Uma Análise Contextual de Conteúdo Duplicado no YouTube *

Tiago Rodrigues
UFMG, Belo Horizonte/Brasil
tiagorm@dcc.ufmg.br

Fabrício Benevenuto
UFMG, Belo Horizonte/Brasil
fabricio@dcc.ufmg.br

Virgílio Almeida
UFMG, Belo Horizonte/Brasil
virgilio@dcc.ufmg.br

Jussara Almeida
UFMG, Belo Horizonte/Brasil
jussara@dcc.ufmg.br

Marcos Gonçalves
UFMG, Belo Horizonte/Brasil
mgoncalv@dcc.ufmg.br

ABSTRACT

Videos have become a predominant part of users' daily lives on the Web, especially with the emergence of online video social networks such as YouTube. Since users can independently share videos in these systems, some videos can be duplicates (i.e., identical or very similar videos). Despite having the same content, there are some potential differences in duplicates, for example, in their associated metadata (i.e., tags, title) and their popularity scores (i.e., number of views, comments). Quantifying these differences is important for three reasons. The first is related to the necessity of understanding how users associate metadata to videos on YouTube, which is crucial for video information retrieval mechanisms and recommendation systems. The second is associated with understanding possible reasons that influence on the popularity of videos, essential to the association of advertisements to videos and performance issues related to the use of caches and CDNs. The third comes from the necessity to detect opportunistic actions, which pollute and compromise the use of the system. This work presents a wide characterization of the differences among identical contents in online video sharing systems. Using a large video sample collected from YouTube, we construct a data set of duplicates. Besides quantifying contextual differences among duplicates, our results also reveal the presence of suspect behavior in the creation and association of metadata to videos.

RESUMO

Vídeos se tornaram uma parte predominante da vida diária dos usuários da Web, especialmente com o surgimento de redes sociais de compartilhamento de vídeos como o YouTube. Como usuários podem compartilhar vídeos independentemente nesses sistemas, alguns vídeos podem ser duplicados (vídeos idênticos ou muito similares). Apesar de duplicatas possuírem o mesmo conteúdo, existem potenciais diferenças em seus metadados (tags, categorias, etc.) bem como em seus indicadores de popularidade (número de visualizações, avaliações, etc.). Quantificar essas diferenças é importante por três motivos. O primeiro está relacionado à necessidade de se entender como usuários associam metadados a vídeos no YouTube. Tal entendimento é importante para máquinas de busca e sistemas de recomendação. O segundo está associado ao entendimento de possíveis motivos que influenciam na popularidade dos vídeos, essencial para a associação de propagandas a vídeos e questões de desempenho ligadas ao uso de caches e CDNs. O terceiro vem da necessidade de se detectar ações oportunistas e maliciosas, que poluem e comprometem o uso do sistema. Este trabalho apresenta uma ampla caracterização das diferenças entre duplicatas em redes sociais de compartilhamento de vídeos. Utilizando uma ampla amostra de vídeos do YouTube, construímos uma grande base de dados de duplicatas. Além de quantificar várias diferenças entre as duplicatas de nossa base de dados, nossas análises revelam a existência de comportamento malicioso de alguns usuários na criação de conteúdo duplicado.

WebMedia'09, October 5-7, 2009, Fortaleza, CE, Brazil.

Categories and Subject Descriptors

J.4. [**Computer Applications**]: :Social and behavioral sciences Miscellaneous;; H.3.5 [**Information Storage and Retrieval**]: :Online Information Services-Web-based services

General Terms

Human factors, Measurement

Keywords

duplicates, metadata association, social network, videos

1. INTRODUÇÃO

Vídeos vêm ocupando cada vez mais espaço na *Web* e no dia a dia de seus usuários. Com a crescente popularidade de redes sociais de compartilhamento de vídeos, usuários se tornaram verdadeiros canais para distribuição de vídeos, compartilhando milhões de novos vídeos todos os dias. Como um exemplo, de acordo com a empresa comScore, 74% de toda a audiência da Internet norte americana assistiu a vídeos online em maio de 2008, sendo que 12 bilhões de vídeos foram exibidos em apenas 1 mês. O YouTube, um dos pioneiros no compartilhamento de vídeos, proveu sozinho 34% desses vídeos [9].

Como usuários podem compartilhar vídeos de forma independente em sistemas como o YouTube, vários vídeos podem ser duplicatas, ou seja, vídeos de conteúdo idêntico ou bastante similar. A existência de duplicatas é um problema em vários outros sistemas, como por exemplo na *Web* [17], em blogs [1] e em sistemas de compartilhamento de imagens [25]. Duplicatas podem trazer diversos tipos de problemas para o sistema. Como um exemplo, é indesejável obter como resultado de uma busca por vídeos apenas uma lista de vídeos idênticos. Em termos de armazenamento, espaço de disco poderia ser economizado mantendo-se apenas uma das cópias do conteúdo no sistema. Além disso, duplicatas podem dividir a audiência de um determinado conteúdo, o que pode

impactar no desempenho de caches e redes de distribuição de conteúdo (CNDs) [8]. Neste contexto, abordagens para identificação e agrupamento de duplicatas foram implementadas no YouTube [10] e recentemente propostas pela comunidade científica [24].

Entretanto, a existência de grupos de duplicatas abre espaço para o entendimento de questões importantes de sistemas de compartilhamento de vídeos. Apesar de possuírem conteúdos iguais, existem várias potenciais diferenças contextuais entre duplicatas, como por exemplo, em seus metadados (*tags*, categorias, etc.) e em seus indicadores de popularidade (número de visualizações, avaliação, etc.). Quantificar tais diferenças é de fundamental importância por três motivos. O primeiro está relacionado à necessidade de se entender como usuários associam metadados a vídeos no YouTube, essencial para máquinas de busca e sistemas de recomendação. O segundo está associado ao entendimento de possíveis motivos que influenciam na popularidade dos vídeos, importante para a associação de propagandas a vídeos. O terceiro vem da necessidade de se verificar a existência de ações oportunistas muitas vezes imperceptíveis para o sistema e para os demais usuários.

Este trabalho aborda as diferenças contextuais existentes entre vídeos duplicados. Através de uma amostra baseada em buscas realizadas no YouTube, criamos uma grande coleção de vídeos considerados como duplicatas pelo YouTube. Com base nesses dados, apresentamos uma ampla caracterização de diversas diferenças entre duplicatas em termos de seus metadados, características de suas popularidades e características de seus donos. Além de quantificar tais diferenças e revelar possíveis fatores que influenciam na popularidade de vídeos, nossos resultados também revelam a existência de comportamento malicioso por parte dos usuários na criação intencional de duplicatas.

O restante deste trabalho está organizado da seguinte forma. A próxima seção discute trabalhos relacionados. A seção 3 descreve como coletamos e criamos uma coleção de duplicatas a partir do YouTube. A seção 4 caracteriza diferenças entre vídeos com conteúdos iguais. A seção 5 investiga possíveis ações maliciosas associadas a criação de duplicatas. A seção 6 apresenta um resumo das nossas principais descobertas e contribuições enquanto a seção 7 conclui o artigo e oferece direções para trabalhos futuros.

