Cutting without Pain: Mitigating 3G Radio Tail Effect on Smartphones

Fei Yu†, Guangtao Xue†, Hongzi Zhu†, Zhenxian Hu†, Minglu Li†, Gong Zhang‡
†Shanghai Jiao Tong University, China
‡Huawei Research, China
†{fishfly_1008,gt_xue,hongzi,huzhenxian,mlli}@sjtu.edu.cn; ‡nicholas.zhang@huawei.com

***Abstract*—3G technology has stimulated a wide variety of high-bandwidth applications on smartphones, such as video streaming and content-rich web browsing. Although having those applications mobile is quite appealing, high data rate transmission also poses huge demand for power. It has been revealed that the tail effect in 3G radio operation results in significant energy drain on smartphones. Recent fast dormancy technique can be utilized to remove tails but, without care, can degrades user experience. In this paper, we propose a novel scheme *SmartCut*, which effectively mitigates the tail effect of radio usage in 3G networks with little side-effect on user experience. The core idea of SmartCut is to utilize the temporal correlation of packet arrivals to predict upcoming data, based on which unnecessary high-power-state tails of radio are cut out leveraging the Fast Dormancy mechanism. Extensive trace-driven simulation results demonstrate the efficacy of SmartCut design. On average, SmartCut can save up to 56.57% energy on average while having little side-effect to user experience.**

I. Introduction

Smartphones equipped with powerful CPUs and 3G wireless communication modules have gained a large popularity nowadays, which stimulates a large spectrum of high-bandwidth applications such as video streaming, online games and content-rich web browsing booming on smartphones. As the capability of smartphones keeps soaring, battery technology remains a bottleneck. Recently, the tail effect of 3G interface has gained much attention, which refers to the interface will keep at high-power states for a long time even after the completion of data transmission. Studies[1][2] have reported that the tail effect of 3G interface can consume up to 60% of the total energy.

To eliminate the tail effect in 3G networks, however, is very challenging due to two reasons. First, the tails are designed to achieve fast response to upcoming data transmissions. Simply cutting the tails without care would cause non-negligible delays since the interface needs time to switch from low-power (idle) states to high-power (data-transmission-ready) states each time when the interface is not ready for the required data transmission. Long promotion delay would seriously degrade user experience. Second, in order to determine energy-efficient tails, it is inevitable to know the precise data work load in the future. In practise, this is very hard, if not impossible.

In this paper, we first collect a real trace of 3G network traffic from a metropolis in China, involving more than 65,000 users over a month. By analyzing the empirical data, we confirm that the tail effect contributes more than 60% to the total power consumption used for data communication. Furthermore, we examine the packet arrival time from traffic generated by three categories of popular applications, i.e., video streaming, web browsing and instant messaging and find that 3G traffic aroused by all those applications shows strong temporal correlation.

Inspired by this observation, we propose an innovative scheme, called SmartCut, which integrates two key techniques to mitigate the tail effect of 3G networks while keeps most applications running on smartphones un-affected. The core idea of SmartCut is to train an autoregressive move average (ARMA) model using historical 3G traffic trace, based on which SmartCut consecutively predicts the arrival time of upcoming packets. With the estimated packet arrival time, SmartCut first adopts the fast dormancy mechanism to cut the unnecessary high-power-state tails off. Moreover, in order to reduce the side-effect on user experience, it also promotes the 3G interface in advance before the next packet arrives. Trace-driven simulations show that SmartCut can save 56.57% energy on average among different applications.

The remainder of the paper is organized as follows. Section 2 presents the related work. Section 3 introduces the system model. In Section 4, we describe the collected 3G data trace and then analyze the impact of the tail effect. The SmartCut scheme design is presented in Section 5. In Section 6, we describe the methodology to evaluate the performance of SmartCut and present the result. Finally, we give a conclusion of our work in Section 7.

