A Importância das Localidades Geográficas na Difusão Online de Informação

Rodrigo Marotti Togneri, Bruno Tadeu Caetano, Carlos Alberto Kamienski

Centro de Matemática e Ciência da Computação (CMCC)
Universidade Federal do ABC (UFABC) – Santo André, SP – Brasil

rodrigo.togneri@ufabc.edu.br, bruno.caetano@ufabc.edu.br, cak@ufabc.edu.br

Resumo. *A localização geográfica dos usuários é subestimada como fator importante para a difusão de informação em redes sociais, em ambiente online. Isso porque os ambientes virtuais causam a impressão de que as afinidades de interesse são tão mais importantes que outros fatores perdem completamente a relevância. Assim, através da análise de uma rede real de recomendações online, este artigo propõe que a localização geográfica é importante para a difusão de informação em redes sociais virtuais. Sugere-se então que as estatísticas públicas, indexadas por localização geográfica, são uma maneira eficaz e de baixo custo para conhecer melhor os usuários e, consequentemente, melhorar a informação e a difusão.*

1. Introdução

A localização geográfica dos usuários é subestimada como fator importante para a difusão de informação em redes sociais. Sobretudo em ambientes virtuais, nos quais existe a impressão de que as afinidades de interesse são tão mais importantes que outros fatores perdem completamente a relevância. Segundo Manuel Castells (1996-98), estudioso da Era da Informação, a relevância da localização espacial é diminuída em favor da relevância da afinidade de interesses entre as pessoas, pois a Internet realmente tem o poder de facilitar a comunicação à distância e aproximar as pessoas que tem interesses comuns. A relevância da localização geográfica pode ter sido diminuída, como é intuitivo, mas continua relevante, como será visto aqui.

A não observância de um fator relevante como a localização geográfica implica em perda de oportunidade de conhecer melhor os usuários da informação. Sem conhecer os usuários, perde-se a oportunidade de melhor adequar a informação ao seu público-alvo, e, assim, perde-se a oportunidade de melhorar a difusão da informação. E isso pode ser decisivo quanto ao sucesso ou fracasso de uma iniciativa viral, seja ela a promoção de um produto por uma empresa ou a veiculação de uma política pública. No caso de empresas, a situação é ainda mais crítica, pois além de se promover a difusão, deve-se ser mais eficiente que a concorrência.

O objetivo principal do artigo é buscar comprovar a hipótese de que a localização geográfica é importante para a difusão de informação em redes sociais online. Ou seja, que existem características intrínsecas de cada localidade que influenciam a difusão de informação. Essa conclusão deriva do estudo da rede real de recomendações online de um provedor de acesso à Internet. Verificou-se que as localidades possuem intensidades diferentes de utilização do objeto da difusão (no caso,

o acesso à Internet) e funções distintas no fluxo de informação, o que vem a ratificar a hipótese principal.

Dessa forma, a principal contribuição deste trabalho é a identificação do indício de que a difusão de informação online se dá de maneira diferente em cada local. E que essa descoberta, se ratificada por outros casos reais, pode apresentar grande impacto entre os gestores responsáveis pela elaboração e execução de políticas públicas e pelos responsáveis por campanhas de marketing viral, por exemplo. Ainda existem outras contribuições. Uma delas é a própria metodologia para análise de recomendações: a ideia de transformar uma rede de recomendações entre pessoas em uma rede de conexões entre localidades (seção 4). Além disso, dado o pressuposto de que as localidades geográficas são diferentes e suas características intrínsecas importam para a difusão viral, outra contribuição deste trabalho é a sugestão da utilização das estatísticas públicas para a melhoria da difusão de informação em redes sociais online. Primeiro, porque as estatísticas públicas são convenientemente organizadas segundo a localização geográfica, devido aos governos serem organizados desta forma. Historicamente, os governos se desenvolveram em torno da ideia de lideranças fisicamente locais, devido à falta de meios adequados de comunicação a longas distâncias. Também sabemos que as estatísticas públicas nasceram para amparar os governos em suas decisões. Logo, os dados estatísticos tiveram suas principais quebras estruturais também por localidades. Segundo, porque são, em sua maior parte, gratuitas, possibilitando informação relevante a baixo custo. E terceiro, porque tem dados os mais diversos, tanto de caráter demográficos quanto socioeconômicos.

O artigo segue com a seção 2, que o situa em meio aos trabalhos acadêmicos relacionados, ressaltando as contribuições aqui propostas. A seção 3 contextualiza o caso real estudado. A seção 4 dá o entorno metodológico do trabalho. Na seção 5 são apresentados e discutidos os resultados, seguidos pela conclusão (seção 6) e pelas referências bibliográficas.

2. Trabalhos Relacionados

Nos últimos anos, mediante a evidência do papel relevante das redes sociais virtuais na difusão de informação, foram publicados trabalhos importantes sobre assuntos correlatos.

Hill et al. (2006), Minhano et al. (2010), Adamic e Adar (2003) e Ma et al. (2011) mostraram, através da análise de casos reais, que informações das redes sociais online são relevantes para a difusão de informação. O presente trabalho, embora também faça uso de um caso real, tem um enfoque diferente: mostra que informações dos locais geográficos (de um caso real em particular) são relevantes para a difusão de informação.

Centola (2010) e Bampo et al. (2008) investigaram com competência como a estrutura da rede social se relaciona com a difusão de informação, fazendo, para isso, uso de estruturas de rede artificialmente criadas. Essa abordagem é bastante útil para as aplicações em que se pode escolher de antemão qual a estrutura da rede social virtual. Tratam-se, então, de trabalhos complementares a este, pois exploram as estruturas artificiais, são especulativos, enquanto este explora uma rede real, são descritivos da realidade.