2. TRABALHOS RELACIONADOS

Vários trabalhos recentes caracterizaram diversos aspectos de sistemas de compartilhamento de vídeos, especialmente do YouTube. Em particular, a referência [11] apresenta uma caracterização do tráfego do YouTube do ponto de vista do campus de uma universidade e compara suas propriedades com propriedades do tráfego da *Web* e de servidores de vídeos. Em [26], os autores também caracterizam o tráfego coletado do campus de uma universidade. Baseado nos dados coletados, eles realizam simulações e mostram que *caching* de vídeos tanto no cliente quanto no *proxy* e distribuição P2P podem reduzir tráfego de rede consideravelmente e permitir acesso mais rápido a vídeos em sistemas de compartilhamento de vídeos. Em [24], os autores propõem um mecanismo para filtrar duplicatas de resultados de busca. Eles criaram uma coleção de duplicatas baseados em 24 buscas no YouTube, Google Vídeo e Yahoo! Vídeo. Utilizando um algoritmo de clusterização hierárquica, eles detectaram grande parte das duplicatas dentre os resultados de busca por vídeos.

Outro importante estudo sobre sistemas de compartilhamento de vídeos é apresentado em [8]. Os autores analisam a distribuição de popularidade, evolução, e características dos vídeos do YouTube e de um sistema de compartilhamento de vídeos Coreano. Além disso, eles apresentam as primeiras evidências sobre a existência de duplicatas através de uma pequena base de duplicatas do YouTube, construída manualmente. Através de análises sobre a criação das duplicatas eles discutem os potenciais problemas que duplicatas podem causar ao sistema.

Nosso trabalho é diferente de todos esses trabalhos anteriores já que caracterizamos as diferenças contextuais (metadados, popularidade) existentes entre vídeos com conteúdo idêntico, sendo complementar a esses esforços.

O trabalho [23] quantifica duas suposições comuns sobre social *tagging* em páginas *Web*: que tags são significativas e que o processo de *tagging* é influenciado por sugestões. Como resultado mostram que as *tags* mais populares de uma página tendem a ser as mais significativas. Outro estudo relacionado é apresentado em [20], que propõe uma abordagem para o descobrimento de interesses sociais baseados nas *tags* geradas pelos usuários. Baseados na análise de uma grande quantidade de dados reais coletados do del.icio.us[1], mostraram que, no geral, as *tags* geradas pelos usuários são consistentes com o conteúdo *Web* que elas estão associadas. Nosso trabalho é complementar a esses dois no sentido de que quantificamos a similidade com que os usuários associam metadados a vídeos de mesmo conteúdo, mostrando que há diferenças de percepção dos usuários e motivando a busca de novas alternativas na recuperação de informação em conteúdo multimídia.

Em trabalhos anteriores, as propriedades da rede social formada por interações de vídeos resposta no YouTube são analisadas, revelando a existência de vídeo spammers, usuários que postam vídeos resposta não relacionado ao tópico discutido na tentativa de promover um conteúdo [4]. A referência [5, 6] aborda o problema de identificar os vídeo spammers utilizando-se uma abordagem de classificação. Nosso trabalho revela a existência de outros tipos de poluição e comportamento malicioso em relação à criação de duplicatas. Um arcabouço teórico para se detectar *spam* em sistemas de *tagging*, um tipo de comportamento malicioso que procura aumentar a visibilidade de um objeto enganando mecanismos de busca com *tags* não relacionadas foi proposto em [18]. Os autores propõem métricas para avaliar o nível de poluição num sistema de *tagging*. Nossos estudos revelam a presença de *tag-attack* em um sistema real, mostrando propriedades de como as *tags* são associadas aos vídeos utilizados no ataque.

3. COLETA DOS DADOS

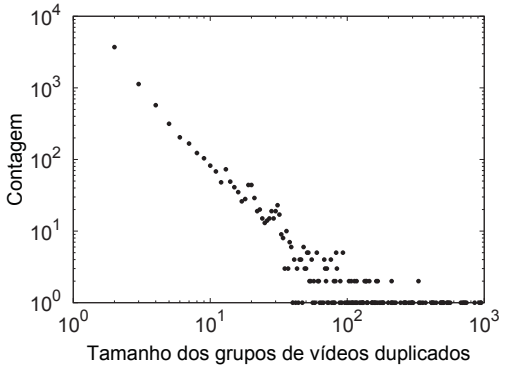

Figura 1: Tamanho dos Grupos de Duplicatas

Para estudarmos as características dos vídeos com conteúdo duplicado foram coletados dados de vídeos do YouTube, um dos mais populares sistemas sociais de compartilhamento de vídeos criados por usuários[2]. A estratégia utilizada para coletar duplicatas é baseada em buscar palavras aleatórias no YouTube e coletar os vídeos que aparecem nos resultados das buscas. Ao mostrar o resultado de uma busca, o YouTube filtra as duplicatas e exibe apenas um vídeo. Entretanto, o YouTube disponibiliza um elo para as duplicatas de cada vídeo, excluídas dos resultados da busca[2]. Nosso

[1] www.delicious.com

[2] Recentemente, o YouTube retirou o elo para as duplicatas, mas

coletor segue esses elos, coletando informações de todos os vídeos encontrados. No decorrer do artigo, chamamos de grupo de duplicatas cada conjunto de vídeos idênticos ou similares agrupados pelo YouTube e disponibilizados através desses elos.

Algoritmo 1 Coletor de Conteúdo Duplicado

```
1: Obtenha uma palavra W do servidor
2: Busca no YouTube usando W
3: for each vídeo v que aparece nos resultados da busca do
4:     if v tem uma lista de duplicatas DV then
5:         for each vídeo d em DV do
6:             Coleta informações de d
7:             Coleta informações do usuário dono de d
8:             Coleta informações de todos os vídeos do dono de d
9:         end for
10:    end if
11: end for
```

Nosso coletor foi construído de forma distribuída (um servidor e dez clientes), seguindo o algoritmo 1. O servidor seleciona aleatoriamente palavras de um dicionário em inglês, obtido a partir de uma ferramenta para correção ortográfica chamada *ispell*[14], e passa as palavras para as máquinas clientes, que realizam buscas no YouTube e coletam informações de todos os vídeos encontrados, suas duplicatas e de seus donos.

Após executar por uma semana, nosso coletor encontrou mais de 100 mil duplicatas, de 80.297 usuários, agrupadas em 9.178 grupos. Notamos que alguns grupos de duplicatas possuem vídeos com durações diferentes (ex: uma versão completa de um vídeo e outro com somente algumas das cenas foram considerados como duplicatas pelo YouTube). Então, como nosso objetivo é ter um conjunto de dados com grupos de vídeos com conteúdo idêntico, filtramos vídeos com a duração diferindo em mais de 2% da duração mediana do grupo. No total filtramos 31.709 duplicatas, reduzindo o número de grupos para 7.330 já que alguns grupos tiveram todas as duplicatas filtradas. A tabela 1 apresenta um sumário de todos os dados coletados e dos dados filtrados.