II. Related Work

The tail effect, where the cellular radio remains in a high power state for a long duration, contributes a large fraction of total radio energy consumption[2]. How to decrease the tail effect is arousing significant attention. In the literature, several schemes have been proposed to reduce the impact of tail effect, trading off between minimal energy waste and good user experience. Basically, those schemes fall into two categories: *tail-cutting* and *tail-sharing*. Given the knowledge of future packets, tail-cutting schemes[3] remove the high-power-state tail by adopting the fast dormancy mechanism, under which a smartphone can ask for an immediate release of radio links and leave the high-power-states rapidly. Knowing

978-1-4673-5946-7/13/$31.00 ©2013 IEEE

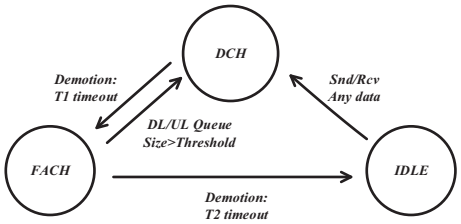


Fig. 1. state transition graph

the information of future packets, however, is infeasible in practice. In contrast, tail-sharing schemes[1][4] re-schedule packets from different up-layer applications and pack packets together in order to share tails. Bartendr[5] further predicts an appropriate time to transfer packed data under better signal strength. These schemes can cause a great deal of packets delayed and therefore are effective only when the applications running on a smartphone are delay tolerant but fail for real-time or time sensitive applications.

III. System Model

In 3G networks, the establishment and release of a radio link between the interface of user equipment (UE) and the base station (BS) are controlled by the Radio Resource Control (RRC) protocol. In RRC, a state machine is established to manage the status of a physical radio link, which consists of three states, i.e., IDLE, FACH and DCH. The RRC state transition graph is shown in Fig. 1.

IDLE: If there is no data to transfer for a sufficiently long time, the interface will switch to the IDLE state. At this state, there is no radio link established between the device and the base station. Therefore, the interface is unable to transfer any data and does not consume any energy either.

DCH: When the interface requires to transmit data, it will switch to the DCH state. At this state, the interface is allocated with a dedicated transport channel. It can make full use of the radio link for high speed transmission but the power consumption at this state is also the highest, e.g., about 800mW.

FACH: When the interface completes a data transmission at the DCH state, it will switch to the FACH state. At this state, the interface is allocated a shared transport channel. The power consumption at this state is about 400mW.

There are two kinds of transitions among the three states, namely, *promotion* and *demotion*.

Promotion: It refers to that the interface switches from a state to another state with higher power consumption, such as from IDLE to DCH or from FACH to DCH. The promotion transition requests for more radio resource from the base station, which will introduce large delays and additional energy consumption.

Demotion: It refers to that the interface switches from a state to another state with lower power consumption. Demotions between states are controlled by *inactivity timers*, namely, T_1 and T_2. First, when the interface completes a data transmission at DCH, the T_1 timer starts. If T_1 expires before any data transfer request, the interface will be demoted to the FACH state. After that, the T_2 timer starts and if it expires, the interface will drop to the IDLE state. Note that during the procedure, any data transmission request will reset both timers. The two timeouts are known as *tail time*.

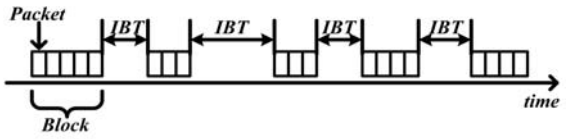


Fig. 2. An example of data blocks and inter-block times.

Tail time is designed in 3G networks so that instant packets can be transferred immediately. On one hand, long tail time can provide good user experience of applications since no packets will experience the promotion delay. On the other hand, long tail time will also result in a great waste of energy if there are no packets to send or receive. Therefore, the tail-time selection strategy is the trade-off between good performance and energy efficiency. An optimal strategy should adaptively meet the data transfer requirements of applications.

IV. Empirical 3G Data Analysis

A. Collecting 3G Trace

We collect 3G traffic trace data from a southern metropolis in China, which contain more than 65,000 3G users for one month from Nov. 25 to Dec. 24, 2011. A record in the trace contains the following information of a packet: User ID, timestamp, packet size, source and destination IP addresses, transportation-layer port, etc. Furthermore, we infer the corresponding application generating the packet using a deep packet inspection algorithm. We choose three categories of applications, i.e., web browsing, streaming and instant messaging to study since they contribute the majority of the traffic.

B. Revealing the Impact of Tail Effect

In order to reveal the impact of tail effect in real scenarios, we apply the above RRC model we have learnt to our trace data.

