



Ministério da  
Ciência e Tecnologia



sid.inpe.br/mtc-m19/2010/10.27.12.54-RPQ

## ANÁLISE DE PADRÕES EM CHAMADAS TELEFÔNICAS

Giscard Fernandes Faria

Relatório final da disciplina Princípios e Aplicações de Mineração de Dados  
(CAP-359) do Programa de Pós-Graduação em Computação Aplicada, ministrada  
pelo professor Rafael Santos.

URL do documento original:

<<http://urlib.net/8JMKD3MGP7W/38FUMP2>>

INPE  
São José dos Campos  
2010

**PUBLICADO POR:**

Instituto Nacional de Pesquisas Espaciais - INPE

Gabinete do Diretor (GB)

Serviço de Informação e Documentação (SID)

Caixa Postal 515 - CEP 12.245-970

São José dos Campos - SP - Brasil

Tel.:(012) 3208-6923/6921

Fax: (012) 3208-6919

E-mail: pubtc@sid.inpe.br

**CONSELHO DE EDITORAÇÃO E PRESERVAÇÃO DA PRODUÇÃO INTELLECTUAL DO INPE (RE/DIR-204):****Presidente:**

Dr. Gerald Jean Francis Banon - Coordenação Observação da Terra (OBT)

**Membros:**

Dr<sup>a</sup> Inez Staciarini Batista - Coordenação Ciências Espaciais e Atmosféricas (CEA)

Dr<sup>a</sup> Maria do Carmo de Andrade Nono - Conselho de Pós-Graduação

Dr<sup>a</sup> Regina Célia dos Santos Alvalá - Centro de Ciência do Sistema Terrestre (CST)

Marciana Leite Ribeiro - Serviço de Informação e Documentação (SID)

Dr. Ralf Gielow - Centro de Previsão de Tempo e Estudos Climáticos (CPT)

Dr. Wilson Yamaguti - Coordenação Engenharia e Tecnologia Espacial (ETE)

Dr. Horácio Hideki Yanasse - Centro de Tecnologias Especiais (CTE)

**BIBLIOTECA DIGITAL:**

Dr. Gerald Jean Francis Banon - Coordenação de Observação da Terra (OBT)

Marciana Leite Ribeiro - Serviço de Informação e Documentação (SID)

Deicy Farabello - Centro de Previsão de Tempo e Estudos Climáticos (CPT)

**REVISÃO E NORMALIZAÇÃO DOCUMENTÁRIA:**

Marciana Leite Ribeiro - Serviço de Informação e Documentação (SID)

Yolanda Ribeiro da Silva Souza - Serviço de Informação e Documentação (SID)

**EDITORAÇÃO ELETRÔNICA:**

Vivéca Sant´Ana Lemos - Serviço de Informação e Documentação (SID)



Ministério da  
Ciência e Tecnologia



sid.inpe.br/mtc-m19/2010/10.27.12.54-RPQ

## ANÁLISE DE PADRÕES EM CHAMADAS TELEFÔNICAS

Giscard Fernandes Faria

Relatório final da disciplina Princípios e Aplicações de Mineração de Dados (CAP-359) do Programa de Pós-Graduação em Computação Aplicada, ministrada pelo professor Rafael Santos.

URL do documento original:

<<http://urlib.net/8JMKD3MGP7W/38FUMP2>>

INPE  
São José dos Campos  
2010



*“Quando estás certo, ninguém se lembra; quando estás errado, ninguém esquece”.*

*Provérbio irlandês*



## **RESUMO**

Atualmente são realizadas no mundo mais que um bilhão de chamadas telefônicas todos os dias. Assim o registro destas chamadas são armazenados em equipamentos da rede de telefonia e no final são utilizados principalmente para cobrança. Contudo quando tendo todos esses registros agrupados é possível realizar um estudo aprofundado do comportamento de cada assinante; bem como encontrar relações entre os diversos assinantes e fluxo de tráfego na rede de telefonia.



## **ANALISING PATTERNS IN TELEPHONE CALLS**

### **ABSTRACT**

The World wide population generate nowadays more than one billion of phone calls every day. All those call records are stored in network equipments and after all are used mostly for billing purpose only. However, putting all those calls together allows to perform a deep study of the behavior of each telephone subscriber, find out relationship between subscriber and the network traffic flow as well.



## LISTA DE FIGURAS

	<u>Pág.</u>
Figura 1 - Tipo de Cobrança x Categoria do Assinante .....	16
Figura 2 - Tipo de Cobrança x Horário da Chamada .....	17
Figura 3 - Tipo de Cobrança x Duração da Chamada .....	18
Figura 4 - Tipo de Cobrança x Tipo da Chamada .....	19
Figura 5 - Tipo de Cobrança X Classe da Chamada.....	20
Figura 6 - Categoria do Assinante A x Horário da Chamada.....	21
Figura 7 - Categoria do Assinante x Duração da Chamada.....	22
Figura 8 - Categoria do Assinante x Tipo da Chamada .....	23
Figura 9 - Categoria do Assinante x Tipo da Chamada .....	24
Figura 10 - Categoria do Assinante x Tipo de Acesso.....	25
Figura 10 – Horário da Chamada x Tipo da Chamada.....	26
Figura 11 – Horário da Chamada x Tronco de Entrada.....	27
Figura 12 – Horário da Chamada x Tronco de Saída.....	28
Figura 13 – Tipo da Chamada x Tipo de Serviço .....	29
Figura 14 – Tipo da Chamada x Tipo de Acesso .....	30
Figura 15 – Classe da Chamada x Tronco de Entrada.....	31
Figura 16 – Classe da Chamada x Tronco de Saída.....	32
Figura 17 – Classe da Chamada x Código de Falha.....	33
Figura 18 – Classe da Chamada x Tipo de Serviço .....	34
Figura 19 – Classe da Chamada x Tipo de Acesso .....	35
Figura 8 - Grafo de Relação dos Assinantes .....	42



## LISTA DE TABELAS

	<u>Pág.</u>
Tabela 1 - Tipos de CDRs .....	Erro! Indicador não definido.
Tabela 2 - Campos dos CDRs.....	Erro! Indicador não definido.
Tabela 3 - Campos do CDR final .....	<b>6</b>
Tabela 4 - Campos do CDR após Seleção.....	Erro! Indicador não definido.
Tabela 5 - Tipos de Dados por Campos .....	Erro! Indicador não definido.
Tabela 6 - Tarifas Telefônicas .....	Erro! Indicador não definido.
Tabela 7 - Atributos do Cluster I .....	Erro! Indicador não definido.
Tabela 8 - Atributos Classificadores .....	Erro! Indicador não definido.



## LISTA DE SIGLAS E ABREVIATURAS

INPE	Instituto Nacional de Pesquisas Espaciais
SID	Serviço de Informação e Documentação
TDI	Teses e Dissertações Internas
SPG	Serviço de Pós-Graduação
VoIP	Voice Over IP
ITU	International Telecommunications Union
CDR	Call Detail Record
CRM	Customer Relationship Management
Q&A	Quality Assurance
PBX	Private Branch Exchange



## SUMÁRIO

		<u>Pág.</u>
1	INTRODUÇÃO .....	1
2	TELEFONIA .....	2
3	INFORMAÇÕES CONTIDAS NOS CDRs .....	4
4	SELEÇÃO DE DADOS .....	9
5	PRÉ-PROCESSAMENTO DOS DADOS .....	12
6	MINERAÇÃO AD HOC DOS DADOS .....	16
7	CLUSTERIZAÇÃO.....	37
8	MINERAÇÃO VIA CLASSIFICAÇÃO .....	39
9	RELAÇÃO ENTRE ASSINANTES .....	41
10	CONCLUSÃO .....	45
11	PRÓXIMOS PASSOS .....	49



# 1 INTRODUÇÃO

Comunicação trata-se de de uma das características mais antigas do mundo animal (isso é real, estudos já mostram que mesmo macacos e outros animais possuem um meio efetivo de comunicação apesar de não tão complexo ou avançado quanto dos seres humanos); o modelo de comunicação do Homem evoluiu em muito no decorrer de sua história, passando por ruídos, expressão corporal, desenhos e por fim a criação do alfabeto. Em paralelo foram criadas também redes de comunicação, que consiste em criar uma infraestrutura ou modelos que permitia a transmissão de informações de modo eficiente, íntegro e seguro, estas redes também evoluíram no decorrer de nossa história; inicialmente os meios eram restritos a grupos que estavam próximos, depois foram criadas estruturas para o envio e recebimento de mensagens (fossem elas escritas em pedra, madeira ou papel), e por fim o modelo de comunicação atual que consiste em conectarem-se pessoas praticamente em tempo real para que sejam transmitidas as informações.

Este artigo possui como principal objetivo apontar (mesmo que de forma simplificada) algumas estratégias básicas para que se possam obter informações estratégicas analisando os dados gerados durante a comunicação entre duas ou mais pessoas. Devido à restrição de acesso aos dados de chamadas telefônicas (já que o sigilo telefônico é garantido por lei), foram utilizados nesta pesquisa os dados coletados durante um dia para uma única central de telefone (equipamento responsável por controlar a chamada telefônica); também todos os números de telefone aqui utilizados são fictícios e foram alterados para que oculta-se a verdadeira identidade dos assinantes.

## 2 TELEFONIA

O conceito de Telefonia, nada mais é que uma rede de comunicação que fornece infra-estrutura para que as pessoas possam se comunicar de maneira efetiva e praticamente instântanea independente de localidade e distância entre os interessados. Atualmente, com o surgimento da Internet e as tecnologias VOIP (Voice over IP), o conceito de telefonia é ainda mais abrangente, pois envolve não somente o ato de se comunicar através de um telefone, mas também a troca de dados dos mais diversos (incluindo voz, video, mensagens de texto, arquivos, etc). Não existe dúvida alguma hoje, que as operadoras de Telefonia (empresas que provêm a infra-estrutura para o uso de telefone) serão também as provedoras da infra-estrutura para que sejam transmitidos também quaisquer outros tipos de dados futuramente.

Estudos do ITU (International Telecommunications Union) datados de 2008 mostram que 60% da população mundial utilizam a Internet como meio de comunicação, e o que chama mais atenção que 97% da população mundial fazem uso da Rede de Telefonia móvel (comunicação através de celulares). É fácil, portanto, perceber que os dados gerados a partir destas redes são não somente extremamente volumoso, como também de extrema importância e riqueza – já que os mesmos apontam o comportamento de grande parte da população, ao menos no que diz respeito em modo de se comunicar.

Os dados das chamadas, e que são gerados durante a comunicação seja por celulares, computadores ou telefones convencionais são denominados CDR (Call Detail Record). Os CDRs são coletados, e utilizados principalmente pelas operadoras para que seja possível efetuar uma cobrança pelo serviço prestado ao assinante (pessoa que utiliza a rede para se comunicar); no entanto os mesmos não estão restritos a emissão de faturas, os mesmos contém informações da Rede (como por exemplo, motivo de falha para uma chamada não completada, rotas de comunicação utilizadas, etc).