	Coletado	Após Filtragem
Período da Coleta	24/05/2008 - 31/05/2008	-
# palavras buscadas	319	-
# grupos de duplicatas	9.178	7.330
# duplicatas	100.373	68.664
# usuários donos de duplicatas	80.297	58.922
# vídeos coletados	1.884.611	1.321.407

Tabela 1: Sumário dos Dados Coletados

A figura 1 mostra a contagem do tamanho dos grupos presentes em nossa base de dados. Podemos ver que a maior parte dos grupos são de tamanho pequeno. Como um exemplo, 51% dos grupos possuem somente 2 duplicatas e 87% possuem menos de 10 vídeos. O maior grupo possui 947 duplicatas.

Como nossa abordagem para criar a base de duplicatas confia no algoritmo do YouTube para detectar duplicatas, é possível que existam alguns vídeos que não sejam realmente duplicados. Para verificar a acurácia com que o YouTube define um vídeo como duplicata, foram selecionados aleatoriamente 154 grupos de vídeos para inspeção "manual". No total, foram assistidos 1.059 vídeos, sendo que o tamanho da amostra selecionada foi definido de forma a permitir uma confiança de 95% no resultado da verificação [15].

Para minimizar o impacto do erro humano, três voluntários foram utilizados para classificar os grupos de duplicatas em corretos ou incorretos. Todos os grupos foram analisados e independentemente classificados por dois voluntários, sendo que o terceiro foi utilizado para os casos em que as classificações diferiram. Para a classificação consideramos a mesma definição de conteúdo duplicado apresentada em [24], que diz que vídeos com pequenas diferenças, tais como alterações de cor e qualidade, operações de edição e pequenas diferenças em tamanho, são considerados duplicatas. Em casos de dúvida na classificação, os voluntários foram instruídos a classificar o vídeo como incorreto, ou seja, utilizamos uma estratégia conservadora. Com 95% de confiança, apenas 5 $\pm$ 3.4% dos vídeos foram classificados incorretamente como duplicatas pelo YouTube. Sendo assim, nas seções seguintes, assumimos que a porcentagem de vídeos classificados erroneamente como duplicatas pelo YouTube não é suficiente para interferir nas análises.

permite que os usuários repitam suas buscas incluindo as duplicatas nos resultados

4. DIFERENÇAS CONTEXTUAIS ENTRE DUPLICATAS

Duplicatas, apesar de possuírem conteúdos idênticos ou muito semelhantes, podem possuir diferenças em diversos aspectos, tais como em seus metadados (ex. *tags* e categoria), nas estatísticas de acesso aos vídeos (ex. número de visualizações e avaliações) ou na idade do vídeo no sistema. Além disso, diferenças nas características dos donos de duplicatas podem influenciar na popularidade desses vídeos. Esta seção discute e analisa essas diferenças, as quais definimos como diferenças contextuais.

4.1 Indicadores de Popularidade

Inicialmente, vamos analisar as avaliações das duplicatas. No YouTube, usuários cadastrados podem avaliar um vídeo após assistí-lo. As notas da avaliação variam entre 1 e 5. Vídeos sem avaliações não foram considerados nessa análise. Para cada duplicata dentro de um grupo calculamos a diferença absoluta entre as médias de avaliações com os demais vídeos do grupo. Intuitivamente, espera-se que tais diferenças não sejam muito discrepantes, uma vez que os vídeos são iguais. A figura 2 (esquerda) mostra a função de distribuição cumulativa (CDF) dessas diferenças. Podemos notar que 23% dos pares de avaliações não diferem e que 95% dos pares diferem com valores inferiores a 2. Isso mostra que a maior parte dos vídeos com conteúdos iguais são avaliados de forma semelhante pelos usuários. Entretanto, há uma pequena fração dos pares de avaliações (cerca de 3%) que apresentam diferenças superiores a 3. Essas altas diferenças refletem as diferentes percepções dos usuários em relação ao mesmo conteúdo. Interessantemente, essa mesma observação foi feita em sistemas de avaliação e recomendação de filmes [12]. Certamente, um filme considerado bom para um usuário pode ser considerado ruim para outro.

O número de comentários textuais postados a um vídeo pode ser um indicador tanto da popularidade quanto da qualidade de um vídeo. A intuição por trás dessa análise é que um vídeo bastante discutido é um vídeo popular e de qualidade. A figura 2 (meio) mostra a CDF da diferença entre o número de comentários de todos os pares de vídeos formados dentro de cada grupo. Podemos notar que 54% dos pares de vídeos comparados possuem diferenças no número de comentários recebidos divergindo em no máximo 1 (36% deles não diverge). Interessantemente, 3% dos pares diferem em mais de 100 comentários, indicando que existem algumas duplicadas que adquirem mais popularidade do que outras. A maior diferença registrada consiste em 288.885 comentários.

Por último, investigamos as diferenças em termos do número de exibições dos vídeos. De maneira semelhante às análises anteriores, fazemos as diferenças entre todos os pares de vídeos formados dentro de cada grupo. A figura 2 (direita) mostra a CDF dessas diferenças. Apenas 21% dos pares diferem em menos do que 100 exibições e cerca de 46% diferem em mais de 1000 exibições.

De maneira geral, podemos notar a partir dessas análises que, apesar de possuírem conteúdos iguais, existem vídeos muito mais populares do que outros dentro do mesmo grupo. Na seção 4.3 discutimos sobre alguns possíveis fatores que influenciam na popularidade das duplicatas.

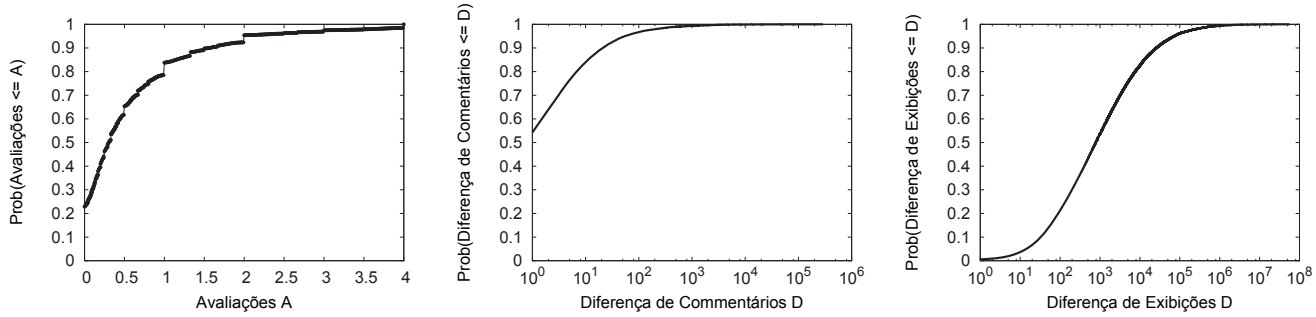

Figura 2: Distribuição de Diferenças entre Pares de Duplicatas para Avaliações (esq.), Comentários (meio) e Visualizações (dir.)

