Caracterização do Comportamento dos Espectadores em Transmissões de Vídeo ao Vivo Geradas por Usuários*

Thiago Silva[1], Vinícius Mota[1], Everthon Valadão[1], Jussara Almeida[1], Dorgival Guedes[1]

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

{thiagohs,vfsmota,evaladao,jussara,dorgival}@dcc.ufmg.br

Abstract. *We have presented what we believe to be the first work on characterization of users behaviors in a live video streaming system based on the Web 2.0. We have analyzed over 1 million of users sessions distributed in more than 7 thousand channels. Our contributions are two-fold. First, we have presented an analysis of the user's interaction model characterizing, among others aspects, the transmissions, and session durations (both follow a Lognormal distribution). Second, we have analyzed the channels popularity distributions and its implications on the system. We found that users are likely to stay longer in popular channels, and the number of simultaneous users in the channel is a factor which attracts new users.*

Resumo. *Apresentamos o que consideramos ser o primeiro trabalho de caracterização do comportamento de usuários de um sistema de transmissão de vídeo ao vivo baseado na Web 2.0. Analisamos mais de 1 milhão de sessões de usuários distribuídas em mais de 7 mil canais. Nosso trabalho pode ser dividido em duas partes. Primeiramente, analisamos a interação dos usuários no sistema caracterizando, entre outros aspectos, a duração das transmissões e sessões (ambas seguem uma distribuição Lognormal). Em seguida, analisamos a distribuição de popularidade dos canais e suas implicações no sistema. Verificamos que usuários tendem a permanecer mais tempo em canais populares e que o número de usuários simultâneos influencia na atração de novos usuários.*

1. Introdução

Nos últimos anos assistimos a uma grande revolução com relação às formas de criação e distribuição de vídeo na Internet. Com a popularização das câmeras digitais e o aumento médio das velocidades de acesso à Internet, a geração e distribuição de vídeos na rede mundial de computadores aumentou consideravelmente. Não somente a oferta de conteúdo aumentou como também a procura pelos mesmos. Foi mostrado que a visualização de conteúdo como seriados de TV, pequenos vídeos (como os disponíveis no YouTube[1]), animações e outras aplicações que geram tráfego através do consumo de vídeo já representam mais de 60% de todo tráfego da Internet [2].

Com o surgimento da Web 2.0 [Oreilly 2007], serviços Web agora contam com a colaboração de usuários para aumentar a quantidade de conteúdo oferecida a seus clientes.

*Esta pesquisa é parcialmente financiada pelo Instituto Nacional de Ciência e Tecnologia para a Web - INCTWeb (MCT/CNPq 573871/2008-6) e pelo Projeto REBU (CTInfo/CNPq 55.0995/2007-2).

[1]http://www.youtube.com - Todas URLs foram acessadas pela última vez em Dezembro de 2008

[2]http://www.cachelogic.com

Essa criação de conteúdo de forma colaborativa se mostra um importante fator na popularidade de diversos serviços como YouTube, Flickr[3] e Delicious[4].

No passado, a realização de transmissão de vídeo ao vivo era acessível somente a grandes empresas de distribuição de mídia, e para isso era necessário um alto investimento financeiro. Atualmente, existem vários serviços gratuitos para distribuição de fluxos de vídeo ao vivo na Internet baseados na Web 2.0. Exemplos desses serviços são Justin.tv[5], Ustream.tv[6], Mogulus.com[7], Stickam.com[8] e YahooLive[9]. Nota-se que o interesse por esse tipo de serviço vem crescendo. Recentemente o sistema Justin.tv anunciou em seu *website* o registro de mais de 101.000 usuários simultâneos.

Para a distribuição de fluxos de vídeos ao vivo na Internet utiliza-se, basicamente, duas arquiteturas de rede: as tradicionais cliente-servidor e as redes Par-a-Par (P2P). No entanto, é possível realizar uma categorização das transmissões de fluxo de vídeo ao vivo considerando se um vídeo é gerado pelo usuário ou pelo provedor de conteúdo. Nesse artigo tratamos de vídeos ao vivo produzidos pelos usuários.

Com isso, esse trabalho analisa um sistema para transmissão de fluxo de vídeo ao vivo gerado e transmitido por usuários oferecido pela Yahoo. Esse sistema, YahooLive, foi escolhido por disponibilizar informações necessárias para realização do nosso trabalho e também devido ao crescimento de sua popularidade, como será abordado na seção 4. Através de uma caracterização realizada neste sistema, foi possível entender melhor as questões referentes ao comportamento e à utilização dos recursos pelos usuários nessa aplicação. Essas informações são úteis no planejamento de carga de provedores de conteúdo, bem como para a geração de cargas sintéticas mais próximas da realidade para esse tipo de serviço, que poderiam ser utilizadas, por exemplo, na verificação da aplicabilidade de arquitetura de rede P2P para atender um serviço semelhante.

O restante do trabalho está organizado da seguinte forma: A seção 2 descreve os principais trabalhos relacionados. Na seção 3 apresentamos os procedimentos realizados na aquisição e tratamento dos dados. Na seção 4 analisamos a participação dos usuários no sistema. A seção 5 analisa o impacto referente à popularidade dos canais no sistema. Finalmente, na seção 6 finalizamos com algumas considerações finais.

2. Trabalhos relacionados

A caracterização de como os usuários utilizam um sistema é importante tanto para obter um melhor entendimento sobre o mesmo, quanto para a produção de cargas de trabalho sintéticas mais realistas e para projeção de mecanismos e soluções mais robustas.

Vários estudos analisaram cargas de trabalho de sistemas que oferecem fluxo de vídeo. É possível classificá-los em três classes de trabalhos. A primeira trata de análise de sistemas em que os fluxos de vídeos são ao vivo, bem como gerados e transmitidos, na Web, apenas pelo provedor de conteúdo. Em [Veloso et al. 2006] os autores estudaram a carga de trabalho de um servidor comercial de fluxo de vídeos ao vivo localizado no Brasil. O foco do estudo foi, basicamente, a caracterização dos processos de chegadas e durações de sessões, com o intuito de utilizar os resultados em um gerador de carga sintética.