### 3 INFORMAÇÕES CONTIDAS NOS CDRs

Como comentado anteriormente, os CDRs são utilizados para armazenar o registro de uma sessão (ou chamada) de comunicação entre assinantes. Contudo, por existirem diversos meios de comunicação (e.g.: telefones fixos, celulares, computadores, etc), as informações contidas nos CDRs podem variar (e de fato variam) de acordo com a necessidade de cada meio de utilizado. Além disso, mesmo quando se trata de meios de comunicação idênticos os dados contidos nos CDRs também podem variar (seja por decisão e estratégia da operadora ou do fornecedor do equipamento onde as chamadas são controladas). Neste artigo, serão retratadas as informações adquiridas de CDRs do equipamento de um fornecedor em específico, para uma operadora em específico e também para uma tecnologia em específico (telefonia fixa).

Para o caso analisado neste artigo, foi possível perceber três variantes de CDRs (ou seja, os CDRs também podem ser distintos dentro de uma mesma operadora, para um mesmo meio de comunicação onde todas as chamadas passam por um mesmo equipamento), são elas:

Tabela 1 - Tipos de CDRs

<b>Tipo de CDR</b>	<b>Descrição</b>
80 bytes	Contém uma quantidade restrita de dados, com o objetivo de diminuir espaço de armazenagem em disco (utilizados para chamadas comuns entre assinantes).
128 bytes	Contém todos os dados (utilizados para chamadas comuns entre assinantes).
192 bytes	Contém todos os dados (utilizados para chamadas via rede inteligente, por exemplo: 0500, 0800, etc).

Cada um dos CDRs mencionados na tabela acima, possuem diversos campos (que nada mais são que os dados das chamadas). Está fora do escopo

deste artigo entrar em detalhe e explicar o significado de cada campo, contudo para que se tenha uma idéia geral pode-se citar alguns exemplos:

Tabela 2 - Campos dos CDRs

Tipo de CDR	Qtd. Campos	Exemplos
80 bytes	18	Assinante Originante Hora da Chamada
128 bytes	31	Assinante Originante Hora da Chamada Tipo do Acesso
192 bytes	45	Assinante Originante Hora da Chamada Tipo do Acesso Identificador do Dígito de Extensão

A tabela acima não possui como objetivo esclarecer o significado dos campos dos CDRs, o importante de se notar na mesma é que existem campos que são exclusivos de alguns CDRs, e que também existem campos que são comuns a todos eles. A importância desta propriedade, consiste no modelo aplicado para uma primeira filtragem dos dados a serem observados em uma análise, em suma foram inicialmente utilizados apenas os dados que são comuns a todos os tipos de CDRs.

$$CDR_f = CDR_{80} \cap CDR_{128} \cap CDR_{192}$$

$CDR_f \equiv$  Campos do CDR final

$CDR_{80} \equiv$  Campos do CDR de 80 bytes

$CDR_{128} \equiv$  Campos do CDR 128 bytes

$CDR_{192} \equiv$  Campos do CDR 192 bytes

Com base no critério de intersecção anteriormente apresentados, teríamos no final o conjunto  $CDR_f$ , composto por 18 (dezoito) campos, são eles:

Tabela 1 - Campos do CDR

<b>Campo</b>	<b>Descrição</b>
Número de A	Assinante que iniciou a chamada telefônica
Número de B	Assinante que recebeu a chamada telefônica
Tipo de Cobrança	Chamada normal ou à cobrar.
Categoria do Assinante A	Nenhuma Assinante Comum Assinante com Tarifação Especial Equipamento de Manutenção Telefone Público Local Telefonista Equipamento de Comunicação de Dados Telefone Público Interurbano Chamada Transferida
Data da Chamada	Dia em que a chamada foi realizada
Hora da Chamada	Horário em que a chamada foi realizada (pode ser nenhum se a chamada não foi completada)
Duração da Chamada	Tempo que a chamada durou (pode ser nenhum se a chamada não foi completada)
Tipo da Chamada	Chamada não Completada Chamada Comum Chamada Estendida Chamada Estendida (fim)
Contador de Chamada Estendida	Contador de CDRs que a chamada foi bilhetada
Tronco de Entrada	Número do tronco por onde a chamada chegou à central
Circuito de Entrada	Número do circuito (do tronco) por onde a chamada chegou à central
Tronco de Saída	Número do tronco por onde a chamada saiu da central
Circuito de Saída	Número do circuito (do tronco) por onde a chamada saiu da central
Classe da Chamada	Chamada completada sem erro Chamada completada com erro Não completada B não atende B ocupado Timeout no tronco de entrada Congestionamento no destino Congestionamento na origem Timeout no junto de saída Erro no tronco de saída A desligou Número mudou Falha em B (tipo 10 ou 15)

Tabela 3 - Conclusão

Código de Falha da Chamada	Nenhum código (chamada correta) FDS=6 Número de A menor que mínimo Número de A maior que máximo Duração maior que máximo Assinante de Tarifação especial Equipamento em manutenção Telefone público Telefonista Equipamento de Comunicação de Dados Outras categorias de assinante
Tipo de Serviço	Sem informação Voz 64 Kbps (dados) 3,1Khz
Tipo do Acesso	Sem informação Analógico Digital
Número Sequencial	Identificador único do CDR

É perceptível que entender detalhadamente do que se trata cada campo, é um trabalho especializado e de responsabilidade do engenheiro de telecomunicações (por exemplo, o Número de A possui um limite máximo de tamanho, o mesmo é válido para o número de B). No entanto, grandes partes dos valores são de conhecimento geral a todos os assinantes que utilizam uma rede de telefonia – o que vai facilitar o entendimento deste artigo.



## 4 SELEÇÃO DE DADOS

Os dados de chamadas telefônicas conseguidos para esse experimento consistem; como comentado anteriormente, de chamadas telefônicas realizadas por uma única operadora em uma única central de uma fornecedora em específico durante o período de um único dia. Apesar de extremamente limitado e focado, o volume do dia chegou a 216.223 chamadas. O Alto volume de dados, somados a extensa quantidade de atributos exigiria uma grande capacidade de processamento para efetuar análise e similaridades de dados através das mais diversas técnicas de Data Mining. Portanto, mesmo que tenhamos efetuado uma redução de atributos (ao criar um conjunto final apenas considerando a intersecção de todos os tipos de CDR) ainda se faz necessário realizar uma pré-análise para maior redução de dados com nenhum ou pouco valor.

Primeiramente foi eliminado o campo que contém a “*Data da Chamada*” em que a chamada foi efetuada (apesar de se tratar de um atributo importante para qualquer análise de CDR), para o conjunto de dados em questão o mesmo não possui valor já que todas as chamadas foram efetuadas no mesmo dia.

Em seguida foi eliminado o campo que contém o “*Número Sequencial*” dos CDRs, já que este atributo é distinto para todo e qualquer chamada o mesmo não permitirá criar correlações com outros dados.

A mesma lógica aplica-se aos *Circuitos de Entrada e Saída*, analisando os mesmos é possível verificar que o mesmo não se repete e, portanto, impedem que seja criada uma correlação (esse atributo em específico poderia ser descartado mesmo que algumas poucas repetições ocorressem, já que os circuitos são utilizados ou de forma aleatória ou sequencial em uma central telefônica).

Os campos número de A e número de B possuem um comportamento similar (ou seja, em sua maioria não se repetem) e, portanto dificilmente serão correlacionados com os outros campos; no entanto, esses atributos não serão

removidos totalmente. Resumindo, eles serão removidos durante uma análise de data mining, no entanto serão utilizados para outra análise de padrões que consiste em encontrar ciclos de chamadas entre assinantes (ou seja, assinantes que possuem uma rede de relacionamento em comum).

Finalmente, pode-se excluir o campo que possui o “*Contador de Chamadas Estendidas*”, para isso basta acumular a duração das chamadas no campo “*Duração da Chamada*” e descartar o mesmo.

Após todas as exclusões, fica-se disponível nos CDRs os seguintes atributos (o que totaliza 13 campos):

Tabela 2 - Campos do CDR após Seleção

<b>Campo</b>	<b>Descrição</b>
Número de A	Assinante que iniciou a chamada telefônica
Número de B	Assinante que recebeu a chamada telefônica
Tipo de Cobrança	Chamada normal - “0” Chamada a cobrar - “1”
Categoria do Assinante A	Nenhuma – “?” Assinante Comum – “01” Assinante com Tarifação Especial – “02” Equipamento de Manutenção – “03” Telefone Público Local – “04” Telefonista – “05” Equipamento de Comunicação de Dados - “06” Telefone Público Interurbano - “07” Chamada Transferida – “11”
Horário da Chamada	Horário em que a chamada foi realizada (pode ser nenhum se a chamada não foi completada). Sem informação – “?” Super reduzida – “S” Reduzida – “R” Normal – “N” Diferenciada – “D”
Duração da Chamada	Tempo que a chamada durou (pode ser nenhum se a chamada não foi completada)
Tipo da Chamada	Chamada não Completada – “?” Chamada Comum – “0” Chamada Estendida – “1” Chamada Estendida (fim) – “2”
Tronco de Entrada	Número do tronco por onde a chamada chegou à central
Tronco de Saída	Número do tronco por onde a chamada saiu da central

Tabela 4 – Conclusão

Classe da Chamada	<p>Chamada completada sem erro – “10”                  Chamada completada com erro – “20”                  Não completada – “44”                  B não atende – “31”                  B ocupado – “32”                  Timeout no tronco de entrada – “40”                  Congestionamento no destino – “43”                  Congestionamento na origem – “46”                  Timeout no junto de saída – “48”                  Erro no tronco de saída – “49”                  A desligou – “50”                  Número mudou - “51”                  Falha em B (tipo 10 ou 15) – “52”</p>
Código de Falha da Chamada	<p>Nenhum código (chamada correta) – “?”                  FDS=6 – “03”                  Número de A menor que mínimo – “04”                  Número de A maior que máximo – “05”                  Duração maior que máximo – “06”                  Assinante de Tarifação especial – “08”                  Equipamento em manutenção – “09”                  Telefone público – “10”                  Telefonista – “11”                  Equipamento de Comunicação de Dados – “12”                  Outras categorias de assinante – “13”</p>
Tipo de Serviço	<p>Sem informação – “000”                  Voz – “001”                  64 Kbps (dados) - “002”                  3,1Khz – “003”</p>
Tipo do Acesso	<p>Sem informação – “00”                  Analógico – “01”                  Digital - “02”</p>

## 5 PRÉ-PROCESSAMENTO DOS DADOS

Antes que seja iniciada a análise dos dados, ainda se faz necessário a definição de alguns outros itens, o principal deles consiste em como será definido o tipo de cada dado. Basicamente pode-se optar por dois tipos distintos cada um com sua vantagem e desvantagem. O primeiro tipo seriam os dados numéricos, onde cada atributo pode ser um valor inteiro (ou real para outros casos) e que varia continuamente ou discretamente em um intervalo pré-definido. Enquanto o segundo trata-se dos dados literais, onde cada atributo possui um conjunto de valores onde apenas um será atribuído (e.g., Sexo: Masculino ou Feminino).

Para os atributos pós-seleção poderíamos optar pelo seguintes tipos de dados:

Tabela 5 - Tipos de Dados por campos

<b>Campo</b>	<b>Tipo de Dado</b>
Número de A	Literal
Número de B	Literal
Tipo de Cobrança	Literal
Categoria do Assinante A	Literal
Horário da Chamada	Literal ou Numérico
Duração da Chamada	Literal ou Numérico
Tipo da Chamada	Literal
Tronco de Entrada	Literal
Tronco de Saída	Literal
Classe da Chamada	Literal
Código de Falha da Chamada	Literal
Tipo de Serviço	Literal
Tipo do Acesso	Literal

Percebe-se que a maioria dos campos (ou atributos) só podem assumir valores literais (por pertencerem a uma classe pré-definida em específico). No

entanto, dois campos chamam a atenção por parecerem a primeira vista numérico e, no entanto foram descritos na tabelas anterior como podendo ser tanto numérico quando literal.