4.2 Diferenças temporais

Outra diferença contextual interessante entre as duplicatas é relacionada à idade do vídeo no sistema. Na figura 3 (esq.) mostramos a diferença de idade das duplicatas em relação ao primeiro vídeo do grupo que foi postado no sistema, o qual chamamos de vídeo original. Podemos observar que há um número significativo de cópias (cerca de 7.000) que são postadas em até no máximo 7 dias depois do vídeo original, mas também encontramos cópias postadas mais de 2 anos depois. Na figura 3 (meio) mostramos a CDF do tempo entre postagens sucessivas de uma mesma duplicata. 39% das duplicatas são inseridas no sistema no mesmo dia que a última cópia do grupo. Observamos também que cerca de 83% foram inseridas num intervalo menor que 30 dias depois que a última cópia foi inserida, enquanto apenas 1% surgiu mais de 1 ano depois que a última duplicata do grupo foi postada. Esses gráficos mostram que a maior parte das duplicatas são postadas em um curto período de tempo, logo após a criação da primeira versão do conteúdo.

Uma questão interessante de ser analisada é a popularidade dos vídeos em função da idade deles no sistema. Na figura 3 (dir.) mostramos a razão do número de visualizações de uma cópia pelo número de exibições do vídeo original como uma função da diferença de idade entre cada cópia e seu respectivo vídeo original. A reta do eixo y = 1 separa duas regiões distintas: a região das cópias mais populares do que os vídeos originais (acima da reta) e as demais cópias (abaixo)[3] No total, 17% das cópias superam a popularidade do vídeo original. Além disso, pouco mais de 2% das cópias não só são mais populares do que o vídeo original, mas também são as mais populares de seus grupos de duplicatas. Algumas dessas cópias ficaram mais populares mesmo tendo sido criadas cerca de 600 dias após a cópia original.

Uma questão muito interessante que surge dessas análises é entender que fatores podem levar um vídeo ficar mais popular do que outro. Na próxima seção analisamos se as características e ações do dono da duplicata podem influenciar na popularidade da mesma.

4.3 Donos das Duplicatas

Intuitivamente, usuários diferentes que criam vídeos iguais possuem características completamente diferentes. Sendo assim, a questão interessante a se abordar em relação aos donos de duplicatas é se características do dono do vídeo influenciam na popularidade das duplicatas. Em sistemas como o Flickr[4], muitos usuários acessam conteúdo através de elos da rede social [19]. Como conseqüência, temos uma grande correlação entre conexões entre usuários na rede social e popularidade dos objetos dos usuários. No caso do YouTube, nossa intuição é que usuários que possuem perfis populares (não só relações de amizade) tendem a ter seus vídeos mais exibidos dentro do grupo de duplicatas.

Para verificarmos se essa tendência existe, consideramos o *ranking* das duplicatas dentro de cada grupo em termos do número de exibições e calculamos a correlação desse *ranking* com o *ranking* dos donos de duplicatas segundo seis diferentes características: (1) Número de exibições dos vídeos do usuário; (2) Número total de comentários recebidos pelos vídeos do usuário; (3) Soma das avaliações recebidas pelos vídeos do usuário; (4) Número de vezes que os vídeos do usuário foram adicionados como favoritos no sistema; (5) Número de amigos do usuário; e (6) Número total de vídeos do usuário. Como um exemplo, para um grupo com 2 vídeos, temos correlação $C = 1$ se o dono do vídeo mais exibido é o usuário mais popular (em termos de alguma das 6 características acima) e $C = -1$, caso contrário. Note que, grupos pequenos tendem a ter uma variabilidade maior na correlação, o que sugere que nossas análises levem em consideração o tamanho do grupo. A seguir discutimos em detalhes a influência de cada uma das características do perfil do usuário na popularidade de uma duplicata.

A figura 4 (esq.) considera o *ranking* dos usuários pelo número de exibições dos demais vídeos do dono da duplicata. Intuitivamente, se outros vídeos do usuário fazem ou fizeram sucesso no YouTube, usuários tendem a ter algum grau de fidelização com os novos vídeos criados pelo mesmo usuário. De fato, cerca de 79% dos grupos possuem correlação positiva. Considerando-se apenas os grupos com mais de dez duplicatas, 92% possuem correlação positiva, sendo 75% maior do que 0,2. O número de comentários, a soma das avaliações e o número de vezes que os demais vídeos do dono da duplicada foram adicionados como favoritos são também métricas que indicam a qualidade dos vídeos criados pelo usuário. De fato, todas essas métricas possuem correlações bem próximas das obtidas com o número de visualizações.

De uma maneira geral, a partir dessa análise dos indicadores de popularidade dos vídeos podemos concluir que usuários com vídeos de qualidade e populares tendem a ter suas duplicatas mais populares, indicando que as características dos vídeos dos usuários tendem a indicar a popularidade de seus novos vídeos. Tal informação é importante não só para a associação de propagandas a vídeos, mas também para a criação de mecanismos de *caching*.

A seguir, analisamos a correlação dos *rankings* considerando-se o número de amigos do dono da duplicata, ilustrada na figura 4 (meio). Podemos notar que usuários com muitos amigos tendem a ter seus vídeos mais populares, indicando que parte dos acessos aos vídeos são oriundos de relações sociais estabelecidas por usuários na rede, semelhante a observações realizadas para o Flickr [19]. Entretanto, podemos observar que tal correlação é menor se comparada aos indicadores de popularidade discutidos anteriormente. Cerca de 73% dos grupos possuem correlação pos-

[3]Para essa análise desconsideramos o vídeo que aparece como resultado da busca e consideramos apenas os vídeos coletados através da lista de duplicados, pois, potencialmente, os vídeo duplicado em evidência no resultado da buscatende a ter mais visualizações. Qualitativamente, os resultados considerando ou não esses vídeos diferem muito pouco.

[4]www.flickr.com

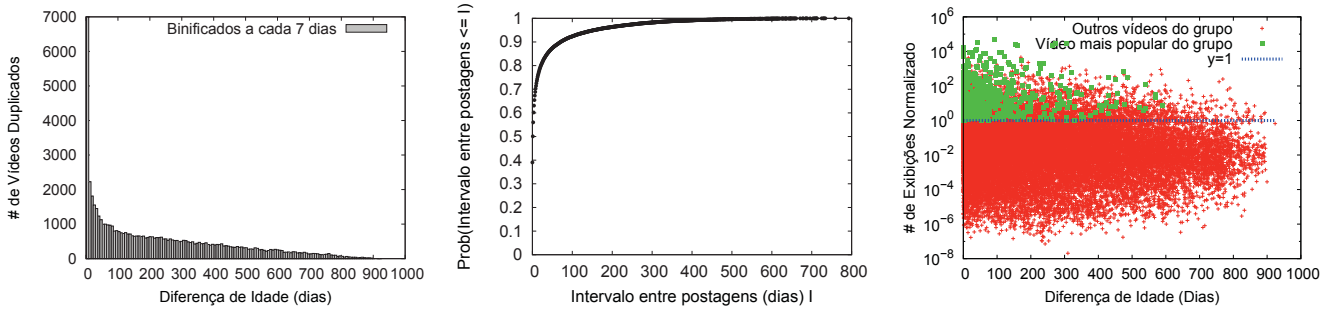

Figura 3: Diferença de idade das cópias em relação a seus vídeos originais (esq.), Distribuição do tempo entre postagem de duplicatas (meio) e Popularidade de cópias em relação a seus vídeos originais (dir.)