Outra visão interessante é fornecida por Yang et al. (2010), que relaciona métricas de redes com redes de recomendação com o objetivo de estimar qual seria o número de iniciadores ideal. Os iniciadores são os primeiros recomendadores. A conclusão é que quanto maior o número de iniciadores, mais rápido a difusão se dá; porém, a partir de um determinado número de iniciadores, o incremento de velocidade de difusão passa a não ser substancial. O mesmo acontece com o tamanho final da rede de recomendações. Portanto, se cada iniciador for custoso a quem queira difundir a informação, pode-se estudar um número ótimo de iniciadores correspondente ao custo-benefício almejado. É uma visão complementar ao presente trabalho, pois, enquanto com as técnicas aqui apresentadas se consegue conhecer melhor os locais que devem ser estimulados, o trabalho de Yang pode dizer quantos iniciadores serão precisos, por exemplo.

Dois trabalhos que corroboram uma das conclusões aqui apresentadas, a de que as estatísticas públicas devem ser mais exploradas em difusão online, são os Scellato et al. (2010) de Kaltenbrunner et al. (2012). Eles estabelecem, através da análise de casos reais de redes sociais virtuais, que a distância física é importante no estabelecimento de conexões, e que, dado que as conexões foram estabelecidas, a distância física não é relevante para a intensidade dessas conexões. Kaltenbrunner et al. descobriram que os usuários tendem a se conectar localmente e sugeriram, da mesma forma que aqui é sugerido, que os profissionais da difusão agrupassem os dados dos usuários por localidade geográfica, afim de conhecerem as características locais do seu público. É importante notar a diferença básica entre os trabalhos citados e este: os trabalhos de Kaltenbrunner et al. e Scellato et al. demonstram a importância das distâncias geográficas nas redes sociais online; já o presente trabalho demonstra a importância da localidade do usuário nas redes sociais online, particularmente em um processo real de difusão da informação. O fato de se estar relacionado com as localidades é ainda mais favorável à conclusão de que as estatísticas públicas são importantes para a difusão online, pois as estatísticas estão diretamente indexadas aos locais. As distâncias geográficas se relacionam apenas indiretamente com as estatísticas públicas.

Outro trabalho que corrobora com as conclusões deste trabalho é o de Marques (2009). Segundo ele, as redes de indivíduos em situação de pobreza são em geral menores, mais locais e menos variadas do que os de indivíduos de classe média. Isso significa que existem relações relevantes entre a geografia da estrutura de rede e fatores socioeconômicos. Assim, Marques incentivou a utilização de informações de rede dos indivíduos para o estudo e para o planejamento de políticas públicas. O presente trabalho também incentiva a utilização de métricas de rede aliada a dados socioeconômicos para fins práticos. A diferença, nesse caso, é que Marques propõe que sejam avaliadas métricas de rede de cada indivíduo, e aqui se propõe as métricas de rede de locais geográficos em que os indivíduos habitam. E que as aplicações práticas dessas propostas podem ser mais amplas, englobando políticas públicas e também iniciativas de marketing de empresas.

Pode-se observar, então, que este trabalho é especial porque é o único que confronta de localidades geográficas com difusão de informação online na análise de um caso real. A demonstração de que existe relação entre geografia e difusão online é particularmente importante do ponto de vista prático, pois permite aos profissionais de difusão não só utilizarem as informações de relacionamento entre os locais, como também toda a gama de informações das estatísticas públicas vinculadas espacialmente. Com mais informação relevante, os profissionais de difusão podem entender melhor seu

público e melhor adequar a informação, bem como estimular regiões com melhor potencial de alavancagem de difusão.

3. Características Gerais do Caso real

Os dados trabalhados são de propriedade de uma reconhecida desenvolvedora de software, atuante no mercado web brasileiro, e foram gentilmente cedidos para esta pesquisa. No caso em questão, a empresa teve atuação como uma provedora de acesso discado à internet que difunde seu discador pelo estímulo de uma rede de recomendações. O cliente pode recomendar o discador a outros indivíduos enviando convites via email por intermédio de um aplicativo no próprio site da empresa. Em os convidados aderindo ao discador através do convite, estes são vinculados à rede de recomendação dos respectivos convidadores.

A expansão da rede de recomendações é incentivada pelo modelo de negócio adotado pela empresa, no qual o usuário não só não paga pela utilização discador (paga apenas pela utilização da linha telefônica), como também tem uma recompensa mediante o uso do discador e mediante o uso dos seus recomendados. Ou seja, quanto mais o usuário e seus recomendados utilizam o discador, maior é a recompensa dele. Essa recompensa ocorre em uma moeda virtual, cujo saldo pode ser resgatado em moeda real.

O caso real em questão tem três características que o tornam particularmente interessante: o tamanho substancial da rede de recomendações, a verificação de todo o ciclo de difusão da informação e o fato de se tratar de um caso raro de difusão viral "puro". Cada uma dessas características é explorada a seguir.

A rede de recomendações possui tamanho relevante, o que permite flexibilidade para subagrupar os dados sem a perda de relevância estatística das características resultantes. E, consequentemente, boa consistência dos resultados obtidos. São aproximadamente um milhão de usuários de todo o Brasil e doze milhões de convites realizados, dos quais 5% foram efetivados perfazendo uma rede de recomendações de 671.045 usuários (68% do total)[1]. Os convites partiram de 123.163 usuários, dos quais a metade deles (59.771) conseguiu converter os convites em conexões. Percebe-se que os convidadores de sucesso foram responsáveis por estabelecer uma rede de recomendações onze vezes maior, uma alavancagem considerável que corrobora a importância de se saber em que locais se concentram esses usuários.