Atributos relacionados ao tempo (como hora) geralmente são numéricos e contínuos, contudo no contexto de telefonia o mesmo pode ser facilmente transformado para literal, a vantagem desta transformação consiste na simplificação e redução do domínio de valores do campo. Quando utilizado em chamads telefônicas, existem basicamente quatro tarifas de cobranças que são relacionadas com horários, são elas:

Tabela 6 - Tarifas Telefônicas

<b>Horário</b>	<b>Tarifa</b>
00:00 às 06:00	Super Reduzida
00:06 às 07:00 21:00 às 24:00	Reduzida
07:00 às 09:00 12:00 às 14:00 18:00 às 21:00	Normal
09:00 às 12:00 14:00 às 18:00	Diferenciada

Independente do horário que se realizou a chamada, no final o impacto no comportamento do assinante será dado pela tarifa a ser cobrada - é claro que esta simplificação reduz o escopo da análise, por exemplo, não se pode analisar comportamento de ligações com base em períodos do dia – no entanto a redução simplifica em muito a utilização do campo em si e permite fazer uma análise quanto a cobrança de forma direta e a baixo custo computacional.

Outro atributo a ser modificado, consiste no atributo de “*Duração da Chamada*”, as cobranças feitas pelas empresas prestadoras de serviço de telefonia ocorrem por minuto, o que permite descartar a quantidade de segundos despendida por chamada, esta simplificação permite, por exemplo, um maior agrupamento na duração das chamadas – e do mesmo modo que no atributo anterior, reduz algumas outras possibilidades de análise – o que

simplifica a utilização do campo e permite uma análise com menor custo computacional.

Por fim, existe outro grupo de dados que devem ser pré-processados antes que possa ser utilizado, são eles o “Número de A” e “Número de B”. Em um CDR de uma chamada telefônica, se o mesmo assinante inicia uma chamada telefônica ou recebe uma chamada telefônica, os valores guardados no campo são distintos. Por exemplo, vamos supor que o assinante Emanuel cujo número de telefone é 5555-3333 pertencente à cidade de São Paulo realizada uma ligação; em seu CDR vai aparecer o campo “Número de A” com o seguinte valor 1155553333 (e.g., o código de área de SP – 11 – adicionado do número de telefone), já quando Emanuel atende uma chamada telefônica interurbano de sua namorada que está em São José dos Campos através da operadora Super Telecom, no CDR vai aparecer o campo “Número de B” com o seguinte valor 0371155553333 (e.g., 0 indicando interurbano, 37 o código da operadora Super Telecom, 11 a área da região, e por fim o número de Emanuel). Para cada tipo de chamada temos um “Número de B” dinâmico, enquanto o “Número de A” permanece estático para qualquer chamada feita por Emanuel. Exemplos de outras chamadas são as locais, locais a cobrar, interurbano a cobrar. Apesar de mais complexa, existe uma regra bem simples que simplifica a relação entre o “Número de A” e o “Número de B”; o campo “Número de A” sempre vai conter o código de área da região mais o número final do assinante, e o “Número de B” sempre vai conter no final da “String” o código de área da região mais o número final do assinante; e é aqui onde se torna possível verificar se um assinante que realizou uma chamada também é o mesmo que recebeu outra.

O processo de simplificação das relações entre “Número de A” e “Número de B” vão bem além (por exemplo temos as chamadas internacionais, além de muitas outras), e a simplificação acima mencionada por um lado reduz em muito a complexidade na manipulação dos assinantes, contudo por outro lado não permite análises mais profundas como por exemplo tipo de uso das

chamadas com base no código da operadora (este tipo de análise fica para um futuro trabalho).

## 6 MINERAÇÃO AD HOC DOS DADOS

Nesta etapa foi realizada uma análise customizada para cada atributo, examinando assim todos os outros pares, com o intuito de encontrar alguma relação entre a distribuição. Esta não é de forma alguma a análise ideal, pois a mesma deve ser feita artesanalmente pessoas, no entanto esta análise cuidadosa permitiu encontrar alguns padrões importantes no uso da rede de telefonia.

### Analizando Tipo de Cobrança (Eixo X)

Categoria do Assinante (Eixo Y)

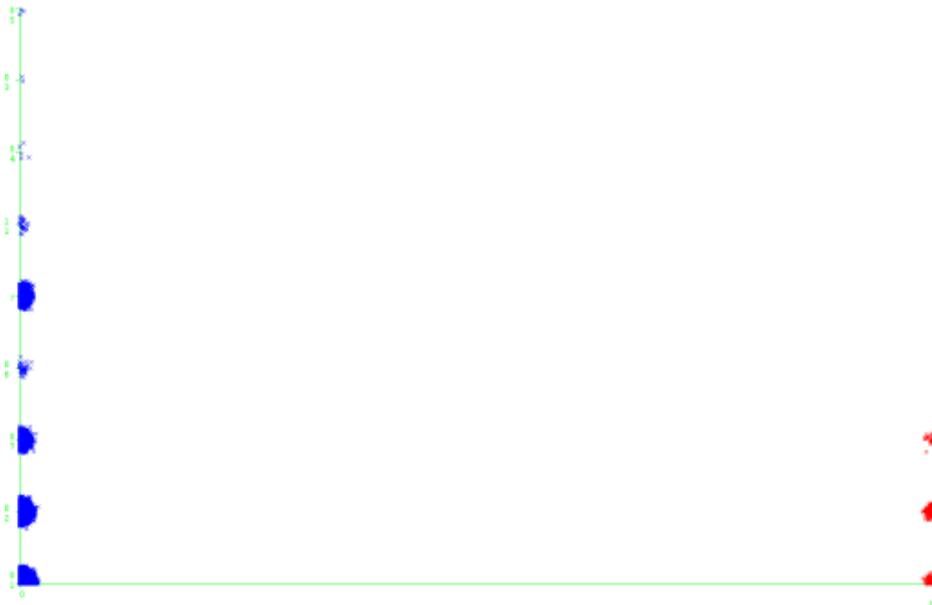


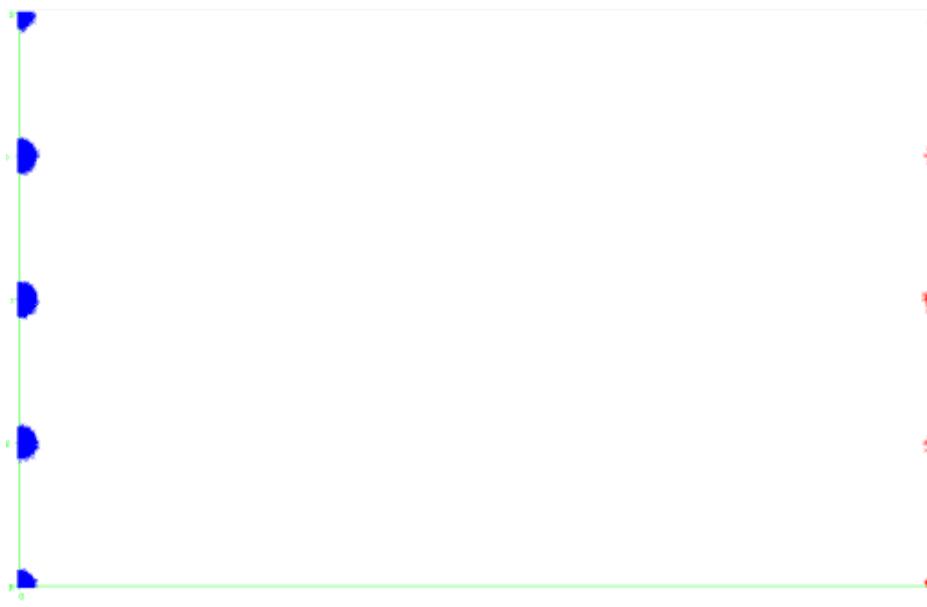
Figura 1 - Tipo de Cobrança x Categoria do Assinante

Através do gráfico plotado é possível verificar que:

- Não são realizadas chamadas a cobrar local através de telefones públicos;
- No entanto o volume de chamadas a cobrar interurbano pelos telefones públicos é bem alto;
- Telefonistas e equipamentos de testes não fazem chamadas à cobrar (o que já era esperado).

Os valores na legenda estão descritos na tabela **Erro! Fonte de referência não encontrada.**

Horário da Chamada (Eixo Y)



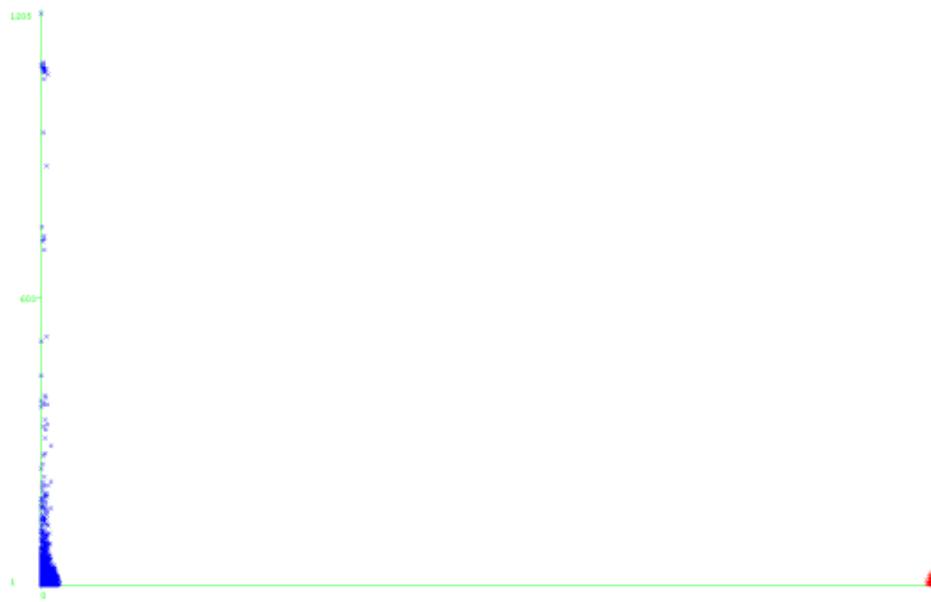
**Figura 2 - Tipo de Cobrança x Horário da Chamada**

Através do gráfico plotado é possível verificar que:

- A maioria das chamadas a cobrar é realizada nos horários mais custosos, ou seja, no horário “super reduzido” existem poucas chamadas a cobrar.

Os valores na legenda estão descritos na tabela **Erro! Fonte de referência não encontrada.**

### Duração da Chamada (Eixo Y)



**Figura 3 - Tipo de Cobrança x Duração da Chamada**

Através do gráfico plotado é possível verificar que:

- Só existem chamadas a cobrar com baixas durações;
- Chamadas comuns estão mais concentradas em baixas durações.

