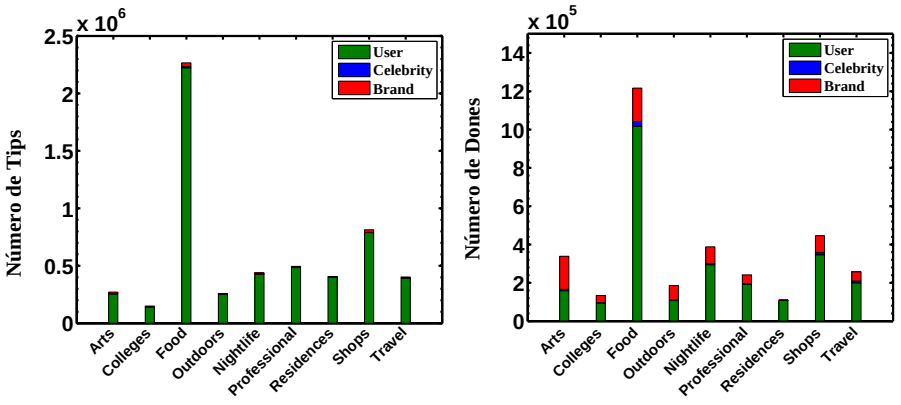
(c) Número de Dones (d) Número de Amigos/Seguidores

Figura 1. Distribuição das características dos usuários (escala loglog)

Ao correlacionarmos o número de tips que um usuário posta e o número de amigos/seguidores de sua rede social foi encontrada uma correlação baixa positiva para o tipo *user* (0.34) e *celebrity* (0.08). Para o tipo *brand* foi observada uma correlação moderada (0.54) a baixa (0.21) se calculada para o top 10% desses usuários com mais tips ou dones. Assim, usuários que postam muitas tips não necessariamente possuem muitos amigos/seguidores. Para os usuários do tipo *brand*, a correlação entre o número de dones recebidos e o número de amigos/seguidores é moderada (0.65) a alta (0.81) para o top 10% dos usuários com mais tips. Para os tipos *user* e *celebrity* a correlação foi baixa. Esse resultado pode ser interpretado de duas formas: uma brand com muitos dones pode estar atraindo mais seguidores para a sua rede social ou seus seguidores estão frequentemente marcado as tips desse usuário com dones. A primeira hipótese não pode ser verificada com base de dados atual, já que não dispomos de quando o usuário se juntou a rede social do *brand*. A segunda hipótese será analisada na Seção 5.1.

Como foi discutido na Seção 3, os venues do Foursquare são agrupados em 9 categorias pré-definidas. As Figuras 2(a-b) apresentam o número de tips e dones recebidos em cada categoria. A maior parte das tips estão na categoria *Food*, assim como os dones recebidos. A categoria *Food* possui também a média mais alta de tips por usuário: 2.77 tips para usuários do tipo *user*, 30.42 para os do tipo *celebrity*, 15.29 para os do tipo *brand*. A média de dones recebidos por usuários do tipo *user* na categoria *Food* foi de 1.16 dones/usuário e 1.84 para os usuários do tipo *celebrity*. Para os usuários *brand* a categoria com a maior média de dones foi a categoria *Colleges* com 3.59 dones por usuário. Assim, podemos observar que cada categoria possui níveis de atividade diferentes e algumas

categorias possuem mais de tips e dones que outras. Na próxima seção, serão analisadas a dinâmica entre os usuários que postam tips e aqueles que as marcam como *done* em cada uma das categorias.