We first arrange packets of each application in a row according to the time when the packet was sent or received. Particularly, we notice that packets are coming in burst in our trace. Since the tail between two immediate packets are pretty short and can be omitted, we refer to a block as a series of consecutive data packets which have small inter-packet time (less than 500 milliseconds in our analysis) and take blocks into consideration instead of individual packets. Accordingly, we define the inter-block time (IBT) as the time interval between any two consecutive blocks of packets (as illustrated in Fig. 2).

Given the RRC model, an IBT larger than the sum of two inactivity timers (i.e., T_1 and T_2) will cause the link between the UE and the BS to be released. In this case, when the next packet block arrives, the UE has to go through a promotion transition, introducing a promotion delay before the real data transfer happens. Besides the promotion delay, waiting for T_1 and T_2 to expire also waste much energy in vain. In addition, as mentioned in Section 3, promotion transition also

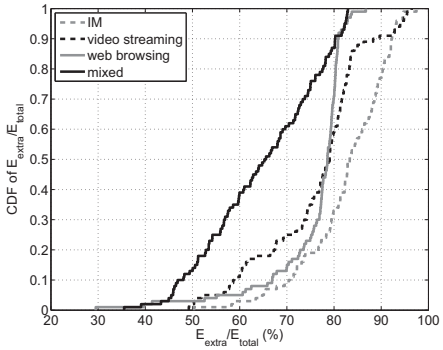


Fig. 3. Ratios of extra energy against total energy among 1,000 users selected randomly from our dataset.

arouses extra power consumption. We consider both the power consumption caused by tails and that of promotions as extra energy cost, in contrast with effective energy cost which is used for the actual data transmission. We plot the cumulative distribution function (CDF) of the ratio of extra energy cost to the total energy consumption over 1,000 randomly-selected users in Fig. 3. It is clear to see that, in general, most of the energy is extra energy cost. For example, over 80% users have more than 70% energy on average spent on tails. It can also be seen that mixed traffic has lighter tail effect than separated ones. The reason is that when several applications access the network simultaneously, the packet arrivals become more intensive, which implies shorter IBTs and duration of tail times. Therefore, the extra energy under mixed traffic is smaller than others.

C. Characterizing Temporal Correlation of 3G data

From the above analysis, the characteristics of the traffic load, especially the layout of packets, plays an important role in determining the amount of power consumed over 3G networks. To understand whether there exist patterns in 3G traffic load, we examine the temporal correlation between IBTs by calculating the marginal and conditional entropy in this subsection.

Let $\{t_i, i = 1..n-1\}$ represent the IBT series where t_i denotes the i^{th} IBT and t_i has m different observed values $\{v_j, 1 \leqslant j \leqslant m\}$ throughout the series. Let x_j be the times of v_j appearing in the series. Then, the probability of $P_{v_j} = x_j/(n-1)$. Therefore, the marginal entropy of the series $\{t_i\}$ can be written as:

$$H(t_i) = -\sum_{1 \leqslant j \leqslant m} P_{v_j} * log_2 P_{v_j} \tag{1}$$

Now, we calculate the conditional entropy of IBTs given their previous M values.

When $M = 1$, let $\{(t_{i-1}, t_i), i = 2..n-1\}$ be the two-dimensional random variable for inter-block time t_i and its immediate predecessor. Suppose that the value space of (t_{i-1}, t_i) is Q. Therefore, the joint entropy of (t_{i-1}, t_i) is:

$$H(t_{i-1}, t_i) = -\sum_{(u,v) \in Q} P_{(u,v)} * log_2 P_{(u,v)} \tag{2}$$

where $P_{(u,v)}$ is the probability of inter-block time t_i being v and its immediate predecessor being u. With $H(t_{i-1}, t_i)$ and $H(t_i)$, the conditional entropy of series $\{t_i\}$ given $\{t_{i-1}\}$ is:

$$H(t_i|t_{i-1}) = H(t_{i-1}, t_i) - H(t_{i-1}) = H(t_{i-1}, t_i) - H(t_i) \tag{3}$$

Similarly, when $M = 2$, the conditional entropy of series $\{t_i\}$ given $\{t_{i-1}\}$ and $\{t_{i-2}\}$ can be written as:

$$H(t_i|t_{i-1}t_{i-2}) = H(t_{i-2}, t_{i-1}, t_i) - H(t_{i-2}, t_{i-1}) \tag{4}$$

Fig. 4 shows the CDFs of the mean marginal entropy and the mean conditional entropy, for $M = 1$ and 2, over 1,000 users in 3G trace. From the CDFs, it can be seen that, in all the application scenarios, the more previous IBTs being given, the smaller entropy we will get. It implies that the uncertainty about IBTs decreases when knowing historical IBTs.