[3]http://www.flickr.com
[4]http://www.delicious.com
[5]http://www.justin.tv
[6]http://www.ustream.tv
[7]http://www.mogulus.com
[8]http://www.stickam.com
[9]http://www.live.yahoo.com

Em [Sripanidkulchai et al. 2004b] foi analisado uma carga de trabalho de fluxos de vídeo e áudio ao vivo de uma grande CDN, o que possibilitou a análise de uma ampla diversidade de conteúdos. Entretanto, mais de 90% do conteúdo analisado foi apenas áudio. Observou-se que a popularidade dos conteúdos segue uma distribuição Zipf de dois modos. Foi também observado que os clientes entram no sistema de acordo com uma distribuição exponencial e que a duração das sessões apresentou cauda pesada.

A segunda classe de trabalhos destinou-se a caracterizar sistemas onde os fluxos de vídeos, também ao vivo, podem ser gerados e/ou transmitidos pelos usuários. Todos esses consideraram sistemas baseados em arquiteturas P2P. Enfatizou-se análises das redes P2P empregadas nesses sistemas, porém foi estudado também o comportamento dos usuários nessas aplicações [Hei et al. 2007], [Li et al. 2007] e [Silerston and Fourmaux 2007].

E por fim, a última classe de trabalhos estudou sistemas para transmissão de fluxos de vídeos armazenados[10] baseado na Web 2.0. Esses estudos objetivam melhor entender e identificar melhorias para o sistema estudado. Em [Duarte et al. 2007] foi analisado características dos vídeos e usuários do YouTube de diferentes regiões geográficas. Foi mostrado evidências de que a localidade dos usuários podem ser exploradas para melhorar a infraestrutura utilizada pelo provedor do serviço.

Os autores em [Gill et al. 2007] analisaram o tráfego do YouTube sob duas perspectivas: local, de uma universidade no Canadá, e global, a partir da lista dos 100 vídeos mais vistos, que é disponibilizada no *website* do YouTube. Os autores mostram que a utilização de cache trazem benefícios para os usuários e provedores de conteúdo. Em [Maia et al. 2008], foram avaliados quatro populares sistemas para transmissão de fluxo de vídeo armazenado, dentre eles o YouTube, DailyMotion[11]. Os autores propuseram uma estratégia de *caching* que explora tanto a popularidade quanto a recência dos vídeos.

Nosso trabalho se diferencia dos demais pois analisamos um sistema que oferece fluxo de vídeo ao vivo, gerado e transmitido apenas pelos usuários, sendo ainda um sistema baseado na Web 2.0. Além disso, focamos principalmente em melhor entender o comportamento dos usuários na aplicação. Até onde podemos verificar, nosso trabalho é um estudo pioneiro com tais características.

3. Metodologia

Nessa seção descrevemos o sistema estudado, bem como a metodologia usada para coletar e processar os dados utilizados neste trabalho. Explicaremos também os procedimentos usados para limpeza dos dados.

3.1. O sistema analisado

Como mencionado anteriormente, o sistema analisado foi o YahooLive. Esse sistema, criado em fevereiro de 2008, é um serviço que permite, de forma bastante simplificada, transmissões de fluxo de vídeo ao vivo na Web, sem nenhum custo financeiro para o usuário. O sistema ainda oferece a possibilidade de interação com os espectadores através de um sistema de bate-papo em modo texto e/ou através de vídeo, caso o espectador possua uma câmera e um microfone. Os usuários não podem pausar, voltar ou avançar o conteúdo de uma transmissão.

[10]Uma mídia (áudio/vídeo) denominada armazenada é uma mídia que durante seu processo de produção/filmagem não é transmitida ao espectador

[11]http://www.dailymotion.com

3.2. Modelo de interação do usuário com o sistema

No YahooLive cada usuário cadastrado é associado a um canal. Esse canal é único e possui o mesmo nome do usuário, dono do canal. Um participante pode disponibilizar conteúdo em seu canal ao vivo a qualquer momento, por quanto tempo quiser, podendo realizar esse processo quantas vezes desejar. Cada intervalo durante o qual um canal fica disponível ao vivo caracteriza uma **transmissão**.

Para a realização de uma transmissão no sistema o participante necessita ser previamente cadastrado. Por outro lado, para participar apenas como espectador isso não é necessário. O intervalo de tempo (entrada e saída do canal), que um usuário assiste uma determinada transmissão é denominado **tempo de sessão** ou **tempo de permanência**. Ao longo de uma mesma transmissão os usuários podem ter inúmeras sessões.

Durante uma transmissão, cada participante possui um identificador (ID) único. No entanto, esse identificador é alocado dinamicamente. Isso significa que, caso um usuário saia de alguma transmissão e em seguida retorne para a mesma transmissão, ou entre em outra, seu ID poderá não ser mais o mesmo, mas será único.

Um usuário pode assistir/se conectar simultaneamente a várias transmissões, porém em cada uma o seu ID será diferente. Sendo assim, nesse trabalho assumimos que, cada ID identificado no sistema representa um **usuário**.

A Figura 1 mostra uma situação hipotética de dois canais C_1 e C_2. Cada canal realizou duas transmissões. Em cada transmissão apresentada nessa figura, podemos observar as sessões dos usuários participantes. Por exemplo, a transmissão Tr_1 do canal C_1 apresenta três sessões, referentes aos usuários A,B,C.

3.3. Coleta e tratamento dos dados

O YahooLive disponibilizou em seu *website* uma API que permitiu o acesso às informações referentes ao número de canais ao vivo em determinado instante, como também os usuários que estão assistindo esses canais. Esse foi um fator decisivo na escolha do sistema a ser analisado, pois essas informações não são fornecidas por outros sistemas similares.

Foi implementada uma ferramenta, que utilizava a API da Yahoo, para a realização da coleta dos dados. A cada 20 segundos o coletor amostrava todos os canais ao vivo no sistema e capturava o ID de todos os usuários participantes, em todos os canais ao vivo, naquele instante. Conseguimos obter uma visão completa de todos os canais e IDs dos usuários, que visualizaram ou transmitiram algum conteúdo, no período analisado. Foi capturado mais de 1 milhão de sessões de usuários em 48.338 transmissões realizadas.