A segunda característica importante é que o objeto, o acesso à Internet, cumpriu praticamente todo o ciclo de sua difusão (figura 1). A vantagem, nesse caso, é que se tem a certeza de que a rede trabalhada é madura, ou seja, já estabeleceu a plenitude de sua atuação.

[1] Observa-se que existem usuários que não fazem parte da rede de recomendações. Isso se deve a estes terem se cadastrado diretamente no site do provedor sem referenciar seu convidador, e, ainda, não terem estabelecido nenhuma outra conexão. Esses usuários foram desconsiderados.

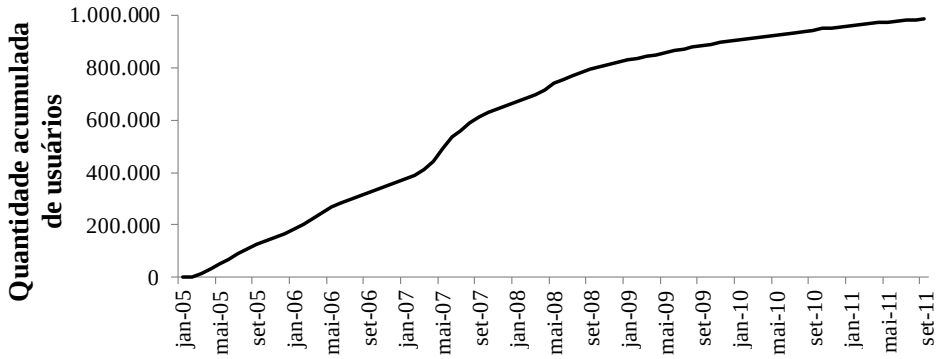

Figura 1. Curva característica da difusão do acesso à Internet oferecido pelo provedor. O período dos dados vai do início de 2005 (lançamento do serviço de acesso à Internet) até setembro de 2011, quando o serviço do provedor ainda estava em operação. A quantidade acumulada de usuários ao longo do tempo evolui para uma assíntota horizontal, indicando saturação do meio. A difusão está perto do fim.

Por fim, a terceira característica importante é que se trata de um caso de difusão viral "puro". Não houve investimento em publicidade por parte do provedor. Nem em qualquer outra estratégia que não o estímulo de uma rede de recomendações. Isso é particularmente interessante porque em casos reais, é muito comum não ser possível isolar um fenômeno como acontece aqui. Neste trabalho, não há deturpações dos resultados oriundas de publicidade de massa, por exemplo. Os resultados remetem exclusivamente ao foco deste trabalho, a difusão de informação, pessoa a pessoa.

4. Metodologia

4.1. Formação de Agrupamentos Geográficos

Para que a importância das localidades fosse avaliada, os usuários foram agrupados em seus locais de origem. Para isso, foi utilizada a informação do código postal brasileiro (conhecido como CEP), que atribuiu uma posição geográfica para cada usuário. A abrangência dos agrupamentos de usuários foi determinada em quatro níveis. O nível mais abrangente é formado pelas cinco macrorregiões geográficas brasileiras. O segundo nível mais abrangente é o formado pelas 27 unidades da federação (UF), constituintes das macrorregiões. Em seguida, um nível mais granular, o formado pelos locais determinados pelos três primeiros dígitos do CEP, o código postal brasileiro. Em geral, os três primeiros dígitos do CEP delimitam bairros onde há maior concentração populacional, e cidades ou agrupamento de cidades onde há menor concentração. Por fim, o último nível avaliado, o mais granular de todos, é formado pelos locais determinados pelo CEP completo, com oito dígitos. Em locais de maior concentração populacional, o CEP completo costuma delimitar uma única rua, enquanto que em locais de menor concentração, chega a delimitar uma cidade inteira. Os objetivos aqui são de avaliar se as conclusões observadas são as mesmas para diversos níveis de granularidade geográfica, e, de confirmar a suposição de que quanto mais granular o nível geográfico mais precisamente se consegue definir onde estão os melhores e piores usuários, através da mensuração da utilização média do discador pelos usuários de cada local.

Dos agrupamentos referentes aos três primeiros dígitos do CEP, foi construída uma rede de locais. A granularidade dos três primeiros dígitos foi escolhida pois ela é

considerada pelos profissionais de marketing como a de melhor relação entre precisão no espaço geográfico com qualidade de informação disponível. Se o espaço geográfico fosse mais granularizado, não se encontraria boa qualidade de estatísticas públicas, principalmente devido às dificuldades de se ter amostras representativas e às dificuldades inerentes às pesquisas de censo e correlatas. Essa rede de locais construída foi estruturada de tal forma que, se um ou mais usuários de um local X tenha convidado e estabelecido conexão com um ou mais usuários de um local Y, foi atribuída uma conexão direcionada de X para Y.

4.2. Atribuição de Métricas de Rede para cada Local da Rede de CEP de 3 Dígitos

Para cada local são atribuídas características quanto à sua participação na rede geográfica de CEP de 3 dígitos, isto é, quanto ao fluxo de conexões. Essas características são advindas das seguintes métricas de rede [Newman, 2003[2]]: grau de entrada, grau de saída, k-core, coeficiente de aglomeração e autorreferência.

O coeficiente de aglomeração foi a única métrica calculada considerando a rede como não direcionada, ou seja, se existirem entre dois locais uma ou duas (uma para cada sentido possível) conexões direcionadas, estas serão reduzidas a uma única conexão sem atribuição de sentido de fluxo da informação. Os locais estão simplesmente conectados de alguma forma. Essa simplificação é adequada, pois a utilidade desta métrica aqui é objetivamente dizer o quanto os conectados diretos de um local são coesos de uma forma genérica. Para os fins deste trabalho, a questão direcional já é abordada pelo grau de entrada, pelo grau de saída e pelo k-core.