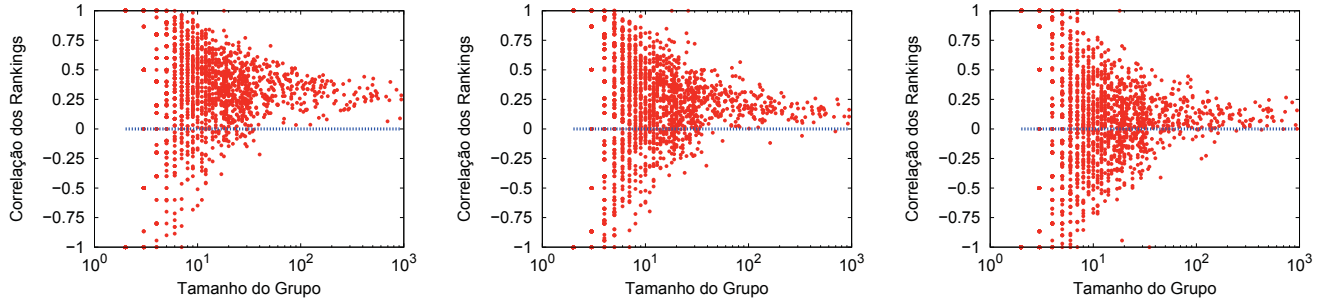

Figura 4: Correlação de Características do Usuário com o Número de Exibições das Duplicatas como Função do Tamanho do Grupo: demais vídeos do usuário (esq.), número de amigos (meio) e número de vídeos (dir.)

itiva e, considerando-se os grupos com mais de 10 duplicatas, 84% possuem correlação positiva, sendo somente 50% maior do que 0,2.

Por último, analisamos o *ranking* dos usuários de acordo com o número de vídeos criados. A correlação desse *ranking* com o das duplicatas está ilustrada na figura 4 (dir.). Novamente, podemos notar uma correlação positiva, especialmente para grupos maiores. Entretanto apenas 58% dos grupos possuem correlação positiva. Dos grupos com mais de 10 duplicatas, 66% possuem correlação positiva, mas apenas 29% maior do que 0,2. Estes números são bastante inferiores a todas as métricas para o *ranking* dos usuários discutidas anteriormente. De fato, intuitivamente, em termos de indicadores da popularidade de um vídeo, a qualidade dos demais vídeos de um usuário é mais importante do que a quantidade.

Em trabalhos futuros pretendemos analisar a influência de outras características dos usuários não coletadas nesse trabalho, como por exemplo, o número de usuários assinantes e métricas básicas de redes sociais como coeficiente de clusterização e reciprocidade [21].

4.4 Uso de Metadados

No YouTube, ao compartilhar um vídeo, o usuário dono do vídeo compartilhado pode associar livremente ao vídeo um título, um texto que descreve o vídeo e *tags* que serão utilizadas para busca e geração de vídeos relacionados. Além disso, o usuário necessariamente precisa associar uma categoria ao vídeo, escolhida dentre um conjunto de categorias pré-definidas pelo sistema. Muitos serviços de informação, tais como busca e recomendação, tem como premissa para uma boa efetividade a associação consistente de metadados a vídeos por parte dos usuários. Nesta seção quantificamos a similaridade existente entre *tags*, descrições e títulos associados a vídeos com conteúdos idêntico ou bastante semelhantes. Em seguida, estudamos também a associação de categorias.

Apesar das buscas terem sido feitas utilizando-se palavras aleatórias de um dicionário da língua Inglesa, podem ter palavras

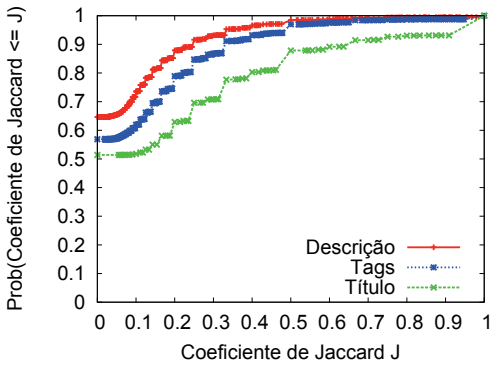

Figura 5: Coeficiente de Jaccard para Tags, Descrição e Título

escritas em outros idiomas. Por esta razão consideramos somente vídeos em que os donos são dos Estados Unidos, Canadá, Reino Unido e Austrália. Esses vídeos correspondem a 33% do total de vídeos na nossa base. Para evitar considerar palavras similares (ex. *house* e *houses*) como diferentes, reduzimos cada palavra ao seu radical. Para fazer isso usamos o método de *stemming* [16][5], um método bastante conhecido para extração de radicais de palavras. Também filtramos palavras sem significado semântico (ex. *the*, *of* e *for*), obtidas de uma lista disponível na referência [22].

Para analisarmos a similaridade entre *tags* de duas duplicatas, utilizamos o coeficiente de Jaccard [3], definido da seguinte forma. Sejam A e B, os conjuntos de *tags* de dois vídeos. O coeficiente de Jaccard, $J(A, B)$, entre A e B é dado pelo número de *tags* em comum entre A e B dividido pelo total de *tags* distintas encontradas na união dos dois conjuntos: $J(A, B) = |A \cap B|/|A \cup B|$. Sendo

[5]Disponível em http://tartarus.org/m̃artin/PorterStemmer

assim, um coeficiente de Jaccard igual a 0 significa que não há *tags* em comum entre dois vídeos enquanto um coeficiente próximo a 1 indica que os vídeos possuem a maior parte das *tags* em comum.

A figura 5 mostra a CDF para o coeficiente de Jaccard para todos os pares de duplicatas dentro do mesmo grupo, para todos os grupos de nossa base de dados. Nós comparamos, separadamente, *tags*, descrição e o título de cada par de duplicatas. Notamos que uma parcela representativa dos pares de duplicatas não possuem *tags* em comum (cerca de 56%), e 87% deles possuem J menor que 0,3. Somente 1,2% dos pares possuem todas as *tags* em comum. Nós também notamos que a similaridade para o título é maior se comparada com a descrição e com as *tags*. Por exemplo, 70% dos pares de duplicatas possuem J para o título menor que 0,3, e 7% possuem J igual a 1. Descrição apresenta a similaridade mais baixa. Por exemplo, 93% dos pares possuem J menor que 0,3, e somente 0,5% possuem toda a descrição em comum.

Como conclusão notamos que, de um modo geral, as duplicatas possuem baixa similaridade em termos de metadados. Título apresenta níveis maiores de similaridade entre duplicatas, indicando que usuários diferentes concordam mais com as palavras escolhidas para representar um mesmo conteúdo. Essa pouca similaridade entre metadados de vídeos com conteúdo iguais reflete que os usuários interpretam um mesmo conteúdo de uma maneira diferente e motiva a busca por alternativas para a recuperação de informação em conteúdo multimídia.