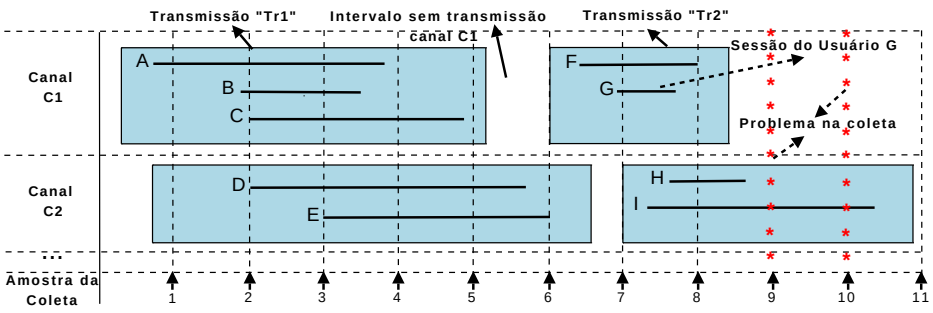


Figura 1. Modelo de comportamento dos usuários e transmissões de canais

A Figura 1 exibe 11 amostras realizadas hipoteticamente pelo coletor. Considerando

Tabela 1. Resumo dos dados coletados

Descrição	Valor
Intervalo da coleta	30-05-08 a 22-06-08 (23 dias)
Núm. de transmissões consideradas	48.338
Total de sessões (ou IDs) consideradas	1.040.688
# de canais únicos ao vivo	7.432
# médio de IDs simultâneos *(C_v) ($max.$)*	719,6 *(0,23) (2.127)*
# médio de canais simultâneos *(C_v) ($max.$)*	56,4 *(0,16) (82)*
# médio de transmissões por canal *(C_v) ($max.$)*	6,6 *(1,72) (248)*
# médio de sessões por transmissão *(C_v) ($max.$)*	30,4 *(3,75) (6.777)*
Duração média das transmissões *(C_v) ($max.$)*	26,3 min *(4,07) (6.055,33 min)*

a situação apresentada nessa figura, serão apresentados a seguir os procedimentos de identificação de duração das transmissões, como também das sessões.

Identificação de duração das transmissões: Identificamos que o canal C_1 iniciou sua transmissão, Tr_1, na amostra da coleta 1 e a finalizou na amostra 5, sendo identificado em 5 amostras da coleta. Assim, consideramos que o tempo total dessa transmissão foi de 100 segundos, pois assumimos que cada evento observado na amostra x dura até imediatamente antes da amostra $x + 1$. Seguindo o mesmo princípio, o tempo da transmissão $Tr2$ foi de 60 segundos.

Identificação de duração das sessões: O processo de identificação de duração das sessões dos usuários, é semelhante ao processo de identificação de duração de uma transmissão. Na transmissão Tr_1 o usuário A foi identificado em 3 amostras da coleta, assim o seu tempo de permanência seria de 60 segundos. Nem todas as sessões menores que 20 segundos são perdidas, por exemplo, a sessão G seria identificada em nossa coleta.

3.4. Limpeza dos dados coletados

Realizamos dois procedimentos de remoção de registros para manter a integridade de nossas análises:

1- Em algumas transmissões não foi possível identificar seu começo ou seu fim. No início e no final da coleta dos dados havia transmissões em execução. Optamos por desconsiderar essas transmissões de nossas análises.

2- Em alguns intervalos durante a coleta enfrentamos algumas falhas na obtenção das informações. Isso pode ter acontecido por problemas na rede ou devido a momentos de instabilidade do serviço oferecido pela Yahoo. Nas amostras da coleta 9 e 10, representadas na Figura 1, é mostrado a ocorrência dessa falha. Como efeito desse problema, não podemos identificar com uma precisão aceitável, por exemplo, em que instante da coleta o usuário H deixou a transmissão. Por esse motivo, optamos por também remover todas as transmissões que estavam ao vivo nos momentos de ocorrência dessas falhas.

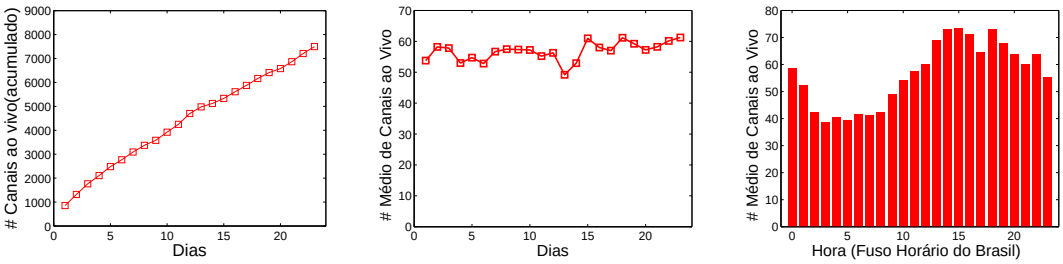
O total de remoções representou uma perda de 1% do número de transmissões observadas e 11% do número total de sessões. A tabela 1 apresenta o resultado após a limpeza dos logs. Quando os valores são apresentados como médias aritméticas, são mostrados também o coeficiente de variação (C_v) e o valor máximo observado ($max.$).

4. Participação dos usuários

Nessa seção identificamos as principais características dos usuários e seus comportamentos no sistema.

4.1. Visão geral

Através de nossa coleta foi possível obter uma visão completa da utilização do sistema por todos os usuários que usufruíram do serviço, durante o período analisado (30/05/08 a 22/06/08) . A Figura 2a mostra o número de canais ao vivo distintos ao longo da coleta. Pode-se observar um crescente aumento na popularidade do sistema, apresentando um crescimento de aproximadamente 8 vezes no número de canais que realizaram alguma transmissão, durante os 23 dias de coleta.