(a) Número de tips postadas (b) Número de dones recebidos

Figura 2. Tips e Dones por categoria

5. Estrutura da Rede de Dones

Para se entender alguns aspectos sobre interação entre os usuários que marcam e/ou recebem dones em suas tips, foi construído uma rede de dones. Nessa rede, cada usuário que marcou ou recebeu algum done foi modelado como um vértice e a aresta que os conecta é direcionada do usuário que marcou done (U_x) para o usuário autor da tip (U_y). O peso v da aresta representa o número de tips do usuário U_y que o usuário U_x marcou como done. Apenas usuários que deram ou receberam algum done estão representados no grafo, sendo assim o grau de cada vértice é maior que zero. A Tabela 2 sumariza as principais características do grafo inteiro e dos grafos por categoria.

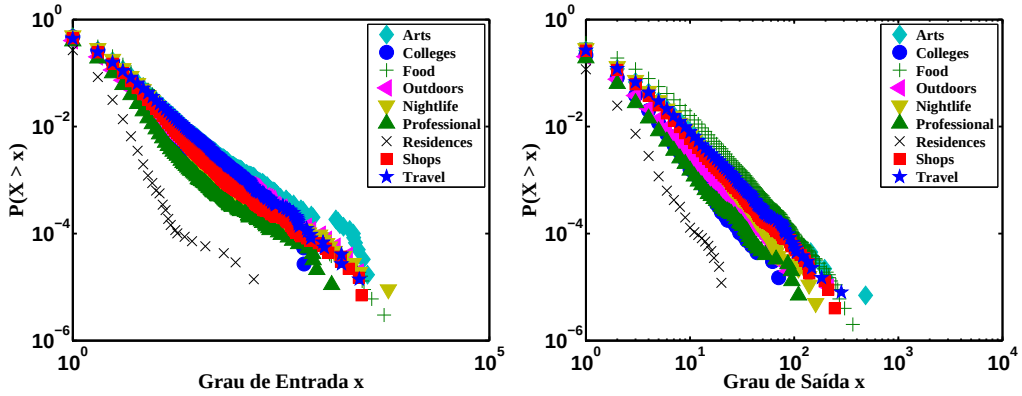
Tabela 2. Sumário das características do grafo de dones por categoria

Categoria	# Vértices	# Arestas	Grau Médio	#Arestas mútuas	# vértices da SCC
Sem categoria	1,143,914	2,283,949	2.0	61,233 (2.75%)	147,282
Arts	177,876	222,952	1.25	1288 (0.60%)	116
Colleges	95,274	94,608	0.99	1766 (1.90%)	16
Food	650,543	964,012	1.48	17,497 (1.85%)	34,117
Outdoors	140,479	140,559	1.00	979 (0.70%)	28
Nightlife	253,631	309,081	1.22	5421 (1.79%)	1559
Professional	221,091	199,578	0.90	4104 (2.10%)	26
Residences	138,320	98,319	0.71	5336 (5.74%)	9
Shops	321,049	362,310	1.00	4094 (1.14%)	31
Travel	186,970	223,881	1.20	1451 (0.65%)	361

Como foi observado na Figura 1(c), poucos usuários dão dones em tips de vários usuários (grau de entrada alto) e poucos usuários recebem dones de vários usuários em todas as categorias (Figura 3). Embora em todas as categorias se veja a presença de distribuições do tipo cauda pesada, os usuários mais ativos da categoria *Residences* dão ou recebem um número menor de dones se comparado às outras categorias. Outra característica do grafo dessa categoria é a maior porcentagem de arestas mútuas. Isso pode

ser explicado pelo fato que os venues dessa categoria representam locais privados, geralmente a residência de um usuário e sua localização geográfica é revelada somente para os amigos do usuário. Assim, apesar das tips serem públicas a todos os usuários, o conteúdo das tips dessa categoria pode não ser interessante a outros usuários do sistema.

Também podemos ver que nos grafos de cada categoria, as distribuições do grau de entrada são mais largas do que as do grau de saída indicando que existem mais usuários que recebem dones (grau de entrada maior) do que usuários que marcam dones (graus de saída maior). Esse desbalanceamento entre a marcação e o recebimento de dones pode ainda ser confirmado através do tamanho da maior componente conectada (SCC) de cada um dos grafos que mostra que poucos usuários dão e recebem dones entre si.