V. SmartCut Design

A. Overview

With the observation of temporal correlation of IBTs, it is possible to infer future traffic, which can be utilized to establish a more energy-efficient RRC state transition strategy. To this end, we design the SmartCut scheme. There are two key techniques integrated in the scheme: *estimating future IBTs* and *cutting unnecessary tails*. Specifically, SmartCut adopts the ARMA model to capture the temporal correlation of recent IBTs, based on which it estimates the next IBT. With such information, SmartCut actively cuts the tails if the expected IBT is larger than the promotion delay and promotes the interface in advance before the real data transmission begins.

B. Estimating Future IBTs

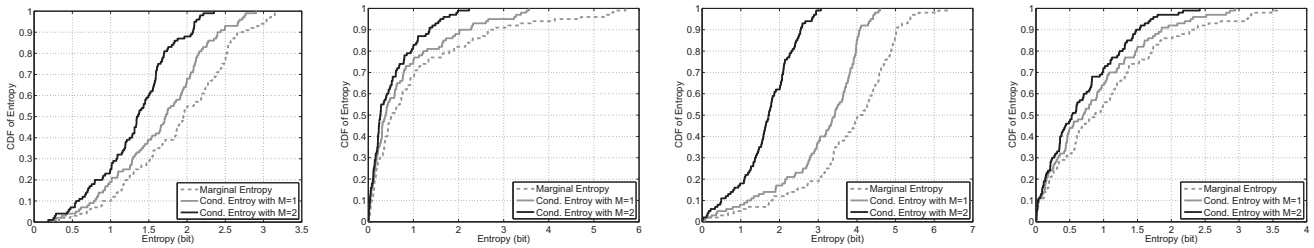
Treating the series of IBTs as a signal in the temporal dimension, we can apply time series analysis models to capture temporal correlation of IBTs and further predict future IBTs. We choose to use ARMA model[6] because it is rich enough to capture a large variety of temporal dependencies. The model consists of two parts: an autoregressive part (AR) and a moving average part (MA). Typically, the model is referred as a two-tuple ARMA(p,q), where p is the order of the autoregressive part and q is the order of the moving average part. It can be defined as follows:

$$\begin{aligned} x_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} \\ + \varepsilon_t - \theta_1 \varepsilon_{t-1} - ... - \theta_q \varepsilon_{t-q} \end{aligned} \tag{5}$$

where $\phi_1, ..., \phi_p, \theta_1, ..., \theta_q$ are the parameters of the model, ϕ_0 is a constant component and $\varepsilon_t, ..., \varepsilon_{t-q}$ are white noise error terms.

The orders p and q can be determined by conducting Akaike information criterion (AIC) test[7], which is a measure of the relative goodness of fit of a statistical model. Using least-square estimation, the AIC can be written as:

$$AIC = nlog\sigma^2 + (p + q + 1)logn \tag{6}$$



(a) Web browsing (b) Video streaming (c) IM (d) Mixed

Fig. 4. CDFs of marginal entropy and conditional entropy of inter-block times in 3G dataset

where n is the number of samples, σ^2 is the variance of the residuals after fitting the model, p and q are undetermined orders of the model. The AIC test will choose the values of p and q which maximize the equation(6).

According to the previous discussion, since different applications have different data traffic pattern, SmartCut maintains a unique model for each one for the purpose of prediction accuracy. After training, future values of IBTs can be calculated using equation (5). With this information, SmartCut is able to make a better decision on which state the 3G interface should stay. Thereby giving increase to energy efficiency with little side-effect on user experience.

Fig. 5 shows an example of IBT prediction. The grey curve represents an IBT series of a randomly selected user from our trace and the black curve represents its estimated value. From the graph we can see that SmartCut achieves good accuracy of prediction when using ARMA model.