A seguir, vamos analisar como os usuários associam categorias às duplicatas compartilhadas no YouTube. O YouTube permite aos usuários escolher entre 14 categorias pré-definidas: *Música (Mus), Entretenimento (Ent), Pessoas & Blogs (P&B), Humor (Hum), Notícias & Política (N&P), Filmes & Desenhos (F&D), Esportes (Esp), Viagens & Eventos (V&E), Veículos (Vei), Instruções & Estilo (I&E), Educação (Edu), Animais (Ani), Ciência & Tecnologia (C&T), e Sem fins lucrativos & Ativismo (S&A)*.

Com o intuito de entender as diferenças nas categorias das duplicatas, nós comparamos as categorias de cada par de duplicatas dentro do grupo, para todos os grupos, contando as ocorrências para cada par possível de categorias. A tabela 2 mostra essa distribuição. Cada linha se refere à distribuição dos vídeos de uma categoria específica e reporta a fração dos pares comparados que caem nas categorias possíveis (a soma de cada linha representa 100%). Como podemos ver, cerca de 89% dos pares de duplicatas da categoria *Viagens & Eventos* são também associados a *Viagens & Eventos*, indicando que os usuários dessa categoria geralmente concordam com a associação de vídeos a essa categoria. Entretanto, para a maior parte das categorias, uma fração não desprezível dos pares de duplicatas são associadas a diferentes categorias. Particularmente, as categorias *Instruções & Estilo*, *Educação*, e *Sem fins lucrativos & Ativismo* possuem baixas frações de pares de duplicatas na mesma categoria (12,7%, 5,4% e 1,5%, respectivamente), indicando que os usuários não concordam com elas. Em geral, duplicatas dessas categorias são majoritariamente associadas com *Humor* e *Entretenimento*, *Música*, e *Pessoas & Blogs*, que são as categorias mais populares de nossa base, como mostrado na figura **??**. De fato, observando as colunas da tabela 2, podemos notar que essas categorias são populares em termos de ocorrência em pares de duplicatas.

Esta análise reflete como os usuários associam de forma diferente tópicos ao mesmo conteúdo. De fato, a associação de categorias a vídeos é subjetiva e um vídeo pode ser naturalmente associado a diferentes categorias. Por exemplo, um dos vídeos mais vistos de todos os tempos no YouTube, chamado de *Evolution of Dance*, mostra um homem dançando músicas famosas através das décadas. Ele foi associado à categoria *Humor* mas pode naturalmente ser associado a categorias como *Entretenimento* e *Música*.

5. CRIAÇÃO DE DUPLICATAS

Redes sociais têm sido alvo freqüente de comunicação não solicitada e de ações maliciosas [27, 13]. De fato, redes sociais de compartilhamento de vídeos parecem ser um ambiente atrativo para usuários gerar propagandas, disseminar pornografa (geralmente como propaganda) ou simplesmente comprometer a reputação do sistema [7]. De certa forma, parte do conteúdo duplicado existente no YouTube pode ser o resultado de ações maliciosas por parte dos usuários. Esta seção investiga as evidências de possíveis comportamentos maliciosos na criação das duplicatas.

Como os usuários podem criar conteúdo livremente nos sistemas de compartilhamento de vídeos, é esperado a criação acidental de algumas duplicatas. Para verificar isso, analisamos o número de duplicatas criadas por cada usuário do nosso conjunto de dados. A figura 6 (esq.) mostra o *ranking* dos usuários ordenados de acordo com o número de duplicatas criados. Notamos que a maior parte dos usuários criaram poucas duplicatas (98% postaram menos de 3)[6], como esperado quando a criação é acidental. Entretanto, encontramos usuários que criaram um grande número de duplicatas, ou seja, o primeiro e o segundo do *ranking* possuem 714 e 103 duplicatas. Verificamos manualmente esses usuários e discutimos sobre eles em seguida.

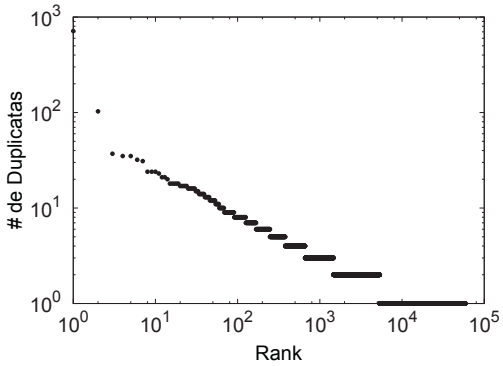

Figura 6: Ranking dos Usuários pelo Número de Duplicatas

O primeiro usuário do *ranking* é um promotor de vídeo resposta. O YouTube possui uma ferramenta que permite aos usuários enviarem um vídeo como resposta a outros vídeos. Um promotor é um usuário que posta um grande número de vídeos respostas para um único alvo, numa tentativa de inserir o vídeo alvo rapidamente nas listas dos vídeos mais respondidos e ganhar visibilidade [4]. Os vídeos postados por esses usuários consistem em vídeos idênticos, geralmente com curta duração (menos que 5 segundos).

Interessantemente, encontramos alguns casos em que as duplicatas de um mesmo usuário possuem diferentes metadados. Em particular, notamos algumas situações em que os usuários criam vídeos iguais mas utilizam *tags* diferentes e populares em cada vídeo. Esse tipo de comportamento oportunista é conhecido na literatura como *tag attack* [13, 18] e é utilizado para enganar mecanismos de buscas para promover algum conteúdo no sistema. O segundo usuário do *ranking* postou vídeos contendo propagandas. Notamos que esse usuário utilizou diversas duplicatas de uma mesma propaganda, mas assinalando diferentes *tags* para cada vídeo, realizando um *tag attack*.

Em comum, notamos que esses 2 usuários criam diversas cópias idênticas de um mesmo conteúdo. Então, investigamos se há outros usuários que criam mais de uma duplicata por grupo. Para isto, calculamos a razão do número de donos pelo número de vídeos em cada grupo de duplicatas da nossa base de dados. Intuitivamente, podemos esperar que a maior parte dos grupos possuam

[6]Entre os outros vídeos do usuário podem existir mais duplicatas além das coletadas, já que coletamos somente aquelas que apareceram como resultado das buscas