Os valores na legenda estão descritos na tabela **Erro! Fonte de referência não encontrada.**

Tipo de Chamada (Eixo Y)

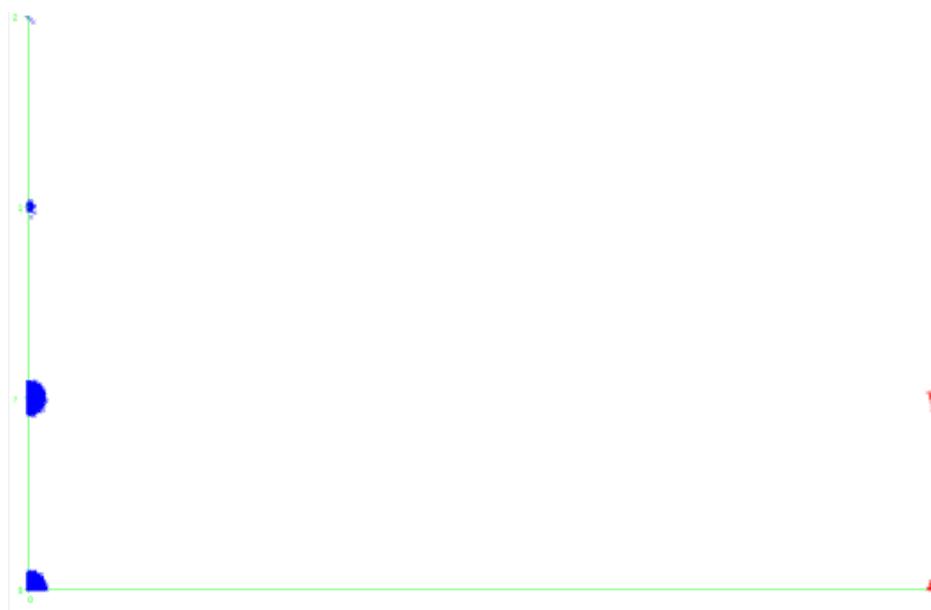


Figura 4 - Tipo de Cobrança x Tipo da Chamada

Através do gráfico plotado é possível verificar que:

- Não existem chamadas a cobrar estendidas (o que está de acordo com a análise anterior).

Os valores na legenda estão descritos na tabela **Erro! Fonte de referência não encontrada.**

Classe da Chamada (Eixo Y)

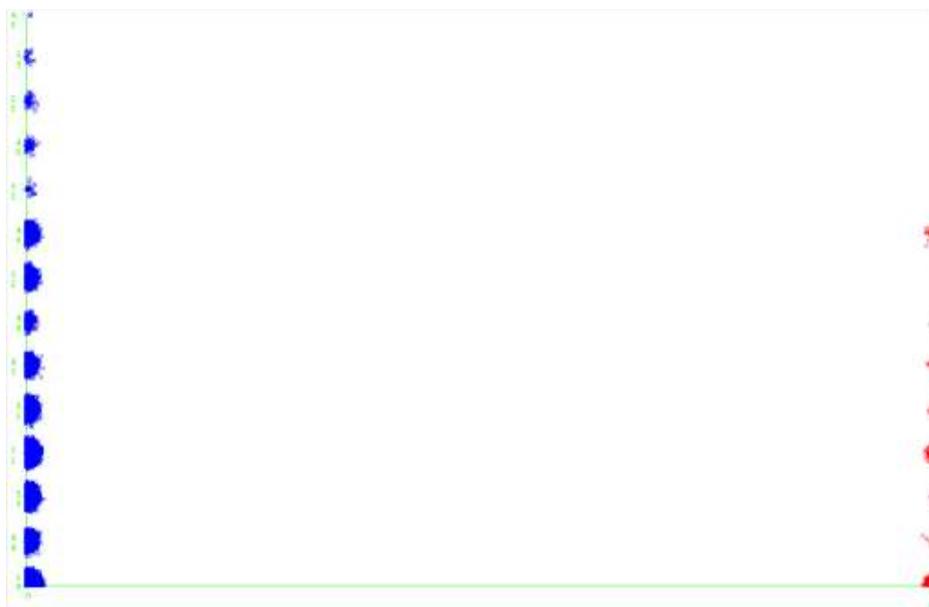


Figura 5 - Tipo de Cobrança X Classe da Chamada

Através do gráfico plotado é possível verificar que:

- A maior parte das chamadas a cobrar é completada;
- As que não são completadas se devem ao fato do número chamado não estar disponível para atender;
- O que mostra ainda um grande aceite dos usuários da rede de telefonia em atender chamadas à cobrar.

Os valores na legenda estão descritos na tabela **Erro! Fonte de referência não encontrada.**

## Analizando Categoria do Assinante A (Eixo X)

Horário da Chamada (Eixo Y)

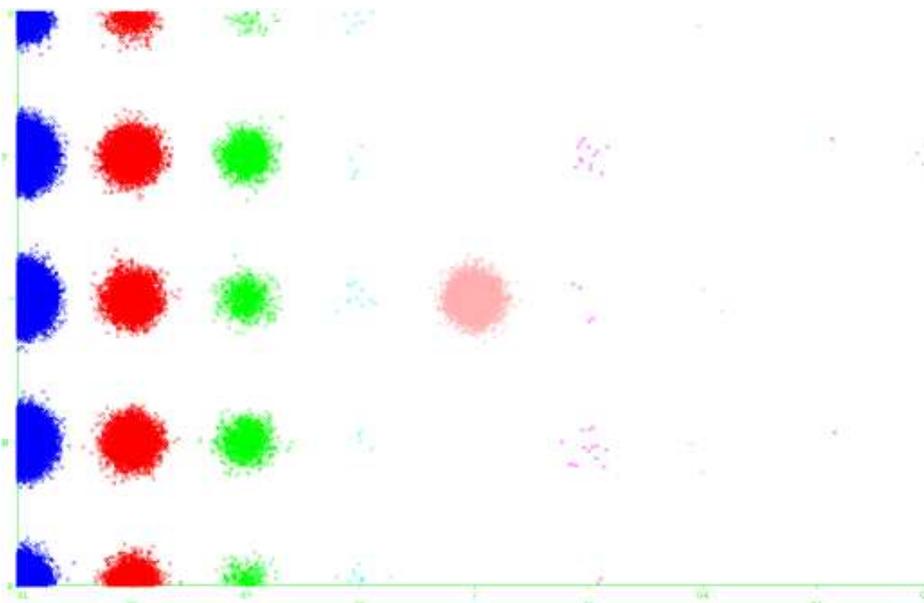


Figura 6 - Categoria do Assinante A x Horário da Chamada

Através do gráfico plotado é possível verificar que:

- O uso de telefones públicos para chamadas interurbanas são menores nos horários mais reduzidos;
- Serviços de transferência de chamadas só são utilizados em horários comerciais, o que indica que o serviço não é utilizado em contexto pessoal, mas à trabalho apenas;
- Chamadas para telefonistas ocorrem apenas em horários diferenciados;
- Equipamentos de manutenção só são utilizados em horários comerciais (como já era esperado).

Os valores na legenda estão descritos na tabela **Erro! Fonte de referência não encontrada.**

### Duração da Chamada (Eixo Y)



**Figura 7 - Categoria do Assinante x Duração da Chamada**

Através do gráfico plotado é possível verificar que:

- Apenas assinantes comuns fazem chamadas com grande duração.

Os valores na legenda estão descritos na tabela **Erro! Fonte de referência não encontrada.**

### Tipo de Chamadas (Eixo Y)

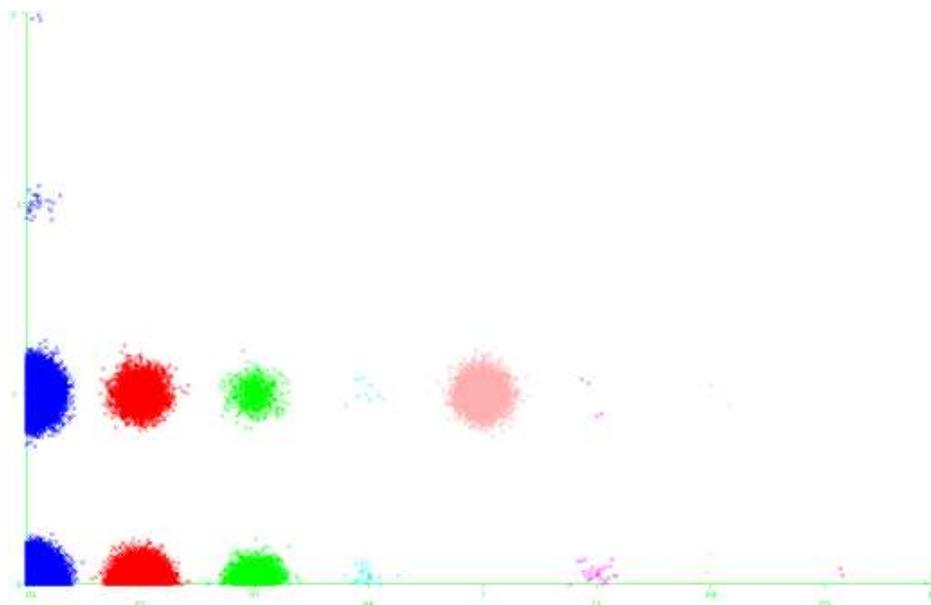


Figura 8 - Categoria do Assinante x Tipo da Chamada

Através do gráfico plotado é possível verificar que:

- Apenas assinantes comuns fazem chamadas estendidas (o que está de acordo com a análise anterior).

Os valores na legenda estão descritos na tabela **Erro! Fonte de referência não encontrada.**

Classe da Chamada (Eixo Y)

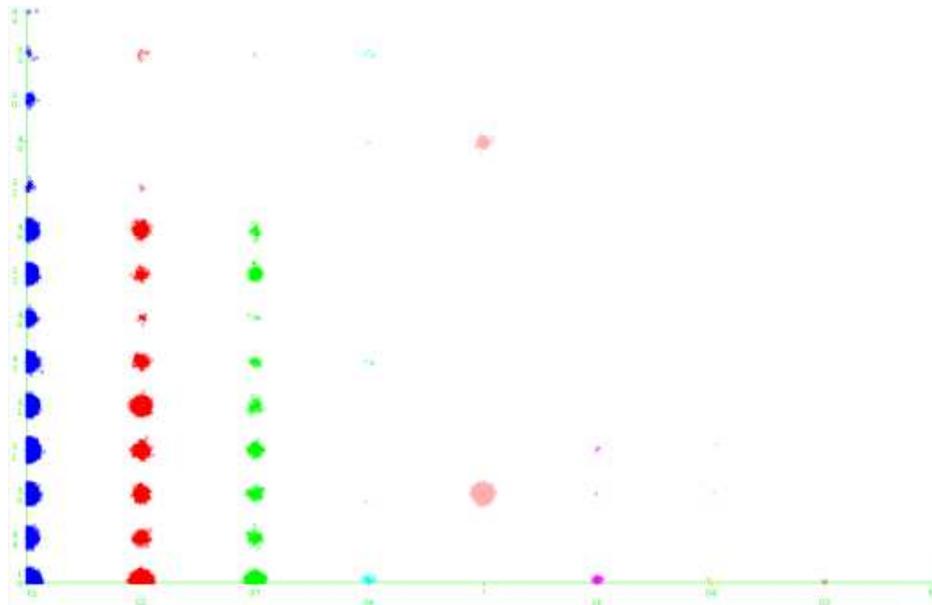


Figura 9 - Categoria do Assinante x Tipo da Chamada

Através do gráfico plotado é possível verificar que:

- Chamadas locais de telefones públicos não completados se devem ao fato do assinante B não estar disponível ou então por desistência do assinante realizando a chamada;
- Equipamentos de manutenção e chamadas a telefonistas sempre são completadas.