(a) Grau de Entrada (b) Grau de Saída

Figura 3. Distribuição dos graus de entrada e saída no grafo de dones por categoria (escala loglog)

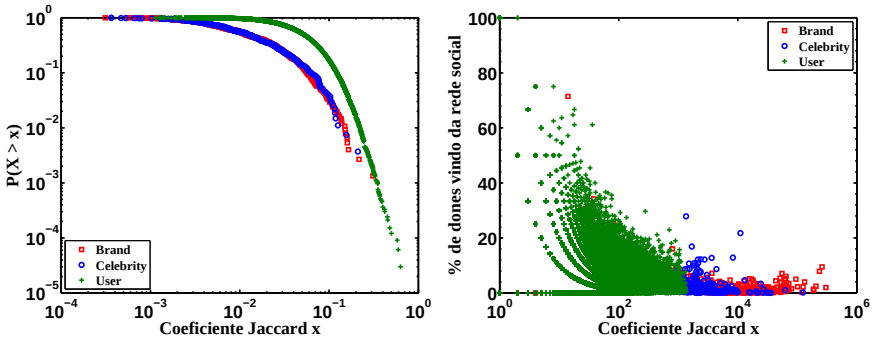
5.1. Influência da Rede Social

Nessa seção será analisada a origem dos dones recebidos por um usuário, ou seja, como sua rede social influencia na popularidade de suas tips e consequentemente no número de dones acumulados por esse usuário. É importante ressaltar que o venue onde a tip foi postada também influencia no número de dones que uma tip pode receber. Deixamos o estudo do impacto da venue no # dones de uma tip para o futuro.

Para quantificar o número de dones recebidos dentro da rede social foram utilizadas duas métricas: a distribuição do coeficiente de similaridade Jaccard [Hand et al. 2001] entre as duas redes (social e dones) e a porcentagem de dones recebidos que vieram de dentro da rede social. O coeficiente de Jaccard representa a fração do número de usuários da rede social de um usuário que deram done em alguma de suas tips pelo tamanho da união da a sua rede social e a sua rede de dones. Pela Figura 4(a), podemos observar que grande parte dos usuários possui um baixo coeficiente de Jaccard o que mostra que grande parte dos dones recebidos venha de fora da rede social do usuário. A Figura 4(b) mostra que mesmo para usuários *brand* que possuem um grande número de seguidores, o número de dones que vem de fora da rede social é maior que os que vem de dentro dela.

Baseado nessas evidências, podemos observar que mesmo usuários bem conectados não necessariamente recebem dones de sua rede social. Assim, ao contrário de outras redes sociais, onde o grafo social é importante para se estimar a influência de um usuário,

no Foursquare, isso não é necessariamente verdadeiro. Na próxima seção é proposto um método para o ranqueamento dos usuários influentes baseado no grafo construído a partir dos dones dados e recebidos por usuários do sistema.



(a) Distribuição do Jaccard dos usuários (escala loglog)

(b) Distribuição da % de dones que vieram da rede social por usuário (escala log no eixo x)

Figura 4. Influência da rede social (usuário com mais de 10 dones)

6. Influência dos Usuários

Nessa seção é introduzido o método proposto para identificação dos usuários influentes no Foursquare. A abordagem mais intuitiva e simples seria a medição através do número de dones recebidos por usuário. No entanto, um usuário que recebeu 100 dones em uma única tip seria ranqueado igualmente a um usuário que postou 100 tips e recebeu um done em cada uma. Acreditamos que o primeiro usuário deva ter mais influência que o último.