	Hum	N&P	I&E	Ent	Edu	C&T	Mus	Vei	Ani	V&E	P&B	F&D	S&A	Esp
Hum	**61,8**	1,4	0,4	14,4	0,1	0,0	3,3	2,1	2,0	0,4	8,3	4,2	0,0	1,6
N&P	35,4	**17,7**	1,9	10,8	1,7	0,5	3,0	6,7	0,7	1,0	14,1	4,1	0,4	2,0
I&E	21,0	4,0	**12,7**	21,5	1,6	2,2	3,2	10,1	1,1	1,2	11,1	5,5	0,1	4,6
Ent	33,7	1,0	0,9	**33,8**	0,3	0,2	4,9	3,5	1,5	0,5	7,9	8,9	0,0	3,0
Edu	14,4	6,3	2,9	14,5	**5,4**	1,5	5,6	13,5	0,6	0,8	24,4	7,4	0,4	2,3
C&T	7,8	3,3	7,2	17,9	2,9	**45,2**	1,2	3,7	0,1	1,0	5,7	3,2	0,1	0,6
Mus	18,3	0,6	0,3	11,4	0,3	0,0	**50,6**	0,6	1,0	0,3	6,5	8,4	0,0	1,5
Vei	21,6	2,7	1,9	15,4	1,4	0,2	1,2	**38,1**	0,3	2,4	9,8	3,6	0,1	1,2
Ani	38,9	0,5	0,4	11,9	0,1	0,0	3,5	0,5	**30,4**	0,3	5,1	7,5	0,0	0,9
V&E	3,2	0,4	0,2	2,0	0,1	0,0	0,5	2,1	0,1	**89,5**	1,0	0,6	0,0	0,5
P&B	40,8	2,7	1,0	16,8	1,2	0,2	5,8	4,7	1,3	0,5	**15,7**	5,1	0,1	4,1
F&D	27,6	1,1	0,6	25,2	0,5	0,1	10,2	2,3	2,6	0,4	6,8	**20,8**	0,0	1,7
S&A	4,0	22,5	3,1	10,2	5,8	0,8	5,6	14,0	0,5	1,6	23,4	6,3	**1,5**	0,7
Esp	7,4	0,4	0,4	6,1	0,1	0,0	1,3	0,6	0,2	0,2	3,8	1,2	0,0	**78,3**

Tabela 2: Distribuição Percentual da Ocorrência de Pares de Categorias

o mesmo número de vídeos e donos de vídeos (razão igual a 1). De fato, nós encontramos razão 1 para 92% dos grupos. Contudo, interessantemente, há uma fração significante de usuários que criaram mais de uma duplicata de um mesmo conteúdo. Chamamos de vídeos suspeitos todas as duplicatas de um grupo criadas por um único usuário. O restante das duplicatas chamamos de vídeos legítimos para as análises seguintes. No total temos 2.668 vídeos suspeitos criados por 608 diferentes usuários.

Já que queremos estudar a associação de metadados a vídeos suspeitos, consideramos somente vídeos cujos donos são dos Estados Unidos, Canadá, Reino Unido e Austrália, ficando um total de 1.086 vídeos suspeitos de 316 usuários. Nós analisamos manualmente vídeos suspeitos de 29 usuários selecionados. Além dos vídeos dos 2 primeiros usuários do *ranking* na figura 6, selecionamos aleatóriamente mais 27 usuários. Como resultado, os vídeos suspeitos analisados foram divididos em três conjuntos básicos: (1) 714 vídeos utilizados para promoção, que são os vídeos criados pelo usuários que postou o maior número de duplicatas em nosso conjunto de dados; (2) 128 vídeos com *tag spam*, criados por 12 usuários diferentes, incluindo o segundo do *ranking*; e (3) duplicatas-próximas. Por duplicatas-próximas nos referimos a vídeos idênticos mas com diferenças na qualidade ou nas legendas em línguas diferentes, vídeos idênticos mas com comentários embutidos, e mesmo vídeos idênticos mas com áudio diferente. Temos 58 vídeos de 16 usuários nesse grupo.

Em seguida, nós comparamos a similaridade das *tags*, descrição e título para os vídeos suspeitos selecionados e para os vídeos legítimos. Cada gráfico na figura 7 mostra a CDF do coeficiente de Jaccard para cada um desses metadados. Foram comparados, para cada metadado, quatro conjuntos de duplicatas: legítimos, vídeos utilizados para promoção, vídeos com *tag spam*, e duplicatas-próximas. Para cada conjunto de vídeos comparamos todos os possíveis pares de vídeos de um mesmo grupo de duplicatas.

Observando a figura 7 (esq.), podemos notar que o conjunto de vídeos utilizado para promoção apresenta o nível mais baixo de similaridade de *tags*. Por exemplo, menos de 8% dos vídeos possuem um coeficiente de Jaccard maior que 0,1. Analisando esse conjunto de vídeos suspeitos, notamos que cada vídeo possue somente algumas *tags* que parecem ter sido geradas automaticamente, levando a um baixo nível de similaridade.

Por outro lado, o conjunto de duplicatas-próximas apresenta um nível maior de coefficiente de Jaccard, se comparado com os outros conjuntos de vídeos. Por exemplo, 55% dos pares de vídeos possuem um coeficiente de Jaccard maior que 0,3 e 80% maior que 0,1. Notamos que esses usuários que criam vídeos com somente diferenças na qualidade, ou nos comentários e legendas embutidos, tendem a utilizar basicamente o mesmo conjunto de tags.

Interessantemente, a maior parte dos vídeos com *tag spam* apresentam uma alta concentração do coeficiente de Jaccard entre 0,3 e 0,6 (90% deles). Analisando esse conjunto de vídeos suspeitos notamos que a maior parte deles possui um conjunto principal de *tags* que são relacionadas ao conteúdo do vídeo e utilizadas em todas as duplicatas, aumentando o coeficiente de Jaccard, e cada vídeo possui um outro conjunto de *tags* variadas e não relacionadas ao vídeo, numa tentativa de enganar mecanismos de buscas e ganhar visibilidade com o conteúdo.

Nós também realizamos a mesma comparação para as palavras utilizadas na descrição e no título, nas figuras 7 (meio) e 7 (dir.), respectivamente. De uma maneira geral, observações similares a *tags* se mantém para descrição e título. Em resumo, podemos concluir com essas análises que, dependo do objetivo dos usuários que criam duplicatas propositalmente, a forma com que eles associam os metadados a esses vídeos variam bastante.

6. PRINCIPAIS DESCOBERTAS

A seguir, sumarizamos as principais descobertas desse trabalho.

- A maior parte das duplicatas são criadas em um curto período de tempo, quando o conteúdo do vídeo está em alta na rede. No entanto, ainda é possível encontrar duplicatas criadas mais de 2 anos depois que o conteúdo surgiu no sistema.

- Grande parte dos vídeos com conteúdos iguais são avaliados de forma semelhante pelos usuários. Apenas uma pequena fração dos vídeos são avaliados de forma muito discrepante. Além disso, apesar de iguais, alguns desses vídeos se tornam muito mais populares do que outros.

- Duplicatas possuem poucos metadados (*tags*, descrição, título e categoria) em comum. Isso pode ser um reflexo da diferente percepção dos usuários com relação ao conteúdo do vídeo, uma mídia rica de informações. Essa baixa similaridade motiva a busca por alternativas ao uso de metadados na recuperação de conteúdo multimídia.

- Usuários com vídeos de qualidade e populares tendem a ter suas duplicatas mais populares, indicando que as características dos vídeos dos usuários tendem a indicar a popularidade de seus novos vídeos. Além disso, usuários com muitos amigos tendem a ter suas duplicatas mais populares, indicando que parte dos acessos aos vídeos é vinda de relações sociais estabelecidas na rede. Tais informações são importantes não só para a associação de propagandas a vídeos, mas também para a criação de mecanismos de *caching*.