(a) # de Canais ao Vivo Distintos Observados

(b) Canais ao vivo - Dias

(c) Canais ao vivo - Horas

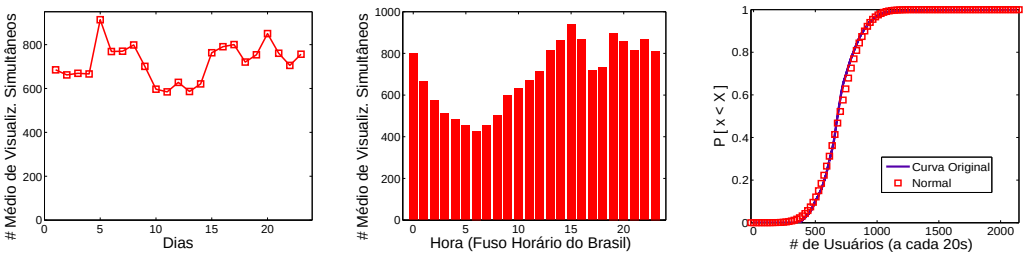
Figura 2. Visão geral - Volume de canais

Canais simultâneos: A Figura 2b mostra o número médio de canais ao vivo identificados no decorrer dos dias de realização da coleta. Podemos observar que o sistema apresentou uma média diária de 56 canais simultâneos realizando pelo menos uma transmissão ao vivo. A Figura 2c mostra o número médio de canais ao vivo ativos por hora do dia ao longo de todo o período, mostrando um comportamento circadiano que se adapta bem ao horário dos EUA. Vale ressaltar que, não necessariamente com o aumento de canais novos o número de canais simultâneos aumentaria, pois um canal poderia entrar no ar em algum dia X e só realizar alguma transmissão novamente no dia $X + 20$.

Usuários simultâneos: A Figura 3a informa o número médio de usuários simultâneos para cada dia da coleta. Podemos observar que o sistema apresentou um número médio de 720 usuários simultâneos por dia. A Figura 3b representa o número médio de usuários simultâneos, por hora do dia, considerando todos os dias analisados. Observa-se também uma tendência de maior utilização entre 13 e 22 horas. Podemos ainda observar uma considerável queda na utilização entre 3 e 8 horas. Ainda com relação ao número de usuários simultâneos, a Figura 3c apresenta uma função de densidade acumulada (CDF) da visão instantânea do número de participantes simultâneos, em todos canais, de todos intervalos coletados. Através de um ajuste de curvas, foi identificado que uma distribuição Normal[12] ($\mu = 693,05$ e $\sigma = 160,9$) melhor descreve esse processo.

Nesse trabalho, utilizamos o *Maximum Likelihood Estimation* (MLE) [Myung 2003] para estimativa dos parâmetros das distribuições. A escolha da distribuição que melhor se adequava aos dados foi feita através de uma inspeção visual nas diversas distribuições consideradas, bem como levamos em consideração também o critério AIC [Akaike 1974].

[12] $P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$



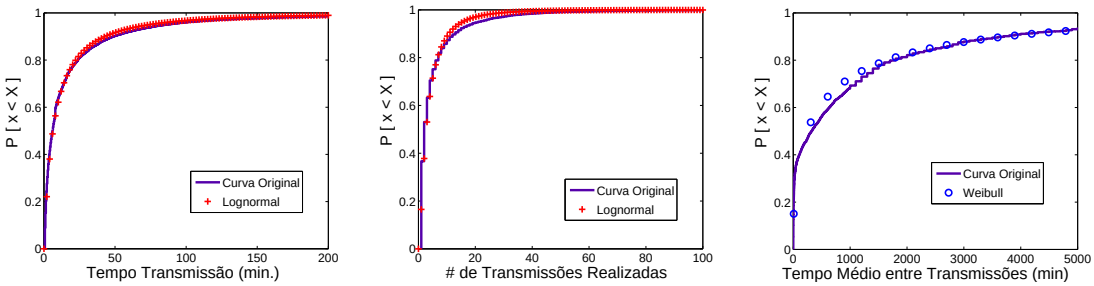
(a) Visualizadores simultâneos - Dias (b) Visualizadores simultâneos - Horas (c) Visão instantânea - 20s

Figura 3. Visão geral - Usuários simultâneos

4.2. Durações e intervalos entre transmissões

A figura 4a mostra a CDF do tempo de duração das transmissões de todos canais observados durante a análise. Nessa figura são mostradas apenas transmissões que duraram até 200 minutos (97,7% de todas as transmissões), a fim de se obter uma melhor visão da parte mais relevante da curva apresentada. Podemos perceber que as transmissões com duração de até 20 minutos representam mais de 70% de todas as transmissões identificadas. Pode-se ainda notar que existe um número considerável de transmissões bastante curtas, pois 15% das transmissões duraram no máximo 1 minuto. Conjecturamos que esse último resultado seja proveniente do período de experiência[13] do usuário.

Identificamos que uma distribuição Lognormal[14] ($\mu = 1.851$ e $\sigma = 1.492$) melhor representa os dados reais encontrados.



(a) Duração das transmissões (b) Núm. de transmissões (c) Intervalo entre transmissões

Figura 4. Análises da duração das transmissões

A Figura 4b mostra a CDF do número de transmissões realizadas pelos canais. É possível notar que pouco mais de 90% dos canais realizaram até 20 transmissões nos 23 dias analisados. Podemos observar também que uma parte considerável (36,6%) dos canais realizou apenas 1 transmissão. Acreditamos que esse resultado tenha duas causas: usuários que testaram o sistema e não voltaram no período analisado e canais novos que entraram no sis-

[13] Período inicial do usuário no sistema, utilizado para testes e aprendizado do serviço

[14] $p(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(\frac{-(\ln x - \mu)^2}{2\sigma^2}\right)$

tema nos últimos dias das coletas. Verificamos que uma distribuição Lognormal ($\mu = 1.03$ e $\sigma = 1.045$) foi a que melhor se adequou aos dados reais encontrados.