O algoritmo de PageRank [Page et al. 1998] tenta minimizar esse problema, levando em consideração a estrutura da rede. Assim a influência de um usuário é computada utilizando-se a influência direta ou indireta de todos os outros usuários através da propagação pela rede. No caso da rede de dones, um usuário se torna influente se outro usuário influente marcou alguma tip desse usuário como done. Uma desvantagem desse método para o contexto do Foursquare é que o número de tips postadas não é levado em consideração. Assim, se um usuário obteve 100 dones através de 10 tips e outro usuário obteve o mesmo número de dones através de 200 tips, o método favoreceria ao último usuário. No entanto, o primeiro usuário deveria ser favorecido já que a aceitação de suas tips é maior que o do último usuário.

Para isto foi proposto um algoritmo baseado no PageRank, o PageRank Normalizado, que leva em conta três idéias chave: o número de dones recebidos (grau de entrada do vértice), o número de tips postadas pelo usuário e o prestígio do usuário que postou. Em nossa definição de influência, um usuário está no topo do rank, quando ele obteve um bom aproveitamento das tips que ele postou e recebeu algum done de algum outro usuário influente. O aproveitamento das tips é medido pela fração entre o número de dones recebidos e o número de tips postadas.

6.1. Page Rank Normalizado

De forma intuitiva, um usuário deve ser influente se ele pode ser interpretado como expert ou autoridade em determinado assunto. No caso do Foursquare, essa influência na

recomendação de venues é medida pelo número de dones recebido pelo usuário em suas tips. Além disso, se um usuário recebe um done de outro usuário influente, isso o torna influente, já que essa tip foi útil para algum usuário. Esse tipo de propagação de influência é feita por algoritmos baseados na estrutura da rede como o PageRank. A equação 1 representa o PageRank tradicional para todos os vértices de uma rede de tamanho N:

$$P(v) = \frac{(1-\alpha)}{N} + \alpha \sum_{u \in M(v)} \frac{P(u)}{N_v} \tag{1}$$

onde v representa um vértice (usuário), $M(v)$ é o conjunto de vértices que apontam para v, P(v) e P(u) representam o valor do PageRank para os vértices u e v respectivamente e N_v é o número de arestas que saem do vértice v. A constante α representa o *dumping factor* (entre 0 e 1) que é a probabilidade da *Random Walk* seguir as arestas que saem de v e a sua complementar $(1-\alpha)$ é a probabilidade da *Random Walk* recomeçar de qualquer outro vértice do grafo.

O PageRank proposto para esse trabalho se diferencia do PageRank tradicional porque leva em consideração o número de tips que usuário postou. Para esse novo PageRank, o peso I de cada aresta (u, v) do grafo é ponderada pelo número de tips do usuário v que recebeu algum done vindo do usuário u:

$$I(u, v) = \frac{d(u, v)}{t(v)} \tag{2}$$

onde $d(u, v)$ é o número de dones dados por u em tips de v e $t(v)$ é o número de tips postadas por v. Essa definição captura e quantifica o aproveitamento das tips postadas pelo usuário v. Assim quanto mais tips de v forem marcadas done por u, maior a influência de v sobre u. O valor do PageRank modificado para um usuário v é definido como:

$$P(v) = \frac{(1-\alpha)}{2}\left[\frac{1}{N} + \frac{t(v)}{\sum_{w \in V} t(w)}\right] + \alpha \sum_{u \in M(v)} I(u, v) * P(u) \tag{3}$$

onde α é um *damping factor*, N é o número de vértices no grafo, V é o conjunto de vértices do grafo e $M(v)$ é o conjunto de arestas que apontam para v, ou seja o conjunto de usuários que marcou alguma tip de v como done. A componente global (fatores multiplicados por $1-\alpha$) foi dividida em duas partes (divisão por 2). A primeira componente $1/N$ tem a primeira interpretação do PageRank tradicional, isto é, fazer com a *Random Walk* recomece de qualquer vértice do grafo com a mesma probabilidade. Já a segunda componente valoriza os usuários que postaram muitas tips. A intuição do algoritmo PageRank Normalizado é a adição ao fator de dumping (1-alfa/N), de um outro fator para privilegiar usuários que postaram mais tips. Assim, a componente local do PageRank dá importância aos usuários que tiveram muitos dones em poucas tip, e a componente global valoriza aqueles que postaram muitas tips.