- Em cerca de 8% dos grupos de duplicatas existem vídeos cujos donos criam mais de um vídeo com conteúdo igual. Dentre as várias razões para isso, notamos usuários que criam vídeos iguais, associando *tags* populares e diferentes a cada um deles,

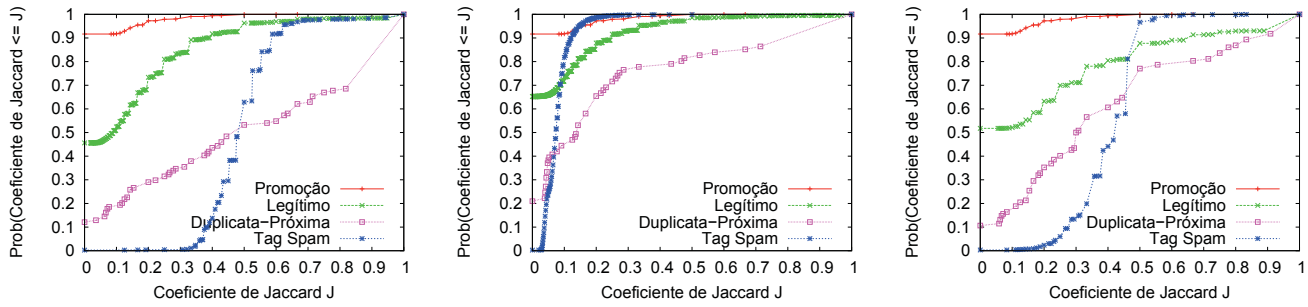

Figura 7: *Tags* (esq.), Descrição (meio) e Título (dir.). Similaridade para os Vídeos Suspeitos

na tentativa de enganar máquinas de busca para divulgar algum conteúdo.

7. CONCLUSÕES

Em sistemas como o YouTube, usuários podem compartilhar vídeos de forma independente, fazendo com que vários vídeos sejam idênticos ou muito similares. A existência de duplicatas abre espaço para o entendimento de questões importantes de sistemas de compartilhamento de vídeos. Esse trabalho apresenta uma ampla caracterização das diferenças contextuais existentes entre duplicatas no YouTube. Inicialmente, coletamos e criamos uma grande e representativa base de duplicatas do YouTube. Nossos resultados analisam e discutem diversas diferenças em relação aos metadados e aos indicadores de popularidade de duplicatas. Além disso, mostramos que popularidade dos donos dos vídeos está correlacionada com a popularidade dos vídeos, indicando que o perfil do dono pode influenciar na popularidade de seus vídeos. Por último, nossos resultados revelam a existência de comportamento malicioso por parte dos usuários na criação intencional de duplicatas.

Como trabalhos futuros, existem três direções para a qual este trabalho pode evoluir. A primeira consiste em aprofundar nossas análises sobre comportamento malicioso por parte dos usuários na criação de duplicatas. A segunda consiste em caracterizar e entender outros fatores que possam influenciar na popularidade de vídeos, essencial para um emergente nicho de mercado que é a associação de propagandas a vídeos. Por último, pretendemos aprofundar os estudos no entendimento da associação de metadados a vídeos pelos usuários.

8. REFERÊNCIAS

[1] E. Adar, L. Zhang, L. Adamic, and R. Lukose. Implicit structure and the dynamics of blogspace. *Workshop on the Weblogging Ecosystem*, 2004.

[2] Alexa. Disponível em: <http://www.alexa.com>. Acesso em: julho de 2009.

[3] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[4] F. Benevenuto, F. Duarte, T. Rodrigues, V. Almeida, J. Almeida, and K. Ross. Understanding video interactions in youtube. In *Proc. ACM Multimedia (MM)*, 2008.

[5] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonçalves. Detectando usuários maliciosos em interações via vídeos no youtube. In *Simpósio Brasileiro de Sistemas Multimída e Web (WebMedia)*, 2008.

[6] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonçalves. Detecting spammers and content promoters in online video social networks. In *Int'l ACM SIGIR*, Boston, USA, July 2009.

[7] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, C. Zhang, and K. Ross. Identifying video spammers in online social networks. In *Proc. Int'l Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2008.

[8] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. In *Proc. Internet Measurement Conference (IMC)*, 2007.

[9] comScore. Americans Viewed 12 Billion Videos Online in May 2008. Disponível em: <http://www.comscore.com/press/release.asp?press=2324>. Acesso em: julho de 2009.

[10] copyright. Youtube copyright policy: Video identification tool. Disponível em: <http://help.youtube.com/support/youtube/bin/answer.py?hlrm=en&answer=83766&hlrm=en_US>. Acesso em: julho de 2009.

[11] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube traffic characterization: A view from the edge. In *Proc. Internet Measurement Conference (IMC)*, 2007.

[12] J. Golbeck. Trust and nuanced profile similarity in online social networks. Technical report, 2008.

[13] P. Heymann, G. Koutrika, and H. Garcia-Molina. Fighting spam on social web sites: A survey of approaches and future challenges. *IEEE Internet Computing*, 11, 2007.

[14] Ispell. Disponível em: <http://www.gnu.org/software/ispell/ispell.html>. Acesso em: julho de 2009.

[15] R. Jain. *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. John Wiley and Sons, INC, 1991.

[16] K. S. Jones and P. Willett, editors. *Readings in Information Retrieval*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[17] T. Kelly and J. Mogul. Aliasing on the world wide web: prevalence and performance implications. In *Int'l conference on World Wide Web (WWW)*, 2002.

[18] G. Koutrika, F. Effendi, Z. Gyöngyi, P. Heymann, and H. Garcia-Molina. Combating spam in tagging systems. In *Proc. Int'l Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2007.

[19] K. Lerman and L. Jones. Social browsing on flickr. In *Proc. Int'l Conference on Weblogs and Social Media (ICWSM)*, 2007.

[20] X. Li, L. Guo, and Y. E. Zhao. Tag-based social interest discovery. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 675–684, New York, NY, USA, 2008. ACM.

[21] M. Newman and J. Park. Why social networks are different from other types of networks. *Phys. Rev. E*, 68, 2003.

[22] C. Rijsbergen. *Information Retrieval*. Butterworth, 1979.

[23] F. M. Suchanek, M. Vojnovic, and D. Gunawardena. Social tags: meaning and suggestions. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 223–232, New York, NY, USA, 2008. ACM.

[24] X. Wu, A. Hauptmann, and C. Ngo. Practical elimination of near-duplicates from web video search. In *Int'l Conference on Multimedia*, 2007.

[25] J. Zhu, S. Hoi, M. Lyu, and S. Yan. Near-duplicate keyframe retrieval by nonrigid image matching. In *ACM Int'l Conference on Multimedia (MM)*, 2008.

[26] M. Zink, K. Suh, Y. Gu, and J. Kurose. Watch global, cache local: Youtube network traces at a campus network - measurements and implications. In *IEEE Multimedia Computing and Networking (MMCN)*, January 2008.

[27] A. Zinman and J. Donath. Is britney spears spam? In *Proc. Conference on Email and Anti-Spam (CEAS)*, 2007.