Já a autorreferência, A, criada especialmente para a configuração deste trabalho, mede a razão de conexões de usuários do local i que foram estabelecidas com outros usuários do mesmo local, L, pela quantidade total de conexões oriundas de usuários (ou seja, conexões tanto com usuários externos e internos ao local), T, conforme traduz a equação abaixo. Matematicamente, $A_i = L_i/T_i$. Nota-se que a autorreferência, ao contrário das demais métricas, é construída considerando-se as conexões dos usuários individuais e não as conexões entre locais. Isso porque os usuários conectados em autorreferência estão em um mesmo local, só sendo possível a abordagem citada.

4.3. Utilização Média do Discador como Medida de Sucesso de cada Local

O sucesso pode ser medido pela utilização do discador para acesso à Internet. Assim, para cada usuário, foi contabilizado o tempo de uso da Internet desde o seu cadastramento no provedor. Quanto mais tempo um usuário esteve conectado à Internet, mais aderente ele foi ao discador. Como aqui o foco está na avaliação dos locais e não de usuários, a utilização da informação foi definida como sendo a média do tempo de uso da Internet de todos os n usuários de um local i desde o cadastramento desses usuários. Assim, $(\overline{U})_i = \sum_{k=1}^{n} u_k/n$, onde $\overline{U}_i$ é a utilização média do discador de um local, em horas de conexão com a Internet, e u_k é a utilização do discador pelo usuário k, em horas de conexão com a Internet, desde o cadastramento do usuário.

Valor de rede, segundo Domingos (2005), é uma medida genérica que indica a contribuição holística do indivíduo para os objetivos da difusão. Ou seja, soma as

[2] Todas as métricas podem ser consultadas na referência, com exceção da autorreferência, criada exclusivamente para este trabalho.

medidas de contribuição individual (a utilização do discador por um determinado usuário, no caso em questão neste trabalho) e a influência do indivíduo na rede (soma das utilizações dos usuários da rede de influência direta e indireta do determinado usuário). Neste trabalho (seção 5.3), é utilizada uma métrica que utiliza este conceito, denominada aqui de Coeficiente de Valor Agregado (CVA), que é o quociente entre valor de rede de um indivíduo e a utilização do discador do mesmo.

Porém, nem todos os usuários foram considerados para compor a utilização média do discador dos locais, como ilustra a figura 2. Isso porque, para se coletar o tempo de uso da Internet por parte de um usuário, deve-se observá-lo após o seu cadastramento no provedor por um determinado período de tempo, de modo que os usuários que se cadastraram mais tarde não tem tempo hábil para a observação adequada.

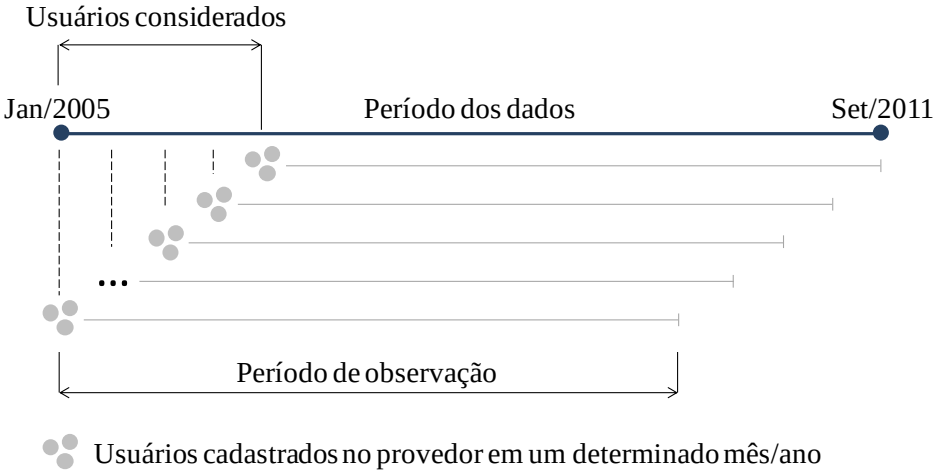

Figura 2. Ilustração da seleção de usuários para composição das medidas de utilização média do discador. Nota-se que se fosse aumentado o período de observação necessário, mais usuários seriam descartados, o que é ruim; porém, a utilização média do discador pelos selecionados seria medida com precisão, o que é bom, pois se pode observar praticamente toda a vida do indivíduo como usuário do serviço prestado pelo provedor. E que se fosse diminuído o período de observação, ocorreria o contrário: maior seria número de usuários considerados e menor a precisão na medição da utilização média do discador. Deve-se, então, buscar um período de observação que encontre uma relação adequada entre esses dois efeitos concorrentes.

O período adequado de observação foi estabelecido como sendo de 48 meses. Foram dois os motivos principais que motivaram a escolha desse período. Considerando que 60 meses foi praticamente o tempo de ciclo do produto do provedor, o primeiro motivo é que, dos usuários que potencialmente poderiam ser observados durante até esses 60 meses, 98% foi ativo até no máximo o 48º mês após o cadastro; já o segundo motivo é que, dos mesmos usuários, 99% do tempo total de uso da Internet foi consumido até o mesmo 48º mês de cadastro. Isso significa que a precisão na medida do uso da Internet é bastante confiável utilizando esse período.