6.2. Resultados

Nessa seção, são apresentados os resultados da aplicação do algoritmo proposto no grafo de dones da base de dados coleta. O algoritmo é comparado com duas estratégias de

ranqueamento: uma baseada somente no número de dones do usuário e outra baseada PageRank tradicional. A avaliação dos usuários influentes foi feita como em outros trabalhos da literatura: os índices de influência dos usuários não são públicos e logo não existe um *ground-truth* para comparação. O intuito não é apontar qual ranking é o melhor, mas mostrar as diferenças nas três abordagens. A Tabela 3 lista os top-5 usuários mais influentes de acordo com cada algoritmo para o grafo geral e em cada categoria. Os usuários *celebrities* estão marcados com (c) e os do tipo *user* com (u). Para o grafo geral (sem categorias), quatro dos cinco usuários mais influentes apareceram nos quatro rankings. No entanto, alguns desses usuários não aparecem no Top-5 de algumas das categorias, o que sugere que alguns desses usuários podem ser mais influentes em determinadas categorias. A Tabela 4 lista os valores da correlação Kendall-Tau [Kendall and Gibbons 1990] entre os rankings dos diversos algoritmos. O coeficiente medido por esta correlação deve estar entre -1 e 1. Se os dois rankings são exatamente os mesmos, o valor do coeficiente é igual 1, assim quanto mais perto de 1, maior é a concordância entre os rankings.

Tabela 3. Usuários influentes por categoria

Categoria	Número de Dones	PageRank	PageRank Norm
Sem categoria	History Channel, Bravo, MTV, Wall Street J., Visit PA	History Channel,Bravo,MTV, Wall Street J., Zagat	History Channel, Bravo, MTV, Wall Street J., Zagat
Arts	History Channel, Explore Chicago, Visit PA,Wall Street J., NHL	History Channel, Wall Street J., NHL, Visit PA, Explore Chicago	History Channel, Wall Street J., La Vida(u), TLC,Explore Chicago
Colleges	Arizona State U, Mizzou, UW-Madison,Cal, Stanford U	History Channel, Sports Authority, Mizzou, Stanford U., Graphic Master (c)	bookrenter.com, Northeastern U., Sports Authority, Mizzou,old main(u)
Food	Bravo, Zagat, Eater.com, Wall Street J., Thrillist	Bravo,Zagat, Thrillist, Visit PA, Wall Street J	Bravo, Foodspotting (c), Zagat, Britt r.(u), Wall Street J.
Outdoors	History Channel, Wall Street J., Explore Chicago, Windows Live P.G., Visit PA	History Channel, Wall Street J., Explore Chicago, Bravo, Windows Live P.G.	History Channel, Wall Street J, Explore Chicago, Healthy Pools (u)
Night	MTV, Bravo, LogoTV, Thrillist, History Channel	MTV, Bravo, Thrillist, History Channel, LogoTV	Bravo, MTV, Thrillist, History Channel, Logo TV
Professional	History Channel, Wall Street J., Explore Chicago, National Post, Huffpost	History Channel, Wall Street J., Explore Chicago, Visit PA, Graphic Master(c)	History Channel, oregonvotes.org (u), USF Athetics (u), Wall Street J., Liam B.(u)
Residences	History Channel, Greensboro NC(u), Huffpost, Lee(c), Mizzou	History Channel, Huffpost, Lee(c), Mizzou, Supermodelme	David K.(u), Beth D.(u), Lee (c) , sarah w.(u), capture the market (u)
Shops	Bravo, Visit PA, Graphic Master(c), Mazda, (red)	Bravo, Visit PA, Graphic Master(c), History Channel, Mazda	Bravo, A bag´s life (u), AT&T, cew(u), Visit PA
Travel	History Channel, National Post, Wall Street J., National Post,KLM	History Channel, Wall Street J., Bravo, KLM, Explore Chicago	History Channel, Wall Street J. KLM, AT&T, Tubus(u)