A Figura 4c mostra a CDF do intervalo médio entre transmissões sucessivas realizadas por um mesmo canal. Para essa análise, levamos em consideração apenas canais que realizaram mais que 1 transmissão (63,4% dos canais identificados). Podemos observar que boa parte dessas transmissões (75%) aconteceu com um intervalo menor que 1 dia. Acreditamos também que, parte significativa desse resultado é referente ao período de experiência do usuário, onde ainda não se está preocupado com a qualidade do conteúdo e sim com o entendimento do sistema, já que cerca de 21% dos intervalos médios entre transmissões foram de até 10 minutos. Apesar disso, os usuários, além de sua fase de experiência, tendem a transmitir algum conteúdo mais de uma vez ao dia. Assim, é recomendado que os provedores de conteúdo desse tipo de serviço considerem essa hipótese no seu planejamento de carga.

O intervalo médio entre as transmissões de um mesmo canal é melhor representado por uma distribuição Weibull[15] ($a = 556,802$ e $b = 0,438$).

4.3. Processo de chegada de sessões

Para analisarmos o momento em que os usuários entraram em um canal, calculamos o momento de entrada dos usuários relativo ao tempo de duração das transmissões. Se um usuário entrou em uma transmissão já ativa por 40 segundos e essa transmissão durou 100 segundos, então o usuário entrou no intervalo de 40 % da transmissão.

A Figura 5 mostra o resultado dessas análises através de uma CDF. Identificamos que uma distribuição Uniforme melhor descreve os intervalos relativos às entradas dos usuários em determinado canal. Conjecturamos que esse resultado pode ter duas explicações:

1- Usuários podem colocar seu canal no estado ao vivo quando quiserem, e por quanto tempo desejarem. No entanto, o YahooLive não disponibiliza bons mecanismos para divulgação de canais que estão no ar, nem os que planejam iniciar suas transmissões em algum determinado momento. Logo acreditamos que, com o auxílio de tais mecanismos, a tendência seria uma maior aderência dos usuários a alguma transmissão no início da mesma.

2- A Figura 6 mostra o número de canais únicos identificados por dia (d), bem como o número de canais que realizaram alguma transmissão no dia $d-1$ e retornaram no dia d e o número de canais que realizaram alguma transmissão no período anterior (desde o início da coleta) e retornaram no dia d. Podemos perceber uma considerável variabilidade de canais novos por dia. Acreditamos que isso dificulta o processo de fidelização dos usuários em determinados canais, e também aumenta número de vezes que o usuário verifica o conteúdo desses canais, o que pode acontecer a qualquer momento.

4.4. Análises das sessões

Nas análises de sessões dos usuários consideramos duas categorias: transmissões que duraram até 20 minutos (curta) e transmissões que duraram acima que 20 minutos (longa). O limiar de 20 minutos foi definido através de comparação com outros limiares. Especificamente com esse limiar observamos uma distinção maior nas distribuições de tempo de sessão dos usuários. Esta categorização se fez necessária para evitar introduzir ruídos devido à agregação de comportamentos muito distintos, observados em nossos dados.

[15] $P(x) = abx^{b-1}\exp\left(-ax^b\right)$

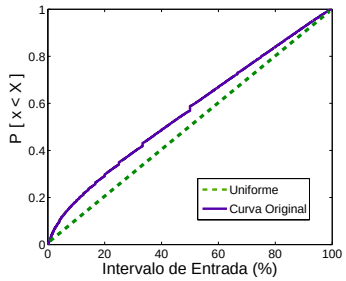


Figura 5. Intervalo de entrada na transmissão

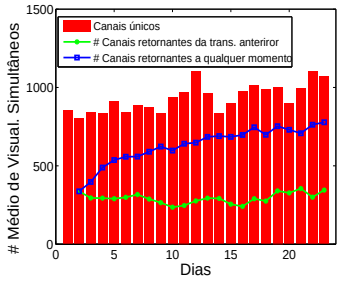


Figura 6. Canais únicos por dia

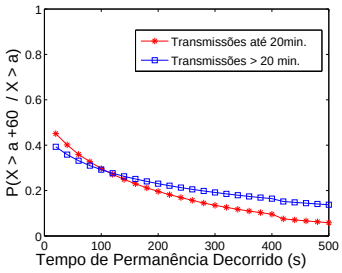


Figura 7. Probab. perman. por mais 60s, em função do tempo já decorrido.

4.4.1. Características das sessões

O tempo médio das sessões nas transmissões da categoria curta foi de 163,7 segundos (com desvio padrão $\sigma = 207,8$) e para as transmissões da categoria longa foi de 253,09 segundos ($\sigma = 501,3$). Nesse último caso consideramos apenas sessões com duração máxima de 1 hora, representando 98,4% de todas as sessões. Os altos valores dos desvios padrões são explicados pela grande variabilidade dos tempos de duração das sessões.

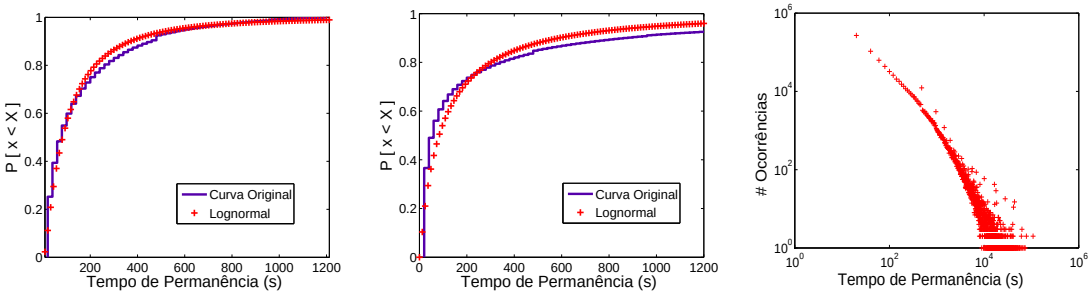
As Figuras 8a e 8b apresentam as distribuições de probabilidade acumuladas (CDF) das durações de sessões identificadas nas transmissões menores que 20 minutos e maiores que 20 minutos respectivamente. Nesse último caso estamos mostrando apenas os tempos de permanência até 1200 segundos, com o objetivo de ressaltar a parte mais relevante da curva. Optamos por representar o tempo de permanência de forma absoluta por melhor representar o comportamento do usuário, que permanece na sessão durante certo tempo, à revelia e sem conhecimento da duração total da transmissão.