Os valores na legenda estão descritos na tabela **Erro! Fonte de referência não encontrada.**

Tipo de Acesso (Eixo Y)

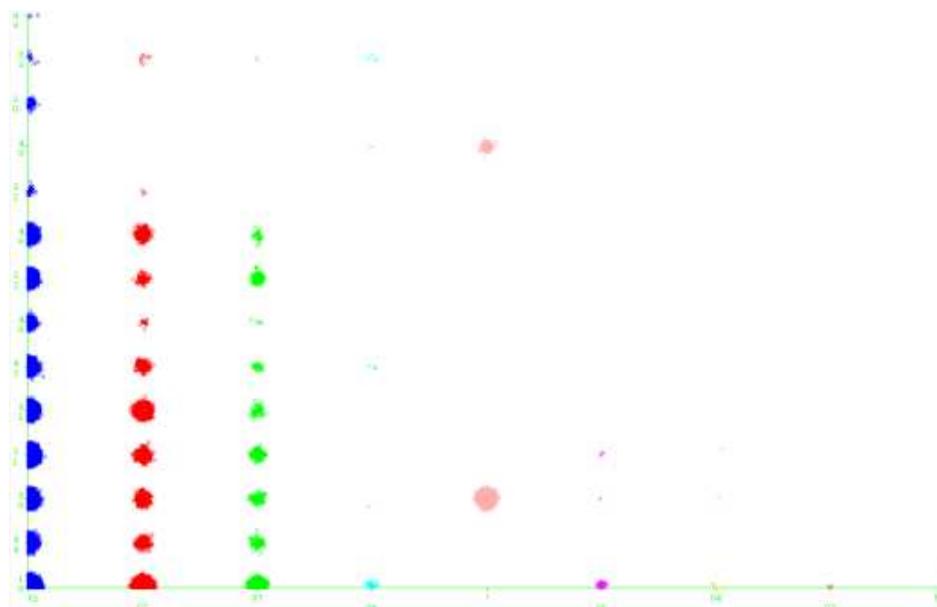


Figura 10 - Categoria do Assinante x Tipo de Acesso

Através do gráfico plotado é possível verificar que:

- Todos os telefonistas e equipamentos de manutenção estão conectados aos acessos analógicos;
- Todo equipamento de comunicação de dados estão conectados aos acessos digitais.

Os valores na legenda estão descritos na tabela **Erro! Fonte de referência não encontrada.**

## Analizando o Horário da Chamada (Eixo X)

Tipo da Chamada (Eixo Y)

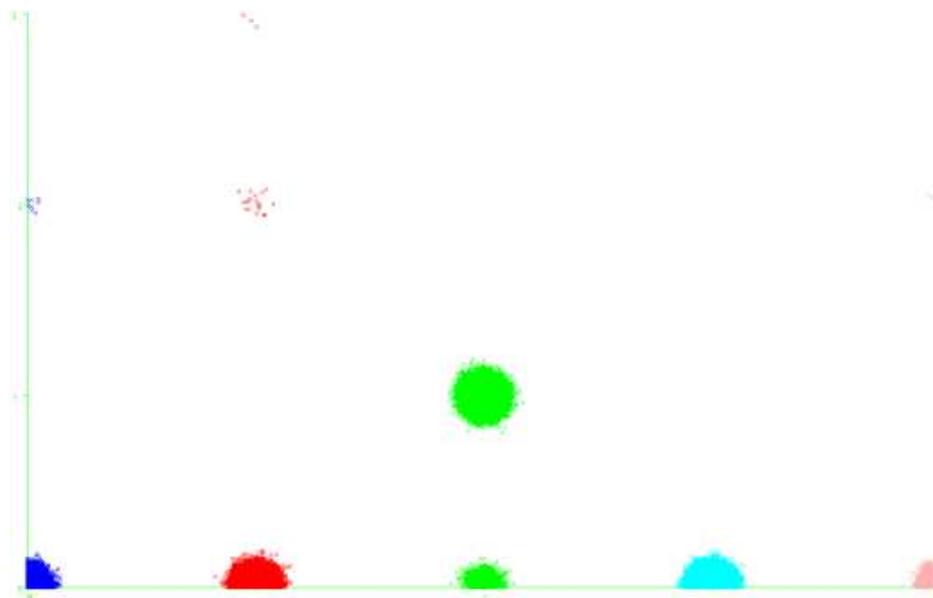


Figura 10 – Horário da Chamada x Tipo da Chamada

Através do gráfico plotado é possível verificar que:

- Chamadas realizadas em horário de tarifação diferenciado nunca são estendidas;
- Chamadas em horário normais estendem mais que no horário super reduzido.

Os valores na legenda estão descritos na tabela **Erro! Fonte de referência não encontrada.**

### Tronco de Entrada (Eixo Y)

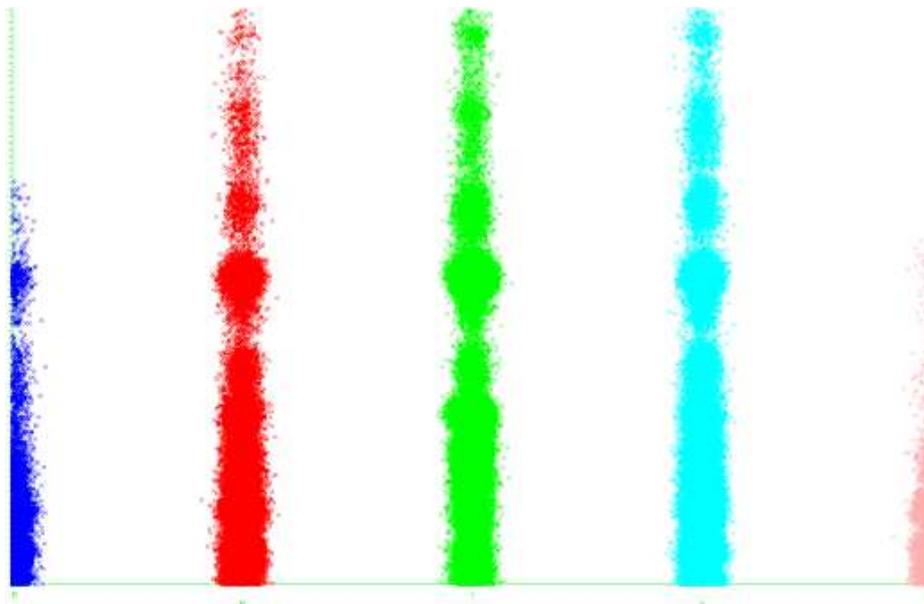


Figura 11 – Horário da Chamada x Tronco de Entrada

Através do gráfico plotado é possível verificar que:

- Chamadas realizadas em horários de tarifação Reduzidos ou Super Reduzido não permeiam todos os entroncamentos disponíveis, ou seja, existe um determinado grupo de assinantes que nunca realizam chamadas nestes horários.

Os valores na legenda estão descritos na tabela **Erro! Fonte de referência não encontrada.**

### Tronco de Saída (Eixo Y)

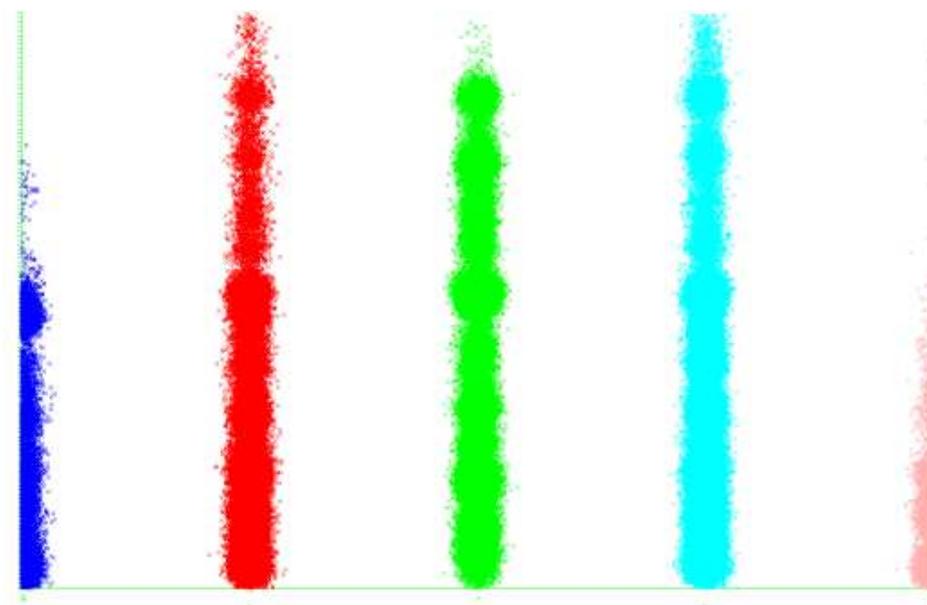


Figura 12 – Horário da Chamada x Tronco de Saída

Através do gráfico plotado é possível verificar que:

- Chamadas realizadas em horários de tarifação Reduzido não permeiam todos os troncos disponíveis, ou seja, existe um determinado grupo de assinantes que nunca são chamados neste horário ou existe alguma configuração que não permite.

Os valores na legenda estão descritos na tabela **Erro! Fonte de referência não encontrada.**

## Analizando o Tipo da Chamada (Eixo X)

Tipo de Serviço (Eixo Y)

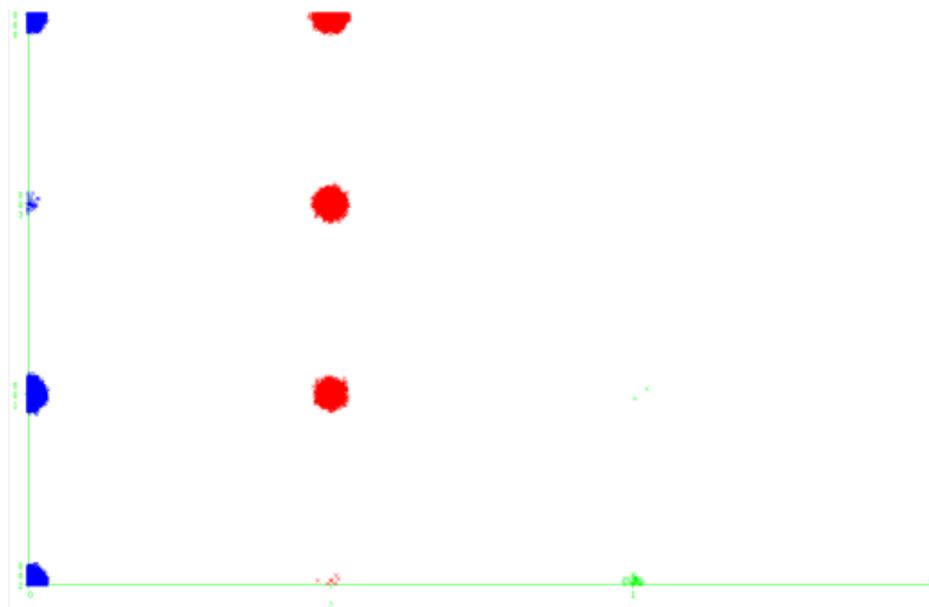


Figura 13 – Tipo da Chamada x Tipo de Serviço

Através do gráfico plotado é possível verificar que:

- Grande parte das chamadas estendidas utilizam o serviço de 64Kbps, o que mostra que grande parte destas chamadas são utilizadas para se conectar na internet;
- Contudo, pode-se perceber que algumas chamadas estendidas ainda utilizam canais simples de voz, mostrando que algumas conexões de dados estão restritas a canais de 56Kbps.