Tabela 4. Correlação entre os rankings das diferentes métricas

Categoria	Dones vs PageRank	Dones vs PageRank Norm	PageRank vs PageRank Norm
Sem categoria	-0.0048	0.0985	0.5041
Arts	-0.0421	-0.0819	0.5100
Colleges	-0.0086	-0.0617	0.5412
Food	-0.0371	0.0220	0.4859
Outdoors	-0.0861	-0.1503	0.4758
Night	-0.0253	-0.0349	0.5134
Professional	-0.1236	-0.1793	0.4654
Residences	-0.3187	-0.3101	0.4296
Shops	-0.0702	-0.0829	0.4637
Travel	-0.0677	-0.0967	0.4711

Para o ranking do número de dones, grande parte dos usuários são *brands* e dois usuários, um do tipo *user* na categoria *Residences* e outro *celebrity* na categoria *Shops* também aparecerem entre os top influentes em suas categorias. As tips do usuário *Greensboro NC* geralmente são sobre locais na cidade de Greensboro e por ser tratar de um

categoria em que os usuários postam poucas tips, o número de dones desse usuário ainda é maior que outros usuários *brands* da mesma categoria. O usuário *Graphic Master*, que representa uma empresa de designers localizada em Singapura, posta tips que promovem seu negócio (spam) e dicas de locais da mesma cidade.

Para o ranking gerado pelo PageRank, observarmos também a prevalência de usuários brands entre os top-5 em todas as categorias. Não houve uma diferença considerável entre os top-5 dos rankings do número de dones e do PageRank. No entanto, para categorias onde o número de usuários que postam tips é menor como *Colleges* e *Residences*, foi observada a presença de usuários muitos influentes como o *History Channel* no ranking do PageRank e usuários do tipo *celebrity* como *Graphic Master* e *Lee*. Se compararmos o coeficiente de Kendall entre esses dois rankings, nota-se que a correlação é baixa ou seja, os rankings gerados são diferentes. Como foi ressaltado na seção anterior, o ranking por dones é sujeito a ataques de spammers. O Page Rank leva em consideração a estrutura da rede, onde dones vindo de outros usuários influentes contribuem para a influência desses top usuários, o que melhor captura a noção de influência.

O Page Rank Normalizado também retornou muitos usuários do tipo *brand* que foram influentes nos outros algoritmos como o *History Channel* e o *Bravo*. Podemos notar também que vários usuários não listados entre o top-5 dos outros algoritmos foram considerados influentes. Dentre esses usuários, podemos ressaltar o usuário *Britt R.* que foi considerado influente pelo algoritmo na categoria *Food*, porque algumas de suas tips foram marcadas como done por outro usuário influente (National Post).

6.2.1. Discussão dos Resultados

A divisão do grafo de dones por categoria mostra que existem usuários que são influentes no geral, mas não são tão influentes em determinadas categorias. É possível observar a presença de usuários experts em determinadas categorias que pode ser identificada através de um grafo especializado por categoria. Dentre os top-5 usuários mais influentes, podemos ver a existência de usuários que já eram especializados fora da rede Foursquare, como por exemplo, universidades dentro da categoria *Colleges* e usuários que representam sites de recomendações de restaurantes (Zagat e Eater.com) na categoria *Food*. Como mostrado para redes de fóruns, as propriedades estruturais também podem ser utilizadas para se identificar usuários influentes (experts) em LBSNs e a suas influências relativas através da utilização de métodos baseados em redes como o PageRank. Como foi descrito, a ausência de moderação no Foursquare facilita a disseminação de spam ou tips que possam difamar o venue e métodos como o proposto nesse trabalho, que levam em consideração a influência relativa e o aproveitamento das tips postadas, podem minimizar a importância desses usuários. A identificação de usuários influentes provê subsídios para vários serviços como busca e recomendação (priorizando conteúdo de usuários mais influentes), organização interna dos servidores, alocação de conteúdo em CDNs e métodos para detecção de usuários maliciosos (pouco influentes). A metodologia proposta nesse trabalho pode ser utilizada em outras redes sociais onde a interação com os outros usuários determina a influência do usuário (e.g. Twitter).