Devido um problema de ordem técnica, o provedor perdeu os dados de uso da Internet de seus usuários referentes aos primeiros meses de operação do serviço (jan/2005 a jan/2006). Com isso e juntamente com o critério de se considerar períodos observados de 48 meses, os 132.304 usuários considerados (20% de toda a rede de

recomendações) para o fim de construir a utilização média da informação são os cadastrados de fev/2006 a jan/2007, um volume satisfatório.

4.4. Análise dos Resultados

Em síntese, a metodologia forneceu quatro elementos complementares: as redes geográficas compreendidas pelo agrupamento de usuários de toda a rede de recomendações original em seus locais de referência, a rede com os agrupamentos da abrangência do CEP de 3 dígitos, as métricas de rede que caracterizam cada CEP de 3 dígitos quanto ao fluxo de informação e uma métrica de sucesso para cada local, que é a utilização média do discador apenas dos usuários cadastrados entre fev/2006 e jan/2007. Da observação desses elementos em conjunto foram possíveis os resultados que seguem.

5. Resultados

5.1. Os Locais Apresentam Intensidades Diferentes de Utilização Média do Discador

Em todos os níveis de agrupamento geográfico construídos, foi verificada a existência de locais com maior utilização média do discador que outras (Tabela 1). Ou seja, de alguma forma, o discador foi mais adequado às necessidades ou às características do público dessas regiões espaciais, o que atesta a relevância da localização geográfica para a difusão de informação.

Os locais constituintes das redes foram divididos em quintis. Isto é, os locais foram divididos em cinco grupos, cada qual contendo aproximadamente 20% do total, ordenados segundo a utilização média do discador. O primeiro quintil com os locais de menor utilização média do discador e o quinto com os de maior utilização média. O motivo de o número de grupos serem cinco decorre do fato de as macrorregiões brasileiras (a menos granular das redes consideradas) serem cinco. Dessa forma, os demais níveis de agrupamento geográfico foram forçados a se dividir nesse número máximo de grupos para que se obtivesse o efeito de comparação direta entre eles.

Nota-se que quanto mais a geografia espacial é subdividida, maior é a amplitude de utilização média do discador verificada. A amplitude, no caso, é a diferença entre a utilização média do discador do melhor quintil e a do pior quintil. A amplitude da rede de CEP é de 302h, praticamente o dobro que a da rede de CEP de 3 dígitos (150h), que é 14% maior que a da rede de unidades da federação (131h), que, por sua vez, é 55% maior que a da rede de macrorregiões. Isso significa que quanto mais recortada a geografia espacial, mais apuradamente consegue-se separar locais “bons” (de maior utilização) de “ruins”.

Para ilustrar de maneira mais detalhada como se dá a diferenciação entre as localidades geográficas, os resultados que vem adiante consideram apenas o nível de CEP de 3 dígitos. Isso porque, como já foi citado na seção 4, esse nível é o mais utilizado pelos profissionais de geomarketing[3], por apresentar uma boa relação entre os níveis de recorte da geografia e de informação disponível.

[3] Vertente do marketing que trabalha a relação entre as características locais e o evento estudado (compra de um determinado produto, por exemplo).

Tabela 1. Diferenças de utilização média do discador entre os níveis estudados de agrupamento geográfico

Utilização média do discador [h]		Quintis					
		1	2	3	4	5	Amplitude
Agrupamentos	Macrorregião	14	22	69	98	99	85
	UF	6	23	67	73	137	131
	CEP de 3 dígitos	0	0	13	59	150	150
	CEP	0	0	0	12	302	302

A figura 3 mostra como se distribui a utilização média do discador pelos 987 locais de CEP de 3 dígitos com usuários representantes. E também estabelece como é a distribuição de usuários pelos mesmos locais, de forma que é possível estabelecer uma relação entre a qualidade (maior ou menor utilização média do discador) e a quantidade de usuários que cada local apresenta.

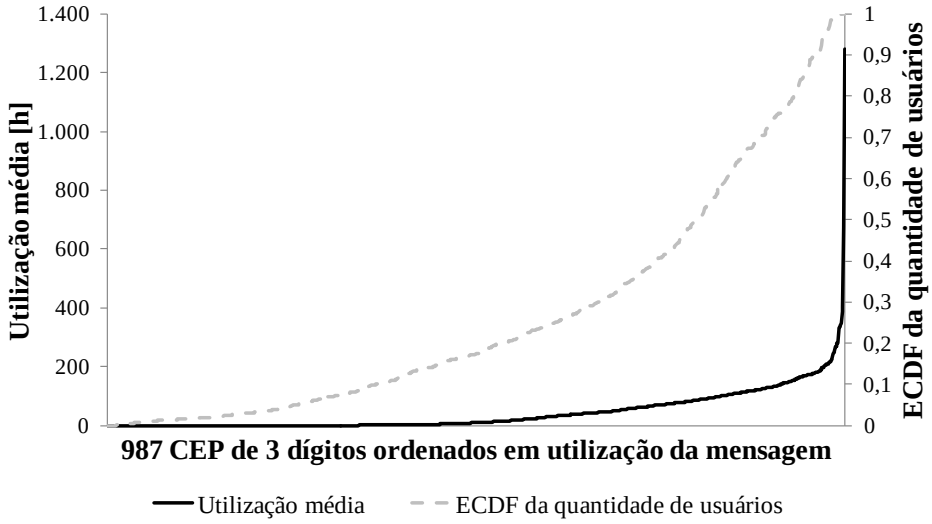

Figura 3. Relação entre utilização média do discador e a quantidade de usuários para a rede de CEP de 3 dígitos. Nota-se que a utilização média é nula até os dois primeiros quintos dos locais, quando começa a crescer de maneira moderada até atingir uma utilização média um pouco abaixo de 200h. No último décimo dos locais, o crescimento se torna muito mais agressivo, atingindo um pico de aproximadamente 1200h.