Os valores na legenda estão descritos na tabela **Erro! Fonte de referência não encontrada.**

Tipo de Acesso (Eixo Y)

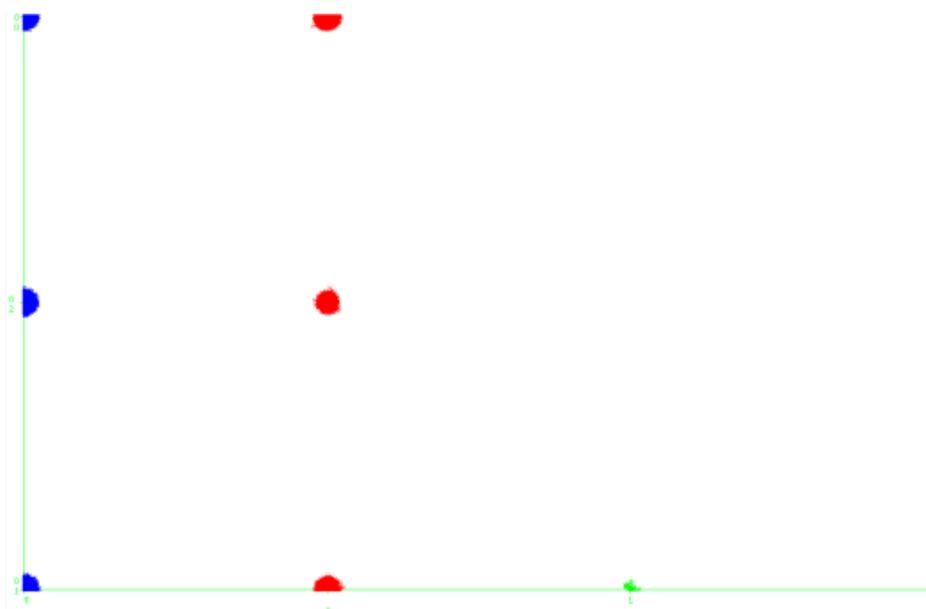


Figura 14 – Tipo da Chamada x Tipo de Acesso

Através do gráfico plotado é possível verificar que:

- Todas as chamadas estentidas são realizadas através de acessos analógicos.

Os valores na legenda estão descritos na tabela **Erro! Fonte de referência não encontrada.**

## Analizando Classe da Chamada (Eixo X)

Tronco de Entrada (Eixo Y)

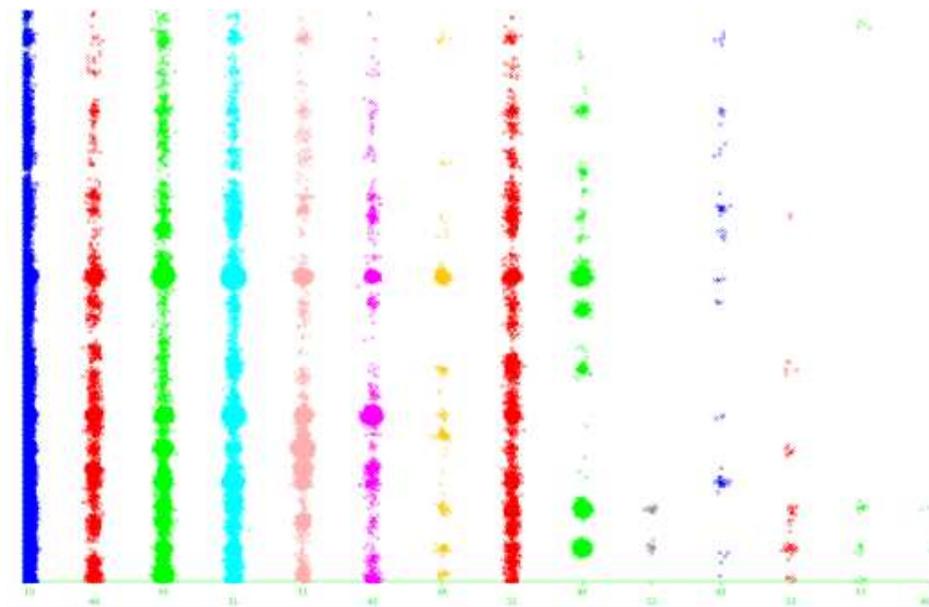


Figura 15 – Classe da Chamada x Tronco de Entrada

Através do gráfico plotado é possível verificar que:

- Chamadas não completadas por timeout no juntor de saída não permeiam todos os entroncamentos;
- Chamadas não completadas por erro B10/B15 acontecem em um conjunto restrito de entroncamentos;

Os valores na legenda estão descritos na tabela **Erro! Fonte de referência não encontrada.**

### Tronco de Saída (Eixo Y)

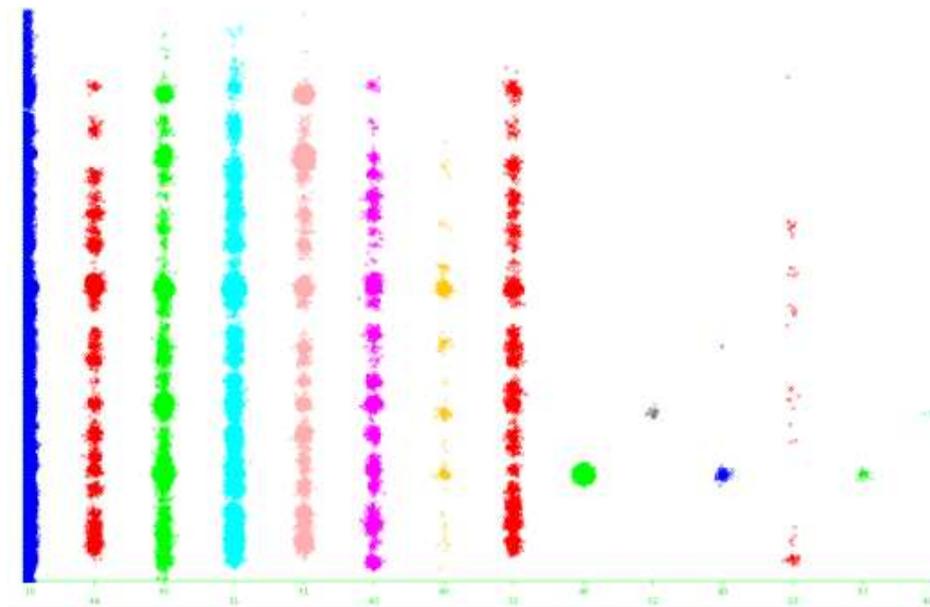


Figura 16 – Classe da Chamada x Tronco de Saída

Através do gráfico plotado é possível verificar que:

- Chamadas não completadas por timeout no junção de saída acontecem a um conjunto restrito de entroncamentos, os mesmos podem estar com algum defeito ou sobrecarga.

Os valores na legenda estão descritos na tabela **Erro! Fonte de referência não encontrada.**

Código de Falha da Chamada (Eixo Y)

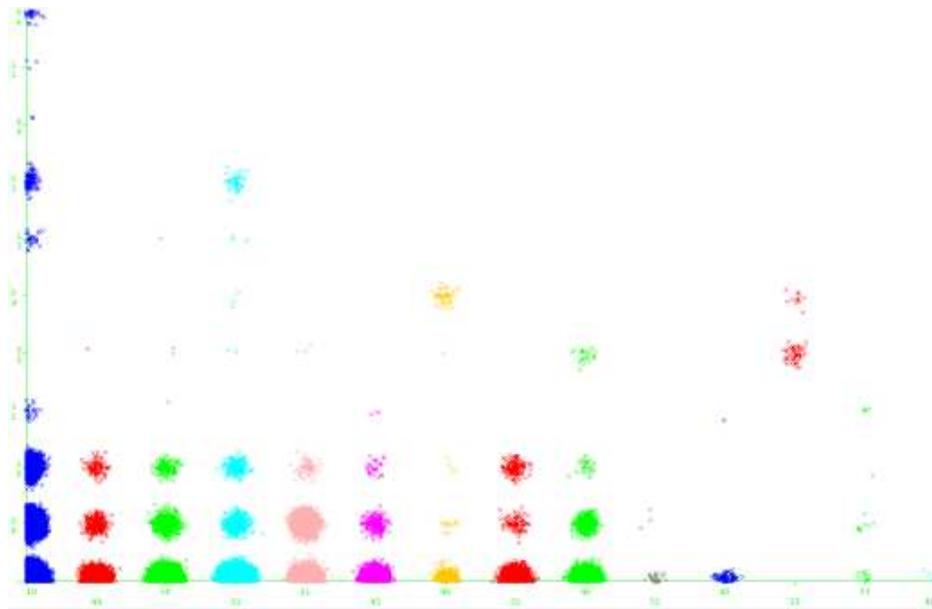


Figura 17 – Classe da Chamada x Código de Falha

Através do gráfico plotado é possível verificar que:

- Existem assinantes que discam um número antigo (sete dígitos) ao invés de utilizar o novo de oito dígitos, o que acarreta em erro da chamada por número alterado;
- Boa parte das falhas por timeout no juntor de saída se deve ao fato de atrasos por parte do assinante ao digitar o número do telefone completo.

Os valores na legenda estão descritos na tabela **Erro! Fonte de referência não encontrada.**

Tipo de Serviço (Eixo Y)

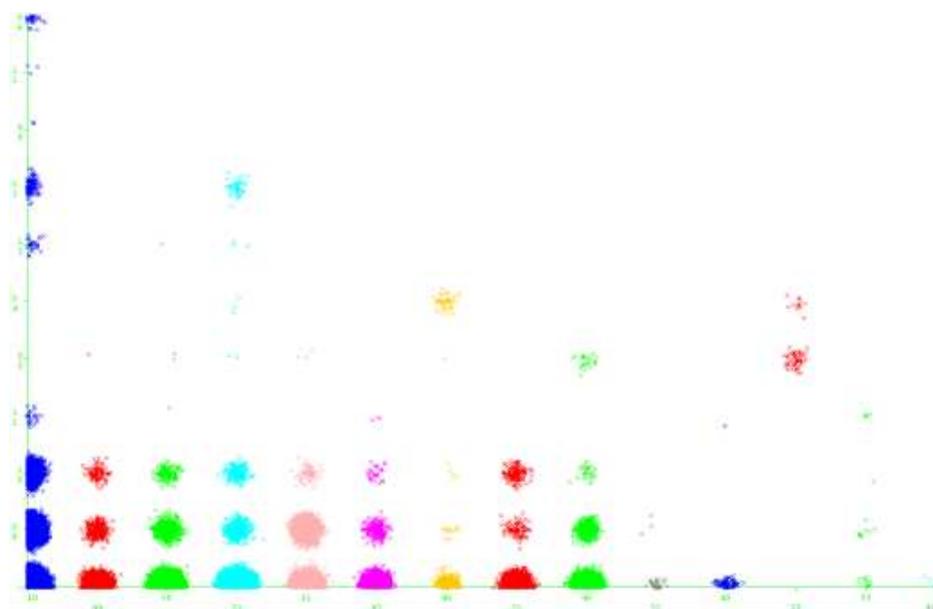


Figura 18 – Classe da Chamada x Tipo de Serviço

Através do gráfico plotado é possível verificar que:

- Chamadas que utilizam canais 64Kbps são menos afetadas por erros de não completamento, o que permite uma melhor qualidade de atendimento do assinante.