Para as duas categorias identificamos, através de um ajuste de curvas, a distribuição que melhor representa os dados observados. Para transmissões com duração de até 20 minutos, constatamos que as distribuições Lognormal e Weibull foram as que mais se aproximaram da curva identificada para os dados reais, porém a distribuição Lognormal ($\mu = 4,424$, $\sigma = 1,161$) representou melhor os dados originais.

Com relação às transmissões com duração superior a 20 minutos foi constatado também que uma distribuição Lognormal ($\mu = 4,421$, $\sigma = 1,527$) melhor representa os dados originais. Estes resultados estão condizentes com os apresentados em [Veloso et al. 2006], que analisou fluxos de vídeo ao vivo gerados pelo provedor de conteúdo.

Ainda com relação aos tempos de sessão dos usuários verificamos qual a probabilidade de um usuário que já permaneceu x segundos permanecer por mais 60 segundos. A Figura 7 mostra esse resultado. Podemos observar que se um usuário permaneceu em uma transmissão da categoria curta por até 100 segundos, sua probabilidade de permanecer 1 minuto adicional é maior do que em transmissões da categoria longa. Após esse intervalo ocorre uma situação inversa. No geral, pode-se perceber que os usuários tendem a permanecer mais tempo em transmissões da categoria longa.

Esse resultado pode ser explorado, por exemplo, na escolha de parceiros em possíveis redes P2P, com a finalidade de servir esse tipo de aplicação, atribuindo uma prioridade maior na escolha de pares que possuem uma probabilidade maior de permanecer mais tempo na rede.



(a) Duração de transmissões menores que 20min.

(b) Duração de transmissões maiores que 20min.

(c) Histograma de duração de sessão

Figura 8. Tempo de sessões dos usuários

4.4.2. Comportamento do usuário

A Figura 8c mostra um histograma da duração das sessões dos usuários, onde ambos os eixos estão na escala logarítmica. Nela pode-se verificar a existência de várias sessões com um curto tempo de duração. Com o auxílio das figuras 8a e 8b (CDFs referentes às durações de sessões) podemos observar que as sessões identificadas em apenas um intervalo de amostra representaram aproximadamente 33% de todas as sessões.

Conjecturamos que existam dois tipos de comportamento do usuário no sistema, que foram denominados **comportamento de análise** e **visualização real**. O comportamento de análise, que poderia explicar essas durações curtas de sessões, se caracteriza pela ação de verificar o conteúdo de um determinado canal por certo período. Já o comportamento de visualização real indica que o usuário escolheu um canal para apreciar seu conteúdo.

Outros tradicionais grupos de pesquisa em multimídia também assumem a existência desses comportamentos. O grupo MSK da Coréia definiu 15 segundos como período de análise. [Cha et al. 2008] acredita que o tempo de análise seja aproximadamente 10 segundos, apesar de utilizar o valor de 1 minuto para fins de comparação com o valor definido por *Nilsen Media Research*[16]. Considerando todas as sessões com duração máxima de 60 segundos essa representatividade sobe para 53,46%.

Esse alto índice de baixos tempos de duração de sessões, que pode representar mudanças de canais, impõe desafios significativos na construção de sistemas P2P para prover infra-estrutura para esse tipo de serviço [Cha et al. 2008]. Para melhor ilustrar, considere que em análise de um popular sistema de transmissão de vídeo P2P, o tempo inicial de configuração[17] pode variar de 10 a 20 segundos para canais populares e mais de 2 minutos para os não populares [Hei et al. 2007].

Como se sabe, as redes P2P não oferecem uma estrutura estável. Quando um nó deixa a rede (interrompe a visualização de alguma transmissão), outros nós da rede podem ser desconectados ou sofrerem perda de qualidade no recebimento do conteúdo. Com um alto índice de sessões de curta duração esse tipo de situação tende a ser mais frequente. Por outro lado, há indícios de que um sistema P2P, para atender esse tipo de serviço nessas condições, pode obter sucesso caso haja nós que permaneçam em boa parte da duração da transmissão

[16]http://www.nielsenmedia.com

[17]intervalo entre a escolha do canal e o início da visualização de seu conteúdo

[Sripanidkulchai et al. 2004a], como foi identificado no sistema analisado.

5. Popularidade de canais

Nesta seção verificamos o que a popularidade dos canais representa na utilização sistema. Analisamos também os fatores que poderiam influenciar na popularidade, média de visualizadores simultâneos, de um canal. Por fim, verificamos as implicações inerentes aos canais com maior popularidade.

Para melhor entender a popularidade dos canais considere a Figura 9. Essa figura mostra o *ranking*[18] de popularidade dos canais (eixo x) ordenado pelo número médio de visualizadores simultâneos (eixo y). Ambos os eixos estão na escala logarítmica. Essa figura nos auxilia no entendimento de como a audiência é distribuída pelos canais. Identificamos que a distribuição de popularidade segue uma Zipf com 2 modos distintos. Esse resultado é semelhante ao de trabalhos que estudaram popularidade de fluxos de vídeos armazenados [Almeida et al. 2001, Chesire et al. 2001] e também ao vivo [Sripanidkulchai et al. 2004b].

Para os 350 canais mais populares, com um número médio de visualizadores entre 26 e 78, pode-se traçar uma reta na distribuição de popularidade, apresentando um comportamento Zipf, com $\alpha = 1,7$. Para o restante dos canais, que são menos populares, a popularidade também segue uma Zipf, porém com $\alpha = 2,6$.

5.1. Audiência

Esta seção analisa a participação dos usuários nos grupos de canais mais populares. Na Figura 10a, eixo y, é mostrado a porcentagem de usuários identificados no sistema. O eixo x representa grupos de popularidade.