7. Conclusões

LBSNs vêem se tornando populares por permitir o compartilhamento de informação social e geográfica. Esse trabalho focou na análise da interação entre os usuários através de tips que são utilizadas para recomendação de locais e dones que servem como *feedback* da tips postadas. A análise foi baseada em uma base de dados coletada do Foursquare, a rede mais popular entre as LBSNs. Nesse trabalho também foram analisadas propriedades estruturais da rede de dones composta por autores e usuários que votaram nas tips postadas. Utilizando essas redes, foi proposto um algoritmo baseado no PageRank para medição da influência levando em consideração o aproveitamento de suas tips do usuário, medida pela fração do número de dones recebidos por tip postada. Os resultados mostram que métodos que se baseiam apenas no número de dones podem não ser efetivos na detecção de influentes e, além disso, podem ser susceptíveis a ações maliciosas.

8. Agradecimentos

Esta pesquisa é parcialmente financiada pelo Instituto Nacional de Ciência e Tecnologia para a Web - INCTWeb (MCT/CNPq 573871/2008-6), CNPq, CAPES e FAPEMIG. Agradecemos a Giovanni Comarela pela ajuda nesse artigo.

Referências

Adamic, L., Zhang, J., Bakshy, E., and Ackerman, M. (2008). Knowledge Sharing and Yahoo Answers: Everyone Knows Something. In *Proc. of WWW'08*.

Cheng, Z., Caverlee, J., Lee, K., and Sui, D. (2011). Exploring Millions of Footprints in Location Sharing Services. In *Proc. of ICWSM'11*.

Cho, E., Myers, S., and Leskovec, J. (2011). Friendship and Mobility: User Movement In Location-Based Social Networks. In *Proc. of KDD'11*.

Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. MIT Press.

Kendall, M. and Gibbons, J. (1990). *Rank Correlation Methods*. Oxford University Press, 5 edition.

Li, N. and Chen, G. (2009). Analysis of a Location-based Network. In *Proc. of IEEE CSE'09*.

News, W. P. (2011). Foursquare grow. http://www.webpronews.com/foursquare-grows-by-3400-are-you-part-of-it-2011-01.

Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. (2011). Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks. In *Proc. of SMW'11*.

Page, L., S. Brin, R. M., and Winograd, T. (1998). The Pagerank Citation Ranking: Bringing Order to the Web. Stanford Digital Library Technologies Project.

Post, N. M. (2011). Business owners try again to sue yelp for payola. http://www.mediapost.com/publications/article/151204/.

Scellato, S., Mascolo, C., Musolesi, M., and Latora, V. (2010). Distance Matters: Geo-social Metrics for Online Social Networks. In *Proc. of WOSN'10*.

Vasconcelos, M., Ricci, S., Almeida, J., Benevenuto, F., and Almeida, V. (2012). Tips, Dones and Todos: Uncovering User Profiles in Foursquare. In *Proc. of the WSDM'12*.

Weng, J., Lim, E.-P., Jiang, J., and He, Q. (2010). TwitterRank: Finding Topic-sensitive Influential Twitterers. In *Proc. of the WSDM'10*.

Zhang, J., Ackerman, M., and Adamic, L. (2007). Expertise Networks in Online Communities: Structure and Algorithms. In *Proc. of WWW'07*.