Os valores na legenda estão descritos na tabela **Erro! Fonte de referência não encontrada.**

Tipo de Acesso (Eixo Y)

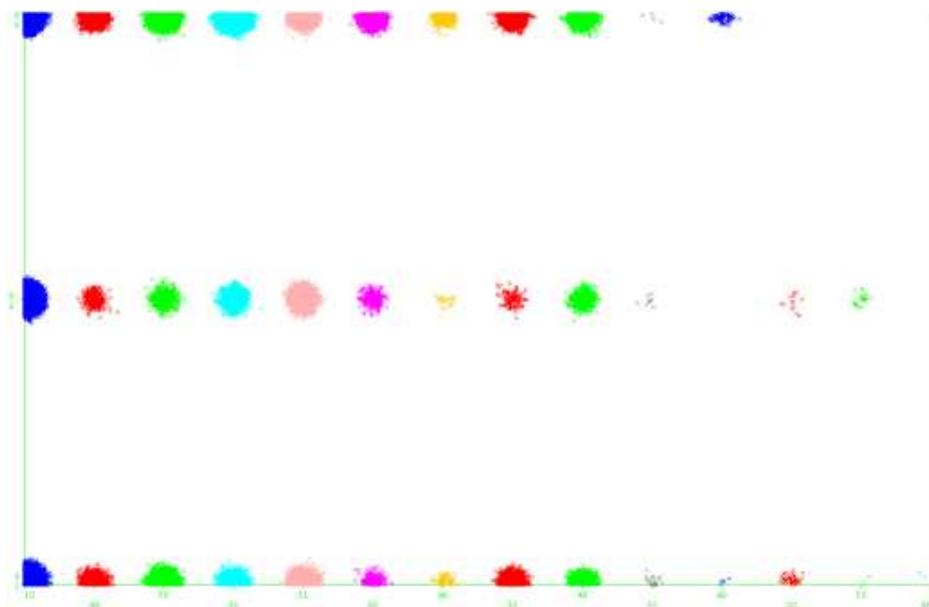


Figura 19 – Classe da Chamada x Tipo de Acesso

Através do gráfico plotado é possível verificar que:

- Chamadas realizadas por acesso digital não possuem falha de erro nos juntores de entrada nem saída, ou seja, possuem melhor qualidade de atendimento aos assinantes.

Os valores na legenda estão descritos na tabela **Erro! Fonte de referência não encontrada.**



## 7 CLUSTERIZAÇÃO

A formação de cluster através da correlação de dados é uma técnica bastante utilizada na mineração de dados; foram então realizadas duas tentativas de formação de clusters distintas com os registros de CDRs descritos anteriormente.

### Cluster I

Na primeira tentativa de clusterização foram utilizados todos os atributos, excluindo-se o “Número de A” e “Número de B”, pois é de conhecimentos que ambos os atributos possuem uma alta covariância – já que estamos manipulando bilhetes de apenas um dia de chamada, e a quantidade de assinantes que chamadas repetidas tendem a não ser muito grande considerando um único dia. O mesmo argumento é válido para a exclusão dos “Troncos de Entrada” e “Tronco de Saída”. A tabela abaixo descreve os atributos utilizados na primeira tentativa de formação de clusters:

Tabela 7 - Atributos do Cluster I

<b>Campo</b>
Tipo de Cobrança
Categoria do Assinante A
Horário da Chamada
Duração da Chamada
Tipo da Chamada
Classe da Chamada
Código de Falha da Chamada
Tipo de Serviço
Tipo do Acesso

Foi utilizado então o algoritmo Simple EM (Expectation Maximization), contudo no final chega-se a um resultado de instâncias.



## 8 MINERAÇÃO VIA CLASSIFICAÇÃO

Inicialmente os dados coletados através das chamadas CDRs não possuíam nenhuma classe, o que implica em um primeiro momento na impossibilidade de se utilizar qualquer algoritmo supervisionado de classificação. No entanto, podemos facilmente criar classes para as chamadas, basta que se selecione qualquer um dos atributos utilizando o mesmo como a classe final da instância. Óbvio que nem todos os atributos podem ser utilizados como classes, a tabela abaixo possui o atributo bem como uma breve descrição se o mesmo pode ou não ser utilizado como classe da instância:

Tabela 8 - Atributos Classificadores

<b>Campo</b>	<b>Descrição</b>
Número de A	Não será classe, pois o mesmo descreve o assinante em si.
Número de B	Não será classe, pois o mesmo descreve o assinante em si.
Tipo de Cobrança	Será classe, aponta todas as chamadas á cobrar ou normal.
Categoria do Assinante A	Será classe, aponta todas as classes de assinantes.
Horário da Chamada	Será classe, pois o mesmo aponta o tipo de tarifação da chamada.
Duração da Chamada	Não pode ser classe, atributo numérico muito extenso, conteria milhares de classes.
Tipo da Chamada	Será classe, pois o mesmo aponta o tipo da chamada.
Tronco de Entrada	Não será classe, pois descreve um tronco em si.
Tronco de Saída	Não será classe, pois descreve um tronco em si.
Classe da Chamada	Será classe, pois o mesmo aponta a classe da chamada.
Código de Falha da Chamada	Será classe, pois o mesmo aponta a classe da falha.
Tipo de Serviço	Será classe, pois o mesmo aponta a classe do serviço.
Tipo do Acesso	Será classe, pois o mesmo aponta a classe do acesso.

Para todas as classificações foi utilizado árvores de decisões (Pruned C4 trees), pois grande parte dos atributos se tratam de literais, o que é melhor manipulado por árvores. Outro ponto consiste na facilidade de visualização do resultado, permitindo depurar e até mesmo ajudar no ajuste dos parâmetros de classificação para alcançar resultados melhores (excluindo ou não os atributos que pouco afetam a separação/classificação).

## 9 RELAÇÃO ENTRE ASSINANTES

Nesta etapa foi realizada uma análise das chamadas em comuns realizadas entre os assinantes telefônicos. A idéia principal consistiu em criar um grafo direcionado, onde os vértices representam os assinantes e as arestas ligações feitas entre assinantes; tendo o grafo em mãos é possível analisar e encontrar a ocorrência de vários subgrafos. Estes subgrafos mostram conjuntos de assinantes que possuem chamadas em comum entre eles, e pode representar alguma forma de relacionamento entre os mesmos.

Como o volume de chamadas é extremamente grande foram aplicados alguns filtros com o intuito de diminuir computacionalmente os recursos necessários para a análise dos dados (alguns desses filtros não seriam muito úteis para um volume maior de dados), o objetivo é diminuir o número de assinantes (ou vértices do grafo) bem como também diminuir o número de relações (ou arestas do grafo):

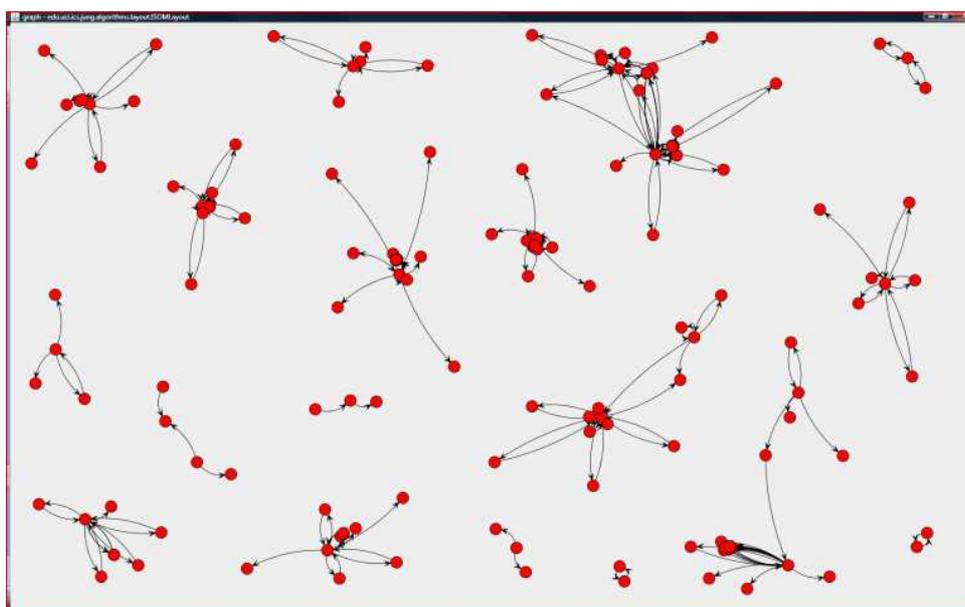
1. O primeiro filtro consistiu em fazer uma varredura por todos os CDRs e selecionar como assinante candidatos apenas aqueles cujos número de lista realizou pelo menos uma chamada telefônica e também receberam pelo menos uma única chamada telefônica. Todos os outros assinantes (sejam os que só realizaram chamadas telefônicas, ou só receberam chamadas telefônicas) simplesmente não podem fazer parte de um grupo de relação, pois isso seria impossível. Percebe-se que este filtro é pouco útil para uma base de dados reais, onde todos os assinantes recebem e realizam pelo menos uma chamada telefônica em um período (de um mês, por exemplo);
2. O Segundo filtro consiste em apenas criar uma única aresta entre dois assinantes não importante se entre eles foram realizadas uma única ou centenas de chamadas. Como nesta análise optou-se por encontrar apenas assinantes relacionados, a força da relação entre eles não traz nenhuma

importância para a análise;

3. O terceiro filtro consistiu em remover todos os assinantes (ou vértices) que não se relacionam (não recebem nem realizam chamadas) com nenhum outro assinante da lista obtida do primeiro filtro.

No final, tem-se uma lista extremamente reduzida, onde constam apenas os assinantes que realmente possuem alguma possibilidade de se relacionarem com quais outros. Apenas como exemplo, inicialmente tinha-se uma lista com cerca de 150.000 assinantes; com a aplicação do primeiro filtro esta lista foi reduzida para 528 assinantes, e após a aplicação do último filtro chegamos a um número de 223 assinantes.