C. Cutting Unnecessary Tails

With the established ARMA models, SmartCut is enabled to predict future data blocks. Based on the estimated IBT values, SmartCut will immediately switch the 3G interface to the IDLE state after a data transmission completes if the expected IBT is larger than the promotion delay. In order for the interface to function properly in future data transmissions, it is also necessary to promote the interface back in advance. For this reason, it promotes the interface back to the FACH state before the estimated arrival time of the next data block. In this way, with an accurate prediction, the promotion can be completed just before the actual data transmission so that users cannot feel any promotion delays and the energy wasted by unnecessary tails are saved.

More specifically, let $\{t_i, i = 1..n-1\}$ be the series of inter-block times and $\{t'_i, i = 1..n-1\}$ be the predicted values of $\{t_i\}$. We assume that the promotion delay is t_{delay}. Then the *cut-and-promote* scheme is described as follows:

- After a new booting of a smartphone, the 3G interface remains at the IDLE state.

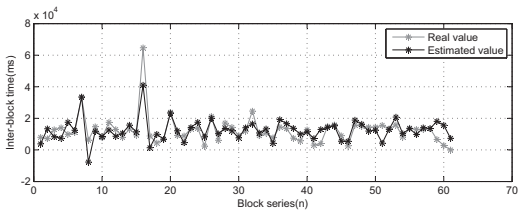


Fig. 5. An example of IBT prediction

- Once a data transmission completes, the interface forecasts the arrival moment of the next data block. Let t'_i be the predicted value of the inter-block time between the current block and the next one. if $t'_i \leq t_{delay}$, the tail gets retained. Otherwise, an immediate demotion from DCH to IDLE will be applied to the interface.
- When at the IDLE state, the 3G interface will be promoted to the FACH state after a time of $t'_i - t_{delay}$, preparing for the upcoming data.

VI. Evaluation

A. Methodology

In this section, we present the trace-driven simulations. We randomly choose a one-month 3G data set including 1,000 users, which contains of data traffic generated by all 3G applications. In realizing the significance of video streaming, web browsing and instant messaging in the traffic distribution, we focus on their traffic to study. We compare SmartCut with several alternative schemes:

1) **Always-on**. In this scheme, the 3G interface remains awake for ever no matter whether there is data to transfer or not.

2) **Always-off**. As an opposite of the Always-on scheme, it always demotes the interface to the IDLE state once a transmission completes.

3) **Fixed-tail**. This scheme is adopted by network operators by default, in which, after each transmission, the 3G interface retains a fixed tail (five seconds for T_1 and 11 seconds for T_2 as we measured).

We evaluate all above radio usage schemes using the following metrics:

1) R_e*:* It is defined as the ratio of the extra energy to the total energy. The larger the ratio is, the less energy-efficient the scheme is.

2) R_p*:* It is defined as the ratio of the number of un-delayed packets to the total number of packets. We use this metric to quantify user experience with certain radio usage scheme.

3) η_e*:* It is defined as the number of un-delayed packets divides the total energy consumed. To thoroughly assess whether a scheme is good, only considering the amount of energy saved is not sufficient. With this metric, we can perfectly depict the trade-off between energy consumption and user experience.

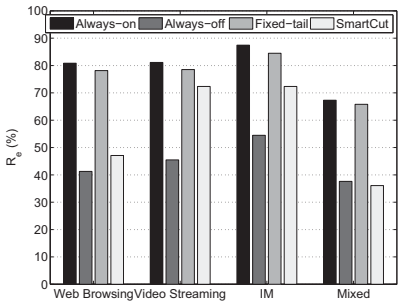


Fig. 6. Comparison of all schemes using metric R_e

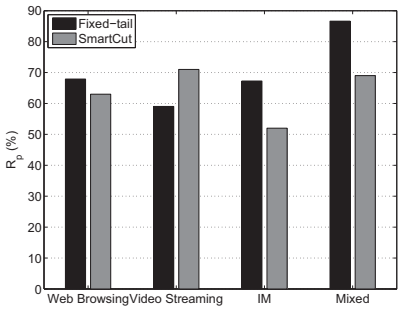


Fig. 7. Comparison of all schemes using metric R_p

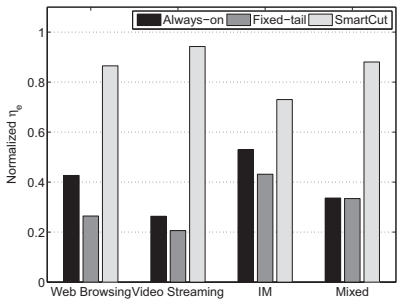


Fig. 8. Comparison of all schemes using metric η_e

B. Comparing with Alternative Schemes

In this subsection, we compare SmartCut against other alternative schemes over all users in the trace.