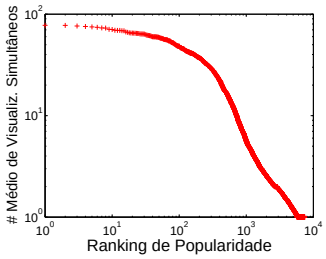
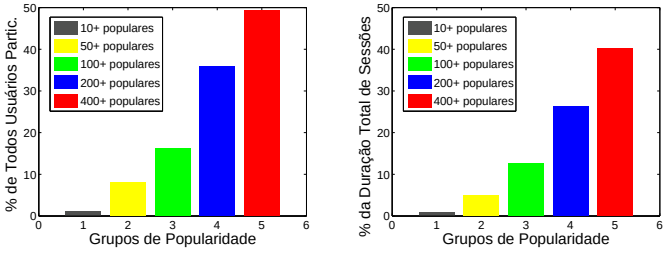


Figura 9. Ranking de popularidade

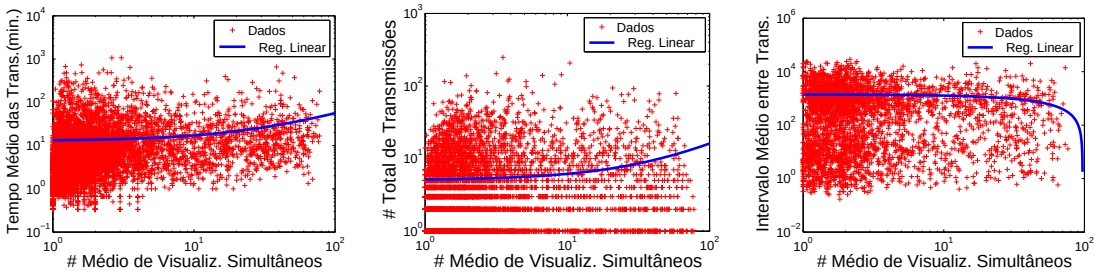


(a) # de usuários (b) tempo de utilização

Figura 10. Representatividade dos populares

Com base nessa figura pode-se perceber a importância de se conceder uma atenção especial aos canais mais populares. Como pode ser observado, os 400 canais mais populares (∼5,38% de todos os canais que realizaram alguma transmissão ao vivo durante a coleta), representaram 49,3% do número total de usuários que participaram de alguma transmissão. A Figura 10b que mostra a porcentagem do tempo de utilização, por todos os usuários de um canal, eixo y, e grupos de popularidade, eixo x, também reforça essa importância. O tempo total de utilização, soma dos tempos de sessões de todos os usuários participantes de alguma transmissão, equivale ao período de 8488,4 dias. Para esse cálculo desconsideramos os tempos de sessões dos donos dos canais. Pode-se perceber que os 400 canais mais populares atraíram usuários por um tempo suficiente para representar 40,1% de todo tempo de utilização observado.

[18] O canal mais popular se localiza no ponto $x = 1$, e o *ranking* aumenta da esquerda para direita.



(a) Média do tempo de transmissão

(b) Num. de transmissões realizadas

(c) Intervalo médio entre transmissões

Figura 11. Análise de fatores que poderiam influenciar na popularidade

A fim de analisar os fatores que poderiam influenciar na popularidade serão utilizadas as Figuras 11a, b e c. Todas elas estão na escala logarítmica. Nas três figuras são mostradas as relações entre o número médio de visualizadores simultâneos, eixo x, com o tempo médio de todas as transmissões realizadas por um determinado canal, o número total de transmissões realizadas por um canal e o intervalo médio de transmissões sucessivas de um mesmo canal no eixo y das Figuras 11a, b e c respectivamente.

Uma inspeção visual das figuras mostra que não existe forte correlação entre os fatores analisados com a média de usuários simultâneos de um canal. Isso pode ser comprovado pelos baixos coeficientes de correlações $r = 0,1$, $r = 0,1$, $r = -0,005$ referentes às análises mostradas nas Figuras 11a, b e c respectivamente. Verificamos também que não existe forte relação entre o tempo total que um canal permaneceu ao vivo, com a sua popularidade ($r = 0,05$). Não mostraremos esse resultado devido a limitações de espaço.

Apenas através desses fatores não foi possível identificar os canais mais populares. Acreditamos que a popularidade dos canais pode estar associada a outros fatores como, por exemplo, qualidade técnica e artística, do conteúdo transmitido.

5.2. As sessões são mais longas nos canais populares?

Para essas análises não consideramos o tempo de permanência do dono do canal na sua própria transmissão, uma vez que os donos dos canais sempre estão presentes na própria transmissão. Desconsideramos também canais que registraram menos de 10 visitantes em suas transmissões, com a finalidade de obter uma análise mais representativa.

Para todos os canais restantes calculamos qual a probabilidade de um usuário permanecer por um tempo superior a 1 minuto. Esse resultado é apresentado no eixo y da Figura 12. O eixo x, dessa mesma figura, representa o *ranking* de popularidade. Observamos que existe relação entre a popularidade e o tempo de permanência dos usuários. Canais mais populares apresentam probabilidades maiores para tempos de sessões superiores a 1 minuto. O coeficiente de correlação entre a posição do canal no ranking de popularidade e as durações das sessões foi bastante expressivo, sendo igual a $-0,6$. Note que canais mais populares ocupam posições mais altas, e, logo, menores no *ranking*. Por isto, a correlação tem valor negativo.

Esse resultado reforça a importância de se dedicar uma atenção especial aos canais populares. Nesse caso em especial, é mostrado que deve-se esperar que, além de mais visitantes simultâneos, os canais mais populares utilizem mais recursos do sistema, devido

ao maior tempo de permanência de seus usuários.

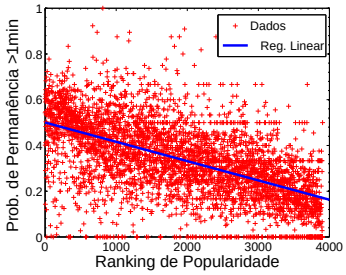


Figura 12. Probab. de permanência maior que 1min.