Obtendo-se o grafo final (onde vértices são assinantes e chamadas entre assinantes são arestas), foi possível desenhar o mesmo, resultando na figura abaixo:



**Figura 8 - Grafo de Relação dos Assinantes**

Analisando a figura é possível apontar algumas estratégias de como utilizar as informações obtidas, os itens abaixo possuem uma descrição das mesmas:

1. Existem grupos de assinantes que realizam chamadas entre si, a partir deste grupo é possível criar promoções instigando os mesmos a

utilizarem mais o telefone para se comunicarem;

2. Existem assinantes que realizam chamadas para alguns assinantes, contudo o mesmo não recebe muitas ligações de volta destes outros assinantes; assim é possível realizar promoções que facilite o assinante que não retornam chamadas a fazê-las;
3. Existem assinantes que recebem várias ligações de outros assinantes, pode-se oferecer a este assinante um serviço do tipo “chamada em espera”, “correio de voz” ou até mesmo algum plano onde o mesmo recebe mais bonus pela quantidade de chamadas atendidas;
4. Existem assinantes que se comportam como centro de atendimento de chamadas, pode então ser oferecido a estes assinantes serviços de chamdas 0800 (para o caso de empresas) ou então um serviço de PBX ou Contact Center;
5. Existem conjuntos de assinantes que estão conectados de alguma forma em subgrafos, além disso, estes subgrafos não possuem todos os nós conectados entre si; o que permite sugerir uma tentativa de conexão entre esses assinantes (totalmente ou parcialmente) seja através de um promoção ou informando o mesmo sobre um número de lista que está presente no círculo de seus amigos e que, no entanto o mesmo não possui o contato.

Lembrando que tendo um grafo direcionado montado das chamadas telefônicas, é possível utilizar qualquer algoritmo de mineração de dados ou da teoria dos grafos para que se encontre qualquer realação existente no mesmo. Este artigo se limitou em uma análise visual do grafo, apenas com o intuito de demonstrar e provar que a partir dos grafos das chamadas telefônicas é possíveis obter diversos resultados bem como características dos assinantes; contudo isso é praticamente nada se comparado com todas as informações que podem ser realmente extraídas.



## 10 CONCLUSÃO

A análise de dados das chamadas telefônicas é muito pouco ou quase nada explorada nos dias atuais. Grande parte dos trabalhos voltados para a área consiste em acompanhamentos simplificados, como por exemplo:

- Acompanhamento de volume de chamadas
- Segmentação dos clientes
- Criação de modelo de vendas por segmentos

As análises mencionadas anteriormente utilizam-se muito pouco ou nada das técnicas e modelos de mineração de dados estudados e desenvolvidos. Grandes partes dessas análises consistem em análise estatística de comportamento com base em número de chamadas por período de tempo, ou na melhor das hipóteses técnicas de clustering também baseada na duração das chamadas realizadas pelos assinantes.

Existe um estudo mais elaborado, onde foram analisados registros de clientes de operadoras utilizando regressão ordinária [1], neste estudo os pesquisadores conseguem mostrar que é possível detectar padrões de rotatividade de clientes das operadoras de telefonia móvel, e que portanto, é possível tomar medidas para que se diminua a taxa de clientes que cancelam seu contrato com uma operadora para aventurar-se em outra. No entanto o estudo não se utilizou dos CDRs, e sim de dados pré-processados e classificados pelas operadoras como clientes que cancelaram seu contrato ou não.

Enfim, grande parte da dificuldade em se estabelecer e analisar os CDRs consiste no caráter sigiloso da informação do mesmo (já que em praticamente todos os países no mundo os acessos a chamadas telefônicas são confidenciais por lei, e só podem ser acessados mediante requisição judicial), seguido pelo grande volume de dados obtidos, e que requer altíssimo custo computacional para a análise. Um ponto extremamente importante na análise dos CDRs consiste em compilar os atributos e dados obtidos para que

seja possível gerar um novo conjunto de dados contendo apenas os atributos específicos para uma análise em específico; a idéia de manter todos os atributos dos CDRs para que seja possível realizar uma mineração de dados mais ampla na esperança de que possa ser encontradas um maior número de relação é fantasioso e incompatível com os recursos computacionais disponíveis atualmente.

É perceptível também que a restrição de acesso aos dados, limita em muito a poder que uma comunidade de pesquisadores pode despende em cima do assunto para que sejam criadas ferramentas poderosas, e por fim que alcance resultados que venham a agradar qualquer Diretor da área de CRM ou Relação com os Clientes de uma empresa prestadora de serviço de telefonia, estamos aí em um empasse (entre necessidade de mercado e restrição judicial) que pode ser facilmente contornado tomando-se como exemplo o estudo realizado pelos pesquisadores Rupesh K. Gopal e Saroj K. Meher [1], onde foram disponibilizadas pelas empresas prestadoras de serviço de telefonia móvel não os dados detalhados de cada chamada mais um conjunto de dados compilados que permita de alguma forma a mineração de algumas informações.

Independente da utilização ou não dos dados de CDRs para análise de comportamento dos assinantes estarem sendo utilizadas ou não, é fato que esta base de dados consiste em uma das maiores fontes de informação disponibilizadas atualmente - o sistema de telefonia é ainda hoje o principal meio de comunicação entre as pessoas (seja para fins familiares, negócio, governamental, amigos, etc). Sistema esse que apesar de estar migrando de forma acelerada para o uso da rede de dados da Internet, não vai perder sua principal característica – que consiste na conexão entre duas ou mais pessoas para a troca de informação/mídia (voz, vídeo, fotos, etc). Esta convergência da telefonia para a rede de internet deve trazer ainda mais assinantes para a comunidade, o que com certeza vai aumentar ainda mais a geração de dados de registro destas chamadas e permitir uma análise ainda maior.

Outro ponto de destaque, consiste no uso restrito dessa massa de dados

apenas as operadoras de serviço de telefonia, tais dados podem ser utilizados por diversas outras instituições que queiram explorar de forma mais apropriada a carteira de clientes existente (através de um melhor conhecimento de seu perfil) e até mesmo por órgãos governamentais responsáveis por regular e monitorar a qualidade do serviço de telefonia nos países, uma ferramenta com tal objetivo poderia se tornar facilmente um software de medição de qualidade no serviço bem como um indicador de que cláusulas contratuais aceitas pelas operadoras (e propostas) pelo governo estão sendo respeitadas e atendidas conforme previamente acordado.

Ressaltando que para os testes realizados neste estudo de caso foram utilizados CDRs referentes a um único dia de chamada de uma única central, o que implica num baixo volume de relações que são realmente possíveis, caso seja possível conseguir uma base de dados mais volumosa englobando mais dias fica evidente que o número de relações tendem a aumentar (ainda mais quando consideramos chamadas em comum entre assinantes).



## 11 PRÓXIMOS PASSOS

Existe uma enorme gama de análise que podem ser feitas nos dados de CDRs, de uma forma geral pode-se criar aplicações baseadas nas seguintes perspectivas:

- Operadora de Serviço (CRM: segmentação, promoções, marketing direcionado).
- Operadora de Serviço (Q&A: comportamento da rede, definição de gargalos, pontos de melhoria, expansão)
- Operadora de Serviço (Segurança: detecção de fraudes no uso da rede)
- Governo (Q&A: verificação de qualidade do serviço)

Qualquer que seja a perspectiva é necessário definir bem o escopo do tipo de análise a ser realizada, pois tendo o escopo bem definido é possível reduzir em muito o grande volume de dados a ser manipulado. Uma segunda opção, não mais simples, consiste em realizar uma análise generalizada e para que se atenda computacionalmente os recursos pode-se atacar o problema utilizando-se de algoritmos de reconhecimento de padrões que possam ser paralelizados e que permitam diluir o custo computacional em um cluster de computadores.

O uso de CDRs focando o serviço para governos ou grandes operadoras de serviço é uma das melhores opções – devido ao grande volume de problemas e benefícios que podem ser garimpados. No entanto, este não é o único mercado que pode ser beneficiado de uma análise de comportamento de assinantes através de CDRs; existem hoje, milhares de centrais telefônicas espalhadas pelo mundo, essas centrais não são utilizadas para prover serviço público de telefonia, mas sim para prover serviço privado para uma empresa ou instituição qualquer. Essas centrais de menor porte e robustez suportam um serviço de telefonia centralizado e são conhecidas como PBX (Private Branch Exchange). Os PBXs geram um volume de CDRs muito menor quando comparado as centrais de telefonia pública concessionadas nacionalmente (é

claro que o volume varia de acordo com o número de funcionários da empresa, mas mesmo uma empresa muito grande não chega perto de atingir a quantidade de habitantes de uma cidade de pequeno porte). Um menor volume de CDRs significa uma redução exponencial do custo computacional para a análise dos CDRs, assim é possível analisar comportamento dos funcionários quanto ao uso do telefone bem como relação da empresa com seus clientes e fornecedores.

Uma outra oportunidade a ser explorada consiste na análise de CDRs gerados em sistema de “contact center” (onde são realizadas vendas, marketing e atendimento de reclamação dos clientes); é possível utilizar os CDRs destes sistemas para que sejam encontrados padrões de comportamento tanto dos clientes que se relacionam com a empresa bem como dos funcionários responsáveis em atender com qualidade os clientes.

Pode-se também combinar os dados de CDRs com outros atributos dos assinantes de telefonia que não estejam diretamente atrelado as chamadas (ou uso de mídia) realizadas. Por exemplo, pode-se adicionar dados como serviços contratados (BINA, transferência de chamada, etc), quantas vezes é colocado crédito no celular por mês (ou valor pago pela conta), etc.

Caso não seja possível realizar uma análise mais complexa dos padrões, pode-se utilizar algoritmos mais simples, bem como o uso da estatística para acompanhamento personalizado de cada assinante (ou funcionário, cliente, fornecedor, etc.) e assim acompanhar a frequência de uso do serviço (ou contato e força do relacionamento). Assim é possível enviar notificações ou alertar quando um assinante está diminuindo em muito o uso da rede (o que pode significar uma mudança de operadora ou redução do custo por algum motivo) ou então até mesmo alertar quando um assinante passa a utilizar a rede em demasia, extrapolando seu comportamento natural (o que pode significar fraude ou clonagem do aparelho).

Finalmente, fica a sugestão para que se utilize os CDRs com a finalidade de criar um engine de classificação e priorização de telefones, assim como o google é capaz de apontar sites, fotos, vídeos e priorizar esses itens utilizando

algoritmos específicos; porque não utilizar a rede de telefonia como uma “página amarela” informatizada. Pode-se construir uma rede onde o volume de chamadas possa ser utilizado como “score/rating” de um estabelecimento comercial, assim fica mais fácil indicar a alguém realizando uma busca qual o melhor estabelecimento seja para pedir uma pizza ou alugar um buffet para que seja realizada a festa de aniversário dos filhos.

Não se deve esquecer também do mais importante, hoje estamos vivendo uma revolução tecnológica onde a rede de telefonia e a rede de dados (Internet) estão se convergindo e cada vez mais se tornando uma única grande rede. Isso significa que em poucos anos, sendo assim, as informações geradas pelos CDRs e também pelo acesso às páginas Web vão possuir uma maior interligação entre si, e uma extensa gama de novas análises de padrões poderá ser realizada, visualizando assim e serviços a serem oferecidos nunca imaginados antes.



## REFERÊNCIAS BIBLIOGRÁFICAS

GOPAL, R. K.; MEHER, S. K. Customer churn time prediction in mobile telecommunication industry using ordinal regression. In: ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING, PACIFIC-ASIA CONFERENCE, PAKDD 2008, 12., 2008, May 20-23, Osaka, Japan. **Proceedings...** Osaka: Springer, 2008. v. 5012, p 884-889. DOI: 10.1007/978-3-540-68125-0\_88. 2008. Lecture Notes in Computer Science.