5.2. Os Locais Possuem Funções Diferentes

Foram feitas duas constatações principais no que se refere a funções que um local pode desempenhar no fluxo de informações. A primeira é que os locais de CEP de 3 dígitos de maior utilização média do discador são mais centrais que outros e a segunda é que são relativamente mais ativos que passivos em relação a conexões. Esta seção aborda as duas constatações, uma por vez.

Para a questão da centralidade, as seguintes métricas de rede foram escolhidas para a investigação dos locais: grau de entrada, grau de saída, k-core (considerando só conexões de entrada), k-core (considerando só conexões de saída) e coeficiente de aglomeração. Todas as métricas consideram o grafo direcionado como de fato o é, com exceção do coeficiente de aglomeração. A figura 4 mostra a relação entre a utilização

média do discador e o grau de entrada de um local de CEP de 3 dígitos, indicando uma correlação direta.

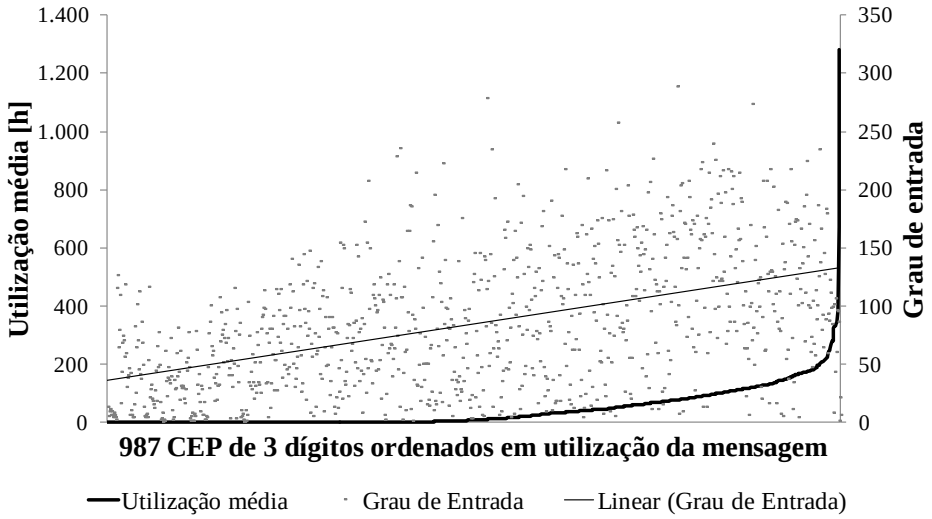

Figura 4. Relação entre utilização média do discador e o grau de entrada do local. Uma linha de tendência foi traçada por uma regressão linear à nuvem de pontos, pares ordenados da ordem dos CEP de 3 dígitos pelo grau de entrada. O coeficiente de correlação de Pearson verificado é relevante, de 0,48. A linha de tendência tem um valor inicial de 36,6 e um valor final de 133,3. Observa-se que existe uma tendência de que os CEP de 3 dígitos com melhor utilização média do discador tenham maior grau de entrada, ou ainda, sejam mais acionados por convites oriundos de outros locais. Isso significa que os locais de melhor utilização média do discador tendem a ser atratores do fluxo de conexões, corroborando para a indicação de que ocupam uma posição mais central na rede.

Para as demais métricas citadas, o procedimento é análogo ao utilizado para grau de entrada, e os principais resultados estão na tabela 2, onde os coeficientes de correlação também são relevantes. Nela, também se percebe que quanto maior a utilização média do discador em um local, existem tendências de que também sejam maiores o grau de saída e os k-cores (de entrada e de saída). Quanto ao grau de saída, isso significa que os locais de CEP de 3 dígitos com maior utilização tendem a ser difusores do fluxo de conexões para outros locais de CEP de 3 dígitos. Considerando o efeito conjunto dos resultados sobre os graus de entrada e de saída, ao serem atratores e difusores do fluxo de conexões ao mesmo tempo, podemos concluir que esses CEP de 3 dígitos são mais que apenas referências estáticas, são "processos por onde a informação flui" conforme sugere Manuel Castells (1996-98). A tendência ascendente verificada nos k-cores ratifica a condição mais central que os locais de maior utilização média do discador exercem com maior frequência, tanto no fluxo de conexões de fora para dentro (quando as conexões se originam em outros locais e se destinam para o local de referência) quanto de dentro para fora.

Tabela 2. Relação entre a utilização média do discador e grau de saída, k-core e coeficiente de aglomeração

	Menor valor da regressão linear[4]	Maior valor da regressão linear	Coef. de correl. Pearson
Grau de saída	-26,2	196,0	0,40
K-core (entradas)	38,0	78,5	0,42
K-core (saídas)	-1,5	48,5	0,59
Coeficiente de aglomeração	0,92	0,63	-0,52

Quanto ao coeficiente de aglomeração, observa-se que ele apresenta a tendência de ser menor quanto maior a utilização média do discador em um CEP de 3 dígitos. Isso significa que os locais de maior utilização média do discador tendem a intermediar a conexão outros CEP de 3 dígitos que não são conectados diretamente. Por exemplo, toma-se um lugar periférico A, de menor utilização. Ele é conectado com alguns outros do mesmo tipo, que, por não serem muito conectados, provavelmente não são conectados diretamente entre si. Mas foi visto que existe uma tendência de centralização de conexões de e para os locais de maior utilização média do discador, de modo que um número significativo dos locais de menor utilização (inclusive A) tem conexões com locais de maior utilização. Assim, a presença dos locais de maior utilização aumenta o número de "triangulações", aumentando o coeficiente de aglomeração local de A. Já no exemplo contrário, toma-se um lugar central B, de maior utilização média do discador. Ele é conectado principalmente com locais de menor utilização, que não são muito conectados entre si, o que, por sua vez, acarreta em um coeficiente de aglomeração mais baixo.