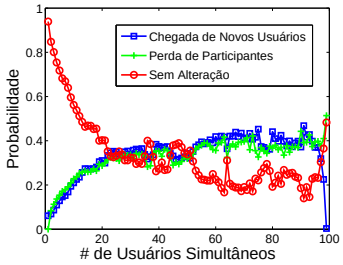


Figura 13. Influência do num. de usuários simultâneos

5.3. O número de usuários simultâneos é um fator que atrai usuários?

Nessa subseção verificamos se o número de visitantes simultâneos atual, de um determinado canal, é um fator que atrai novos usuários. Supondo que em determinada transmissão, no instante t, tenham sido registrado 3 usuários A, B, C e no instante $t + 1$ tenham sido registrado 4 usuários A, B, C, D. Assim, em nossas análises seria identificado o acréscimo de 1 usuário a partir do momento que havia 3 usuários simultâneos. Com esse mesmo procedimento é possível registrar os momentos em que a transmissão não sofreu alteração ou apresentou perdas de usuários a partir de u usuários simultâneos.

A figura 13 mostra o resultado dessa análise. O eixo x representa o número de usuários simultâneos, referentes a todos os canais e transmissões observadas, e o eixo y uma probabilidade. São exibidas 3 curvas, indicando a não alteração, aumento e perda de usuários simultâneos. Podemos observar que a partir de aproximadamente 16 usuários simultâneos a probabilidade de aumentar o número de usuários no instante seguinte, para a maioria dos casos, é maior do que a de diminuir o número de usuários em um determinado canal. Isso dá sinais que o número de espectadores atual de um canal é um fator que atrai novos usuários. Saber que um canal, com um certo número de usuários simultâneos, tende a aumentar seu número de participantes pode ser útil, por exemplo, para se alocar ou desalocar recursos do sistema dinamicamente, com o propósito de atender esse determinado canal.

6. Considerações Finais

Nosso trabalho é o primeiro passo para entender o comportamento dos usuários em um tipo de aplicação que vem recebendo grande atenção ultimamente, os sistemas para transmissão de fluxo de vídeo ao vivo, gerados e transmitidos apenas pelos usuários, baseados na Web. A caracterização foi realizada por um período de 23 dias e foi analisado mais de 1 milhão de sessões de usuários distribuídas em mais de 7 mil canais de transmissão.

Nossa contribuição pode ser dividida em duas partes. A primeira, referente a uma caracterização do modelo de interação dos usuários, pode ser resumida em: A duração das transmissões segue uma distribuição Lognormal. O intervalo médio entre transmissões sucessivas de um mesmo canal é melhor descrito por uma distribuição Lognormal. O número de transmissões realizadas pelos canais segue uma distribuição Weibull. O processo de chegada dos usuários é uniformemente distribuído ao longo da duração de uma transmissão. A duração do tempo de sessão dos usuários é melhor representada por uma distribuição Lognormal (tanto para transmissões curtas como longas).

Já a segunda parte do trabalho, onde foi analisada a distribuição de popularidade dos canais e suas implicações no sistema, pode ser resumida em: A distribuição de popularidade

é representada por uma Zipf de dois modos distintos. Usuários tendem a permanecer mais tempo em canais populares. O número de usuários simultâneos em um determinado canal influencia na atração de novos usuários.

O resultado do nosso trabalho proporciona diversas direções possíveis de se seguir como trabalhos futuros. Uma delas seria a geração de cargas sintéticas mais realistas, com a finalidade de analisar a aplicabilidade de arquitetura de rede P2P para atender um serviço semelhante ao analisado.

Referências

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723.

Almeida, J. M., Krueger, J., Eager, D. L., and Vernon, M. K. (2001). Analysis of educational media server workloads. In *NOSSDAV '01: Proc. of the 11th inter. work. on Network and op. sys. sup. for digital audio and video*, pages 21–30, NY, USA. ACM.

Cha, M., Rodriguez, P., Crowcroft, J., Moon, S., and Amatriain, X. (2008). Watching television over an ip network. In *IMC '08: Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*, pages 71–84, New York, NY, USA. ACM.

Chesire, M., Wolman, A., Voelker, G. M., and Levy, H. M. (2001). Measurement and analysis of a streaming-media workload. In *USITS'01: Proc. of the 3rd conf. on USENIX Symp. on Internet Tech. and Sys.*, pages 1–1, Berkeley, CA, USA. USENIX Assoc.

Duarte, F., Benevenuto, F., Almeida, V., and Almeida, J. (2007). Geographical characterization of youtube: a latin american view. In *LA-WEB '07: Proc. of the 2007 Latin American Web Conf.*, pages 13–21, Washington, DC, USA. IEEE Computer Society.

Gill, P., Arlitt, M., Li, Z., and Mahanti, A. (2007). Youtube traffic characterization: a view from the edge. In *IMC '07: Proc. of the 7th ACM SIGCOMM conf. on Internet measurement*, pages 15–28, New York, NY, USA. ACM.

Hei, X., Liang, C., Liang, J., Liu, Y., and Ross, K. (2007). A Measurement Study of a Large-Scale P2P IPTV System. *Multimedia, IEEE Transactions on*, 9(8):1672–1687.

Li, B., Xie, S., Keung, G., Liu, J., Stoica, I., Zhang, H., and Zhang, X. (2007). An empirical study of the coolstreaming+ system. *Selected Areas in Communications, IEEE Journal on*, 25(9):1627–1639.

Maia, M., Almeida, V., and Almeida, J. (2008). Vídeo gerados por usuários: Caracterização de tráfego. In *XXVI Simpósio Brasileiro de Redes de Computadores, SBRC 2008*, Belo Horizonte, Brasil.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47(1):90–100.

Oreilly, T. (2007). What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *Communications and Strategies, No. 1,*, page p. 17.

Silerston, T. and Fourmaux, O. (2007). Measuring p2p iptv systems. In *Proceedings of NOSSDAV'07*.

Sripanidkulchai, K., Ganjam, A., and Maggs, B. (2004a). The feasibility of supporting large-scale live streaming applications with dynamic application end-points. In *In Proc. of ACM SIGCOMM*, pages 107–120.

Sripanidkulchai, K., Maggs, B., and Zhang, H. (2004b). An analysis of live streaming workloads on the internet. In *IMC '04: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pages 41–54, New York, NY, USA. ACM.

Veloso, E., Almeida, V., Wagner Meira, J., Bestavros, A., and Jin, S. (2006). A hierarchical characterization of a live streaming media workload. volume 14, pages 133–146, NJ, USA. IEEE.