Fig. 6 plots the R_e for all the schemes averaged over 1,000 users. It can be seen that in all application scenarios, Always-on obtains a highest extra energy ratio and in contrast, Always-off has the lowest. It can also be seen that for Always-off, even with all the tails being cut off, the extra energy remains high (up to 50% against the total). It is because that lots of promotions are brought in, which also leads to considerable energy consumption. Note that SmartCut always achieves a more efficient use of energy than Fixed-tail. Table I shows that SmartCut can save up to 56.57% energy for applications compared to the Fixed-tail scheme.

Fig. 7 plots the R_p for all the schemes averaged over 1,000 users. We skip the results of Always-on and Always-off in the figure. The reason is that the former always has a 100% R_p with no packets being delayed and the later always has a zero R_p with all packets being delayed. It can be seen that, in most cases, SmartCut has more delayed packets than Fixed-tail, affecting user experience, which is the cost of improvement of energy efficiency.

Fig. 8 plots the η_e for all the schemes averaged over 1,000 users. Just as in Fig. 7, we skip the result of Always-off for the reason it always being zero. We can see that among all the schemes, SmartCut achieves the highest value of η_e. Especially in video streaming, the η_e of SmartCut is five times as much as that of Fixed-tail. It is implied that SmartCut makes a much more efficient use of energy to improve the user experience.

The simulation results show that in most cases, SmartCut can achieve a significant energy efficiency with acceptable impact on user experience.

VII. Conclusion and Future Work

In this paper, by analyzing large 3G trace, we have confirmed that the tail effect wastes a significant amount of power. We have found strong temporal correlations existing in 3G traffic workload. Based on those key observations, we have proposed a light weighted scheme, SmartCut, which uses historical 3G traffic data to train ARMA models and further utilizes the predicted arrival time of future data transmission to effectively cut unnecessary tails while having little side-effect to user experience. The trace-driven simulation results have shown that SmartCut are energy-efficient. As the power of transmission at the DCH state or FACH state may vary with signal strength, In the future, we are planning to study the impact of signal strength to SmartCut.

TABLE I
Energy savings

Applications	*Energy Savings (%)*
Web browsing	49.19
Video streaming	56.57
IM	30.16
Mixed	49.56

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant No.60970106, No.61170237, No.61170238 and No.61202375, the National High Technology Research and Development Program under Grant No.2011AA010502 and Science and Technology Commission of Shanghai Municipality under Grant No.12ZR1414900.

References

[1] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani, "Energy consumption in mobile phones: a measurement study and implications for network applications," in *IMC*. ACM, 2009, pp. 280–293.

[2] F. Qian, Z. Wang, A. Gerber, Z. Mao, S. Sen, and O. Spatscheck, "Characterizing radio resource allocation for 3g networks," in *Proceedings of the 10th annual conference on Internet measurement*. ACM, 2010, pp. 137–150.

[3] ——, "Top: Tail optimization protocol for cellular radio resource allocation," in *Network Protocols (ICNP), 2010 18th IEEE International Conference on*. IEEE, 2010, pp. 285–294.

[4] H. Liu, Y. Zhang, and Y. Zhou, "Tailtheft: leveraging the wasted time for saving energy in cellular communications," in *Proceedings of the sixth international workshop on MobiArch*. ACM, 2011, pp. 31–36.

[5] A. Schulman, V. Navda, R. Ramjee, N. Spring, P. Deshpande, C. Grunewald, K. Jain, and V. Padmanabhan, "Bartendr: a practical approach to energy-aware cellular data scheduling," in *Proceedings of the sixteenth annual international conference on Mobile computing and networking*. ACM, 2010, pp. 85–96.

[6] MathWorks, "Estimating ar and arma models." [Online]. Available: http://www.mathworks.cn/help/toolbox/ident/ug/bq54wup.html

[7] "Akaike information criterion." [Online]. Available: http://en.wikipedia.org/wiki/Akaike_information_criterion