Em termos práticos para os profissionais de difusão, isso significa que atingir um local de maior utilização média do discador significa atingir vários outros locais, pois as conexões (originadas das recomendações) fluem para esses outros locais. A influência se expande no espaço físico rapidamente. Já atingir um local de menor utilização vai gerar menor área de influência de forma direta. Existe a possibilidade real e comum de um local de menor utilização se conectar a outro de maior utilização e iniciar o processo mais rápido de difusão espacial, mas evidentemente a probabilidade é menor. Assim, quando o objetivo for difundir uma informação com rapidez, aconselha-se aos profissionais de difusão que abordem diretamente os locais de maior utilização.

A segunda constatação desta seção é demonstrada na tabela 3: os locais de CEP de 3 dígitos de maior utilização média do discador são mais ativos que passivos em relação a recomendações, interna e externamente. Ser mais ativo internamente significa que os usuários do local se conectam mais entre si que em outros locais. Essa é mais uma força dos locais de maior utilização: os usuários desses locais tendem a estimular outros usuários do mesmo local. Relembrando, a métrica da autorreferência mede o percentual de conexões ativas de usuários do local que foram estabelecidas com outros usuários do mesmo local. Nota-se que existe a tendência crescente da autorreferência conforme a utilização cresce e que a correlação é significativa (0,51).

[4] As métricas avaliadas não permitem valores negativos. Os valores negativos observados são decorrentes da regressão linear.

Já ser mais ativo externamente significa que um determinado local, em sua relação com os demais locais, possui maior valor do quociente entre grau de saída e grau de entrada. O grau de saída indica a atividade, a orientação da influência para fora, para outros locais, ou seja, quantos locais o determinado local influencia. O grau de entrada indica a passividade, a orientação da influência para dentro, ou seja, quantos lugares influenciam o determinado local. Assim, observa-se que os CEP de 3 dígitos de maior utilização média são mais orientados para fora que outros. Os 222 CEP de três dígitos de menor utilização apresentaram valor 0, devido a não possuírem grau de saída.

Tabela 3. Relação da utilização da informação com autorreferência e balanço atividade / passividade

	Menor valor da regressão linear	Maior valor da regressão linear	Coef. de correl. Pearson
Autorreferência	0,5%	10,3%	0,51
Balanço atividade / passividade	0,07	1,45	0,38

5.3. Utilização Média do Discador e Valor de Rede são Diretamente Relacionados

Os usuários de locais da rede de CEP de 3 dígitos com maior utilização média do discador são também os que atingem maior valor de rede médio. Isso é verificado na figura 5, que mostra a tendência crescente da média local do coeficiente de valor agregado (CVA). Essa é mais uma vantagem dos locais de maior utilização média: além de os usuários utilizarem mais o discador, fazem com que outros o utilizem por influência.

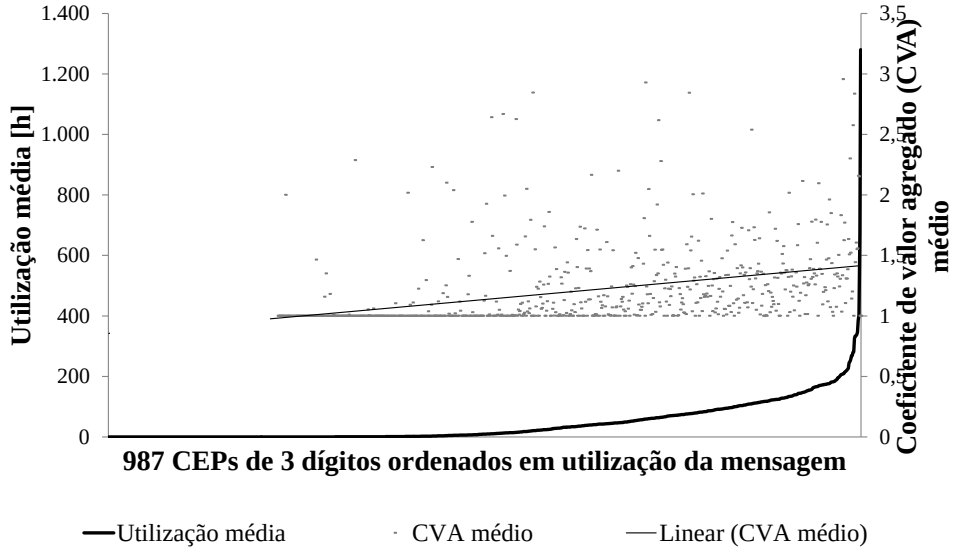

Figura 5. Relação entre utilização média do discador e o poder de influência dos usuários. Nota-se que os primeiros 221 CEP de três dígitos não possuem valor válido do quociente de valor de rede e de utilização individual, devido possuírem valor zero no denominador, a utilização individual. Dentre os pontos restantes, foram excluídos os *outliers*, convencionados como sendo os de valor do quociente superior a três, que corresponde ao 95º percentil. O coeficiente de correlação de Pearson verificado é relevante, de 0,39. A linha de tendência tem

um valor inicial de 0,99, no primeiro CEP de três dígitos válido, e um valor final de 1,45.

5.4. Discussão dos Resultados

Dos resultados apresentados até aqui e tendo em vista a aplicação prática, sugere-se que seja considerada a utilização das estatísticas públicas, que caracterizam as localidades, sob o pretexto de que conhecer o usuário melhora a informação e impulsiona a difusão. Neste contexto, de maneira pragmática, recomenda-se aos profissionais correlatos a captação do CEP dos usuários no momento da adesão destes à informação, ao menos para uma amostra dos cadastrados, de modo que se consiga identificar a importância de cada local, em função e intensidade. O CEP é uma fonte de informação de baixo custo, relevante e desburocratizante. De baixo custo porque através dele tem-se acesso às informações das estatísticas públicas, que em geral são gratuitas. Relevante devido ao que é aqui demonstrado sobre a relação entre redes de recomendações online e a localização espacial dos usuários. Desburocratizante porque é um dado que traz muita informação agregada, evitando que o usuário tenha de preencher extensos formulários para a adesão ao objeto da difusão. O preenchimento de um único campo, o CEP, permite acessar uma ampla gama de informações demográficas, socioeconômicas e de infraestrutura. No momento da adesão, o objetivo do profissional de difusão é obter o máximo de informação relevante do usuário fazendo com que este tenha que fornecer o menor volume de dados possível. A ideia é que quanto mais fácil para o usuário for o processo de adesão, maior é a chance de sucesso.

6. Conclusão

O presente trabalho indica, através do estudo de um caso real, que a localização geográfica é importante para a difusão de informação em redes sociais online, ao menos para o caso real apresentado. E alerta os profissionais de difusão online para a subestimação do fator localização, que, se adequadamente considerado, poderia proporcionar maior assertividade de público-alvo e maior eficiência de difusão. De uma forma mais precisa, foi visto que os locais físicos possuem intensidades de utilização da informação e funções muito diferentes entre si. Com relação às funções que as localidades podem desempenhar, notou-se que as de maior utilização média do discador são mais ativas, implicando que utilização e influência são diretamente relacionadas. E também que são mais centrais que as demais, implicando que são centros por onde a informação flui, centros conectores de locais de menor utilização e, ainda, gatilhos facilitadores de maior cobertura territorial.

Aqui foi sugerido que existem diferenças entre locais físicos no que tange difusão online de informação e foi demonstrado como ocorre o fluxo dessa difusão. O próximo passo seria verificar se as conclusões aqui apresentadas são consistentes em outros casos reais, afim de que possam ser generalizadas. Outro próximo passo seria explorar de fato as estatísticas públicas, ou seja, descobrir que características conseguem diferenciar locais que aderem à informação de outros que não aderem, e mesmo entender porque alguns lugares tem maior intensidade de utilização ou influência que outros. Uma descrição completa do fenômeno deveria envolver a investigação das características demográficas, socioeconômicas e de infraestrutura, advindas diretamente das estatísticas públicas, e das características de rede, já abordadas aqui. Estas características devem ser observadas de forma separada e também

conjuntamente, quando então deve ser percebido como elas se combinam a gerar o comportamento da rede. Dessa forma, os profissionais de difusão de informação conhecerão ainda melhor os seus usuários e poderão tornar a informação ainda mais adequada e a difusão ainda mais eficiente.

Referências

Adamic, L. A., Adar, E. (2003), "How to Search a Social Network", HP Labs 1501 Page Mill Road, Palo Alto, CA 94301.

Bampo, M, Ewing. M. T., Mather, D. R., Stewart, D. e Wallace, M. (2008), "The Effects of the Social Structure of Digital Networks on Viral Marketing Performance", Information Systems Research, Vol. 19, No. 3, p. 273-290, issn 1047-7047, eissn 1526-5536 08 1903 0273.

Castells, M. (1996-98), "The Information Age: Economy, Society and Culture", 3 volumes, Oxford.

Centola, D. (2010), "The Spread of Behavior in an Online Social Network Experiment", Science 329, 1194, DOI: 10.1126/science.1185231.

Domingos, P. (2005), "Mining Social Networks for Viral Marketing", IEEE Intelligent Systems, v. 20, n. 1, p. 80-82.

Hill, S., Provost, F. e Volinsky, C. (2006), "Network-based Marketing: Identifying Liking Adopters via Consumer Networks", Statistical Science, Vol. 21, No. 2.256, p. 256-276.

Kaltenbrunner, A., Scelato, S., Volkovich, Y., Laniado, D., Currie, D., Jutemar, E. J. e Mascolo, C. (2012), "Far from the Eyes, Close on the Web: Impact of Geografic Distance on Online Social Interactions", WOSN 12, ACM 978-1-4503-1480-0/12/08.

Ma, H., Zhou, T. C., Lyu, M. R. e King, I. (2011), "Improving Recommender Systems by Incorporating Social Contextual Information", ACM Trans. Inf. Syst. 29, 2, Article 9, http://doi.acm.org/10.1145/1961209.1961212.

Marques, E. C. L. (2009), "As Redes Sociais Importam para a Pobreza Humana?", Departamento de Ciência Política USP e Centro de Estudos da Metrópole do CEBRAP.

Minhano, R. Q., Fernandes, S. e Kamienski, C.A. (2010), "Revealing Hidden Connections in Recommendation Networks Using Online Social Networks", Santo André, http://hostel.ufabc.edu.br/~cak/hidden_connections_2010.pdf.

Newman (2003), M. E. J., "The Structure and Function of Complex Networks", SIAM Review, v. 45, n. 2, p. 167-256.

Yang, J. e Jones, B. (2010), "A Study of the Spreading Scheme for Viral Marketing Based on Complex Network Model", Physica A 389, p. 859-870.